2024-08-01

Using Causal Inference to Understand Public Perception Towards Electric Vehicle Adoption

Jesus Alejandro Gutierrez Araiza
University of Texas at El Paso

USING CAUSAL INFERENCE TO UNDERSTAND PUBLIC PERCEPTION

TOWARDS ELECTRIC VEHICLE ADOPTION

JESUS ALEJANDRO GUTIERREZ ARAIZA, B.S.I.S.E.

Master's Program on Industrial Engineering

APPROVED:

Sergio A. Luna Fong, Ph.D., Chair

Ivonne Santiago, Ph.D. & P.E., Co-Chair

Tzu-Liang (Bill) Tseng, Ph.D & CMfgE

Stephen L. Crites, Jr., Ph.D.
Dean of the Graduate School

Dedication

The greatest challenges humans face throw-out their lives are two: the challenge of where to start and the challenge of when to stop.

-Sameh Elsayed

After six years of crossing between worlds, Juarez and El Paso, I find this moment an appropriate time to reflect as a new door opens for me beyond the Mexico-USA Borderland.

When one graduates from High School, the future is full of uncertainties. We don't really know what we really want, who will remain by our side the upcoming years, or what the job market will be like when we graduate. There are thousands of questions that we may not have considered at that time, questions we might wish we had asked, regardless of where we come from.

However, there is a special question that no one really asks when moving to a new country as international student, which is: *Are you ready to reject tranquility early?* This question speaks to the commitment of always giving 101% in everything you do—academically and professionally. It means going beyond the standard requirements and staying constantly vigilant for job notifications, whether for internships or co-ops, anxiously wondering if you've passed the prescreening before the interview, all because you answered "Yes" to the question: *Will you need sponsorship now or in the future?*

I wish I had asked myself that question about tranquility back in 2018, but in retrospect, I realize that doing so would have meant missing out on the experiences, mentors, and friends I've gained, especially in the last two years. There is no doubt many of those moments were exceptional, funny and memorable, but also, they were moments to grow, to reflect, and to be transformed into a better person.

There are many people who supported me along this path. While I wish I could name all of them in this document, I will do my best to acknowledge those who were truly outstanding and who made a lasting impact on me personally, academically, and professionally.

First of all, I would like to thank my **parents**, Carla and Jesus, for this opportunity you gave all those years ago. We can agree that we did not really know about the challenges that could carry studying in other country, but here we are, about to end this path in UTEP, being the second master of the family. *Muchas Gracias!*

I would like to thank my **sister**, Ana Paola, for being these last three years my partner-in-crime, for allowing to share those funny but also enlightening moments with you. I am glad I was close during your High School stage and to train you as much as possible to be ready for your new experience in college that you are about to start soon. *Muchas Gracias!*

To my dear ***Abuelita***, Maria del Rosario (Chayo), for your half-month visits to Juarez, for those wisdom moments that were comforting, and for those times you prepared with love the classic and delicious *Tortillas de Harina* with beans and cheese, which I will never be tired of them. To my **great uncles**, Roberto and Adelaida, for allowing me to be a guest in your house in those moments that I was not possible to cross the border to arrive on time to event and of course to never forget to give me a *burrito* as lunch on every moment that it was possible. *Muchas Gracias!*

To my **High School friends**, for never forgetting me despite being in different universities, for allowing me to be present in your success events, and for letting me to have more, if not better moments than the ones we had back in High School. Ivan G., Lizbeth O., Andrea S., Ramses A., Brisa C., Mirab M., Lexi R., and Don Jose Luis P. *Muchas Gracias!*

To my **Church friends**, I know that despite that part of our life's journey ended unfortunately during the COVID-19, I am glad I joined to that path with you, and I shared those moments and even travels before and after our service and young church leader. Sebastián G., Luis R., Héctor A., Briana C., Sergio T., Gema M., Bryan L., Paulina O., Mauricio U., Eduardo G., Roberto V., Jaime M., Elisa D., Roberto L., Sebastián L., José C., Joao R, Sebastián L., Hugo S., and Paola D., *Muchas Gracias!*

To my **IMSE Laboratory friends**, with whom I shared most of my time these last two years, either as research colleagues or student organization partners. All those *Comedy Hours*, lunchtimes, jokes, research and professional development travels were worth it to reduce the stress in our research duties hours. Enoc F., Laura T., Irvyn H., Rene D., Karen G., Carla I., and Juan O. *Muchas Gracias!*

To those **professors and staff in UTEP** who taught me not just about being a better professional engineer, but also being a better person, and trust in me when I was offered either research opportunities or department involvement opportunities for my career since I was an undergraduate student. Dr. Luis Contreras, Dr. José Espíritu, Dra. Ana Cram, Dr. Sergio Luna, Dra. Ivonne Santiago, Dr. Juan Fernández, Dr. Jaime Sánchez, Dr. Bill Tseng, and Ms. Betsy Castro-Duarte, *Muchas Gracias!*

Thank you all for making me know that these six years were the best decision I could have taken, and this thesis is dedicated to all of you!

USING CAUSAL INFERENCE TO UNDERSTAND PUBLIC PERCEPTION TOWARDS ELECTRIC VEHICLE ADOPTION

by

JESUS ALEJANDRO GUTIERREZ ARAIZA, B.S.I.S.E.

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Industrial, Manufacturing, and Systems Engineering

THE UNIVERSITY OF TEXAS AT EL PASO

August 2024

Acknowledgments

This work was supported by the National Science Foundation (NSF) Engineering Research Center (ERC) Advancing Sustainability through Powered Infrastructure for Roadway Electrification (ASPIRE) under Grant No. EEC-1941524. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF.

I would like to express my gratitude to Dr. Sergio Luna and Dr. Ivonne Santiago for their trust on my skills to deliver this important project on the last two years, despite the multiple challenges and commitments each one faced from our battle fronts.

Finally, I would like to thank Dr. Bill Tseng for being part of this committee and providing feedback to improve this work.

Abstract

Why despite all efforts to promote Electric Vehicles (EVs) as an alternative transportation method through strategies such as tax credits on unit purchasing or long-term environmental benefits communication, its market penetration has not reached the expected goals in the United States? Even though there have been important advancements in the EV technical perspective and financial EV purchasing incentives, the final EV customers still face barriers on their scenarios that do not allow them to purchase this type of contemporary transportation means. Not understanding their local barriers could be a mistake that could reduce EVA expectations in the country and keep, if not increase, the environmental impact, which could impact public health.

The application of data collection methods such as surveys and the implementation of Machine Learning algorithms has been considered as the standard to run this type of analysis. However, the sample size bias (or imbalanced data) on the final dataset brings risks such as biased model performance, majority class overfitting, and misinterpretation of the results. This causes the stakeholders to deal with less prominent causes, wasting time and resources.

To address this challenge, it is proposed the fusion of the Sentiment Analysis of social media posts, U.S. census attributes and U.S. charging stations datasets focused on three U.S. cities: Indianapolis, Indiana (IN); Salt Lake City, Utah (UT), and El Paso, Texas (TX) through Causal Inference. This approach results in defining which attributes impact the most public perception (or sentiment) towards EVA.

The study results indicated that social media users' sentiment perceptions in the three cities were predominantly positive, followed by neutral. The diverse Electric Vehicle Adoption (EVA) conversation topics demonstrated empirically that the large volume of social media posts reflects the complexity of the topics discussed, such as EV Equity Awareness, EV Adoption Costs and EV

Charging Infrastructure. In addition, to establish a causality between the sentiment perception and external variables, it was discovered that those have a scarce if not no correlation between them, having to rely on the observed social media posts information in the built of Bayesian Belief Network that allows to know the impact of these alternative variables in the sentiment polarity.

The findings from this study give an addition to the decision-making process of stakeholders and policymakers regarding how to broadcast the EV transition message to communities with diverse backgrounds and develop strategies to achieve the EVA.

Keywords –*Electric Vehicle Adoption, Social media data, Sentiment Analysis, Machine Learning, Feature Selection, Electric Vehicle, Charging station, Causal Inference, Charging Infrastructure.*

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Background and Motivation

Environmental impacts such as Greenhouse gas (GHG) emissions, and pollution have significantly contributed to climate change, affecting our planet's capability to produce natural resources, regulate itself, and dispose of waste. The United States (US) exemplifies the repercussions of these environmental challenges. Notably, the transportation sector significantly contributes to the United States' carbon footprint, accounting for 29% of GHG emissions, caused mainly by Light Duty Vehicles commonly used to commute to school, workplaces, or recreational activities (U.S. Environmental Protection Agency, 2023). Popularity increment (Spencer et al, 2023; Popovich, 2024), public and private fleet electrification programs (The White House, 2023), charging infrastructure investment (Boushey, 2023), public awareness campaigns (Singh et al, 2023), and financial incentives (Hardman et al, 2017) have been some of the strategies implemented in the last years to motivate Electric Vehicle Adoption (EVA) in the United States. However, the adoption rate has not met the expectations of analysts, stakeholders (Morgan, 2023), and car manufacturers (Wayland, 2024) in the country. Pamidimukkala et al (2024) reported that based on a systematic literature review of articles published between 2018 and 2022, 70% of them were focused on the EVA research domain. This fact shows the growing priority of addressing this aspect on the EV research. In addition, the authors discovered that 57% of the published EVA articles used surveys as a methodological approach to obtain their results. Although this method dominates EVA studies, their nature, time-consuming, and costs limit the researchers to focus on certain population groups at a certain period, allowing the introduction of data biases such as the Hawthorne Effect (Zaleznik, 1984) and sampling bias. Therefore, the dataset compromises the performance capabilities and quality of Machine Learning algorithms' results, launching results

that might not concur with reality, and bringing unfair solutions to population groups that may have not been involved in the early design stages (Khan et al, 2022).

Nevertheless, there might still be a way to understand the main motivators of the public toward EV with such dataset limitations. By fusing the U.S. Census, EV charging station locations, and Twitter (now X) posts sentiment score as a consolidated dataset to be analyzed through Causal inference techniques could outstand those features that might have an impact on the EVA perception with the data available by that time.

1.2. THESIS CONTRIBUTION

This thesis provides significant contributions to the EVA field, allowing a better understanding of how barriers should be addressed at a local level. Providing a methodology that considers not only sociodemographic factors or charging infrastructure but also the social media post' may offer not only a different way on getting the main factors affecting the perception but also a way to test *use case* scenarios where this work can focus on variables and test them to see their impact on the response.

1.3. ORGANIZATION AND CHAPTER SUMMARIES

Chapter 2 provides key concepts and terminology used in this work and their use through the most recent literature. The chapter analyzes the current Electric Vehicle Adoption (EVA) Trend, social media as a contemporary data source, Sentiment Analysis as a technique to transform social media text attributes into numerical or categorical attributes, the current application of Machine Learning algorithms on EVA, and Causal Inference. Through this chapter, it is also defined the current potential advantages, disadvantages, and research gaps.

Chapter 3 provides the methodological approach implemented in this work. It introduces the cities where the study was implemented; the datasets required, particularly the ones from

Twitter, the US Census Bureau, and the US Department of Energy Alternative Fuels Data Center, the data collection process, the modeling framework, algorithms, and software implemented to make this study possible.

Chapter 4 offers the deliverable results as a product of the implemented methodology in this work divided into three areas focusing on quantitative analysis: Descriptive Analytics, Diagnostic Analytics, and Prescriptive Analytics.

Chapter 5 discusses the obtained results. This work identifies and interprets the implications of the results, connecting them with previous literature, highlighting their relevance in the area, and providing potential solutions.

Chapter 6 explains the limitations found during the development of this work. Additionally, it addresses potential biases and other factors that may affect the validity and reliability of the findings. By acknowledging these limitations, this chapter aims to provide a transparent and critical perspective on the scope and boundaries of the research.

Chapter 7 offers a guideline regarding future work. Suggestions for future work include expanding the study to different populations, employing alternative methodologies, and exploring related variables that were not covered in this manuscript. This chapter serves as a roadmap for researchers who wish to build upon the foundation laid by this study and advance the field.

Chapter 8 provides a conclusion summarizing the main key findings, possible applications, and upcoming challenges.

Chapter 2: Literature Review

This chapter provides a comprehensive analysis of the existing research on Electric Vehicle Adoption (EVA), social media as a Data Source, Sentiment Analysis, current Machine Learning techniques, and Causal Inference. Additionally, this chapter analyzes and evaluates key findings, implemented methodological approaches, and limitations found in the literature to propose a new methodology.

2.1 Electric Vehicles: Definition and Adoption Current state

According to Semanjski (2023), an Electric Vehicle (EV) is a means of transportation that uses electric motors powered by sources from an off-vehicle source or within the vehicle (such as a battery or electric generator). The International Energy Agency (IEA) (2024) considers EVs to be a key technology for decarbonizing road transportation. However, the IEA acknowledges that the transition has been slow to become a global phenomenon due to factors such as a lack of charging infrastructure and high purchase prices. In other words, while the technical development of EVs is crucial, making EVs accessible to the public must also be a priority in the transition to this transportation technology.

As a result, Electric Vehicle Adoption (EVA) has been gaining importance throughout the EV research trends in the last years. Recently, Pamidimukkala et al (2024) reported that there was an increment of articles addressing EV adoption behaviors from 7% in the 2012-2017 period up to 70% in the 2018-2022 period. Previously, Singth et al (2023) had split their EVA-related papers literature published from 2011 to 2022 into knowledge areas such as Engineering (25.33%), Energy (18.21%), Environmental Science (16.30%), and Social Sciences (14.65%). Both articles' authors demonstrated that the EVA research area has been showing its interdisciplinary nature, underlying the importance of exploring individual-level antecedents (Income, education level, age,

gender, among others) and Social Influence antecedents (personal attitudes, social pressure, public awareness, among others) to address potential barriers or motivators on EVA. The consensus drawn from this exploration is the significance of addressing potential barriers for adoption. Pamidimukkala et al (2024) classified the motivators and barriers into the following factors: Contextual, Situational, Psychological, and Demographic. As shown in Figure 2.1, the number of times (or frequency) that Situational factors were outstanding in the EVA literature was 1208 or 35.62%. However, the sum of Demographic and Contextual factors was 1271, or 37.48%, having a difference of only 1.86% with Situational factors.

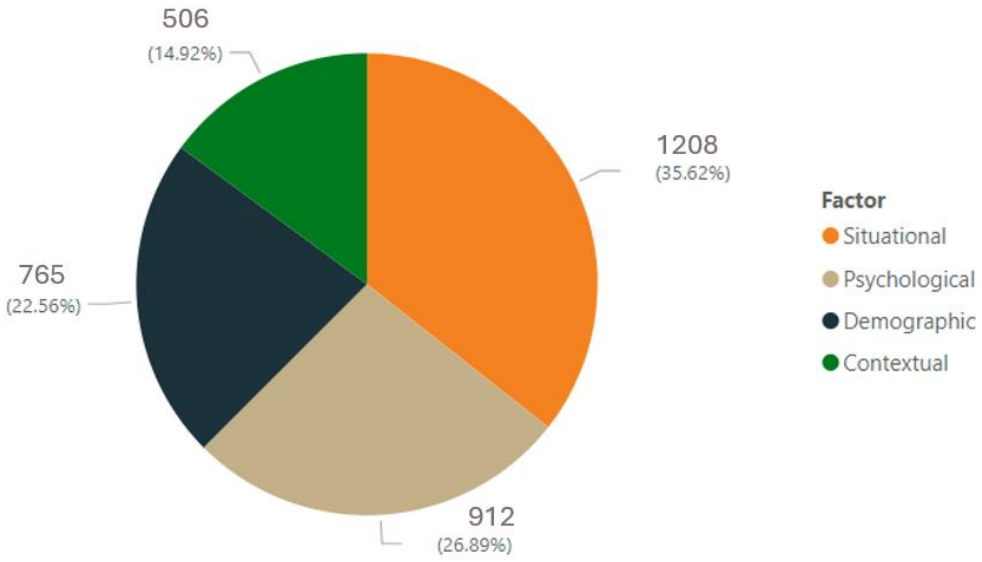


Figure 2.1. Distribution of factors that had an impact on the EVA according to Pamidimukkala et al (2024)

However, if the factors are decomposed into subcategories, as shown in Figure 2.2., it is possible to note that the Individual subcategory from the Demographic factor is the one that is mostly mentioned in the literature, with 16.39% of the total, being followed by the Technological (from the Situational factor) and Policy incentives (from the Contextual factor) subcategories, with 16.24% and 10.82% respectively.

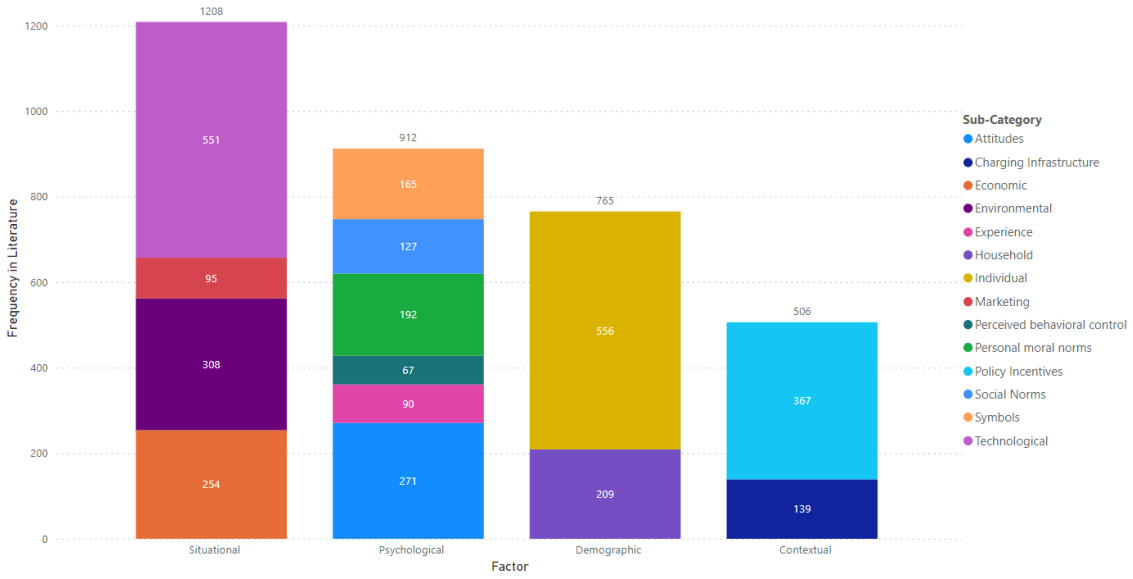


Figure 2.2. Decomposition of factors into subcategories that affected the EVA, according to Pamidimukkala et al (2024)

Considering the causing factors of these motivators and barriers to EVA, numerous research papers have underscored the importance of comprehending the EVA situation within their home countries. From a **global perspective**, Ruoso and Duarte Ribeiro (2022) comprehensively analyzed socioeconomic factors across 28 countries. Their selection criteria focused on nations representing nearly 96% of the global EV sales volume considering both the relevance of their contribution and the availability of the World Bank and Human Development Data Center and the International Energy Agency. The authors aimed to establish correlations between economic variables—such as Gross Domestic Product (GDP) per capita, Human Development Index (HDI), education index, total greenhouse gas emissions per capita, and pump price for gasoline—and the annual Electric Vehicle (EV) market share growth rate in each country. Utilizing regression analysis, including a Multiplicative Nonlinear Regression Model, their findings indicated correlations between renewable energy development and gasoline prices with CA. However, they

emphasized the pivotal role of geographical context and characteristics as outstanding factors in CA studies.

Regarding **country-level** analysis, Stajić et al. (2023) conducted a binational online questionnaire-based study on Croatia and Slovenia, Southeast European countries, seeking insights into the motivational factors and preferences regarding Battery Electric Vehicle (BEV). Surveying 278 individuals from both countries who had already purchased a BEV showed that initial BEV purchase cost, higher education level and income, including purchase incentive, were pivotal in their decision to transition to an Electric Vehicle.

From another perspective, Yang et al. (2023) delved into a country with a mature EV market – Norway. Instead of relying on survey data, they employed statistical data from Statistics Norway and TerraClimate. Employing Game Theory, Negative Binomial Regression Model, and Poisson Regression Mode, their investigation highlighted the positive effects of factors such as Vehicle Mileage, Urbanization, Income, and the number of Charging Stations. Conversely, the proportion of elderly individuals and low minimum temperatures were identified as factors that negatively affect EVA.

Considering the expansive territory of the United States, covering approximately 9.8 million square kilometers with diverse weather patterns, population demographics, and unique problem-solving perspectives, a localized analysis of EVA proves to be more insightful. Carley et al (2013) initiated one of the earliest studies, conducting an online survey in the country's twenty-one largest urban area. The survey, covering topics, such as Personal Attributes, General Beliefs, Vehicle ownership, Vehicle attributes of interest, EV and Infrastructure awareness, and their reaction to claimed EV disadvantages and advantages. Based on the results, they conclude that

despite all the advantages of an EV, disadvantages such as range anxiety and high purchase cost had a higher weight on their decision-taking.

A more focused study within the USA, this time at the **state level**, was undertaken by Gehrke and Reardon (2022). Using inputs such as local state census, passenger vehicle purchases and utilization at state level, they sought to determine predictors for increasing EVA in Massachusetts. Employing Logistic Regression to identify the probability of EV purchase based on statistically significant variables, they concluded that the interaction between housing and neighborhood characteristics, high-income households in single-family homes, and an increase in public charging stations emerged as critical predictors for EVA in the state.

Min et al (2023) explored the context of underserved communities in Seattle, Washington, focusing on significant variables hindering the adoption of Distributed Energy Resources (DERs), with EV chargers as a specific study element. Utilizing datasets from U.S Census Bureau 2014-2018 American Community Survey (ACS), EV Charger installation permit records (or to be specific, spatial data) from Seattle city, they applied Principal Component Analysis (PCA), Structural Equation Modeling (SEM) and K-Means clustering techniques. Their findings emphasized that housing tenure and type variables outweighed race and income considerations in EV charger adoption. Furthermore, advocating for housing-related support incentives emerged as a pivotal strategy to encourage residents to adapt their electric systems to DERs.

Finally, Lozada-Medellin et al (2023) evaluated perceptions, opinions, and knowledge of underrepresented communities (URCs) about electrified transportation technologies and access to EV infrastructure in three selected communities of El Paso, TX. Through focus group sessions and surveys as their methodological approach, they found that those communities had certain knowledge regarding EVs, perceiving the purchase cost, driving range, and charging cost as main

disadvantages. However, their perception was still positive since they considered that EV could enhance their air quality and cost-effectiveness in the long term. With that information, it was possible to create a community profile or community overview.

Finally, recent literature that focused on EVA showed that their methodologies had issues that needed to be considered when evaluating the results, as shown in Table 2.1.

Table 2.1. Recent literature showing Methodology issues on the EVA Analysis

Article/Paper	Problem Statement	Methodology Issue
Higueras-Castillo et al (2021)	Analyze the main factors that impact on the EVA in Spain through an online survey.	The sample was taken throughout the country but there was no grouping based on specific cities and the survey was offered only through an online survey platform.
Ruan and Lv (2023)	Analyze the public perception of Electric Vehicles in public social media posts from Twitter and Reddit	The users from Reddit post as anonymous, therefore it is impossible to connect them with a context, such as a city or user' gender.
Mpoi et al (2023)	Analyze factors and incentives that were affecting the EVA in Greece, focusing on Athens.	The analysis did not focus on the characteristics of the respondents in groups based on their characteristics.
Rye and Sintov (2024)	Analyze the perception of rideshare drivers and commuters regarding EV attributes in the USA.	The rideshare drivers' sample was limited to those driving in Los Angeles (USA) and the commuter population sample was obtained at a national level, having a different proximity level to the researchers among the samples.

As observed, many researchers employed surveys, questionnaires, and interviews as primary methodologies for data collection in EVA studies. While these methods offer valuable perspectives on EVA in specific locations, their datasets may cause a bias in the interpretation of the results due to their sample size bias, as warned in multiple articles (Austmann & Vigne, 2021; Ruan & Lv, 2022; Peng et al, 2024). Hence, new sources of information or methods that consider

the sample selection biases must be explored. In summary, EVA is a research branch at its peak due to its high interest. It requires a new perspective from which it can obtain those motivators or barriers that impact this decarbonization process.

2.2 Social Media as Research Data Source

Integrating social media data into research has become increasingly crucial for understanding public opinions on contemporary topics. According to the Pew Research Center (2021), since 2019 72% of U.S. adults have reported using at least one social media platform, Facebook, Instagram, LinkedIn, Pinterest, and Twitter (now X) as the most popular options. This significant statistic underscores for both learning and sharing information or opinions on real-time issues. Post content, timestamp, and engagement metrics (including the number of reactions and comments), as well as additional data fields such as the geographic location of the post submission and details about the post's author, are accessible through the social media network's Application Programming Interface (API). The availability of these data fields varies depending on social media company’s specific requirements and capabilities. Therefore, an extensive pool of data exists that can be harnessed for multifaceted analyses.

In the realm of Business Intelligence, the importance of social media data is gaining traction (Choi et al., 2020), aiding in the comprehension of customer sentiments and facilitating the design of Business Decision-Making Systems (BDMS) that can adapt to these sentiments (Yang et al., 2022). The manufacturing industry has also recognized the potential benefits of utilizing this data type, with discussions highlighting its capacity to drive innovation and necessitating the development of strategies for market expansion (Karmugilan & Pachayappan, 2020; Borah et al., 2022).

Finally, during the COVID-19 pandemic (Huang et al., 2022), social media data played a pivotal role. It enabled the development of methods for early warning detection alerts by gauging public attitudes and emotions (Luna et al., 2022), modeling the impact of fake news on social media (Frenkel et al., 2020; Pulido et al., 2020), and detecting an unfortunate increment in abusive

or hateful conversations (Babvey et al., 2021). In summary, fusing social media data has evolved into a critical tool in research, enabling better predictions, model development, and the formulation of strategies tailored to real-time environments. The inherent complexity of social media data, characterized by its diverse nature—whether structured or unstructured—and its ever-growing volume, coupled with the multitude of presentation formats (text, video, audio, Graphics Interchange Format), positions it squarely within the realm of Big Data (Hou et al, 2020; Rahman & Reza, 2022), This recognition mandates the application of sophisticated techniques to interpret each facet of this vast and dynamic dataset, with Sentiment Analysis standing out as a crucial method in this analytical toolkit.

2.3 SENTIMENT ANALYSIS

A primary methodology employed in the analysis of social media data is Sentiment Analysis, a branch from Natural Language Processing (NLP). It was defined by Feldman (2013) as "the task of finding the opinions of authors about specific entities by identifying sentences that contain comparative opinions and extract their perceptions". Sentiment Analysis serves the crucial function of converting categorical data (text from social media posts) into numerical data. This transformation enables the determination of the opinion polarity of the message—whether it is positive, negative, or neutral.

This methodology has become indispensable for interpreting the opinions and the evolving needs of social media users over time. A notable instance was during the COVID-19 Pandemic period (2020-2022), where it played a pivotal role in diagnosing public perception across various pandemic stages, such as the social distancing phase (Kaur et al., 2020; Kwon et al., 2020; De Rosis et al., 2020) and the vaccination process (Liu & Liu, 2021; Chinnasamy et al, 2022; Zulfiker et al, 2022). Furthermore, the application of Sentiment Analysis has extended to the business domain, contributing to the creation of forecasting models (Fan et al., 2017) and an understanding of sentiment impact on energy stocks (Reboredo & Ugolini, 2018). Finally, it has proven

instrumental in comprehending people's perceptions and reactions to unexpected events (Birjali et al., 2017; Neppalli et al., 2017; Mansour, 2018), politics (Wenando et al., 2020; Nugroho, 2021; Parra Aramburo et al., 2022) and unfortunate armed conflicts (Nandurkar et al.,2023).

2.3.1 Sentiment Calculation: The VADER method

Most of the techniques tend to evaluate the text to determine whether the sentiment is positive, negative or neutral regarding a certain topic in a text. One of those is known as the Valence Aware Dictionary and sEntiment Reasoner (VADER) (Hutto & Gilbert,2014) method, which is a rule-based tool that relies on a dictionary of words (lexicon) and a set of rules to determine that score. Its score computation is sensitive to both polarity (positive/negative) and intensity (strength) of emotion. In the years since its publication, multiple articles and research papers have put VADER in the spotlight as the contemporary method to calculate text sentiment.

Park and Seo (2018) analyzed social media posts from Twitter to analyze the perception regarding of three Artificial Intelligence Assistants in October 2017 by using VADER since It allowed to separate the tweets into the three opinion polarity types: positive, negative and neutral. After the method was implemented, traditional statistical approaches such as T-test, Kruskal-Wallis test and Mann-Whitney test were used to determine differences between the perception generated on the social media posts, allowing to provide a decision-making tool regarding AI Assistant selection based on their perception in social media.

Borg and Boldt (2020) integrated the method with the two variants of the Support Vector Machine (SVM) algorithm, the original and the *LinearSVC,* to analyze the sentiment of customer support e-mails generated in a Swedish Telecom corporation. After it was implemented, it was reported that the integration of VADER and the original SVM method allowed to obtain F1 score and AUC bigger than the model that was using the *LinearSVM* method only by a difference of 0.146 and 0.091 respectively.

Monselise et al (2021) not only analyzed the sentiment found on Twitter social media posts regarding the COVID-19 vaccination, but they also divided the sample into topics to determine

the sentiment evolution around those topics. It was concluded that the vaccination access and administration were the main concerns among the social media users, being fear the leading emotion in the sample just followed by joy. However, the authors warned that the users from urban areas could have been overrepresented in the dataset.

Long et al (2022) analyzed the posts generated on Reddit in the first two months of 2021 to understand the perception of the users regarding an American well-known videogame and gaming merchandise retailer. Once VADER was implemented, they were compared with the company intraday returns in the stock market. The authors concluded that despite the Reddit forum showing sentiments correlated to the market movements, it was still a risk to rely only on social media posts in the investment decision making process.

Jagini et al (2023) integrated the social media posts sentiment regarding the Bitcoin price and its number of posts produced per day with historical data of the cryptocurrency price per day. Once the posts were analyzed using VADER, the integrated dataset was used to train a Linear Regression model, obtaining a testing R2 score of 97.75%.

As it was shown, the VADER method has been used to transform its value from text into numeric. The literature showed that it has been an important component in the proposed methodologies before other statistical methods were applied, recognizing its utility.

While data visualization tools, including bar charts, heatmaps, and time charts, offer valuable insights, it is expected that the integration of statistical techniques on a comprehensive dataset—encompassing individual sentiment analysis, statistical data from the U.S. Census, and social media data—can yield crucial insights into the contextual landscape surrounding the generation of posts.

2.4. Statistical Approaches

Statistics has been playing a role in decision-making in multiple research fields in most cases. Through the literature, the introduction of methodologies on data collection, preprocessing,

manipulation, analysis, and interpretation has allowed researchers to obtain insights, uncover patterns, confirm hypotheses, and obtain credibility on their results, allowing others to replicate their methodologies.

There are two approaches On Statistics, the Frequentist and the Bayesian, that share a common objective: perform a proper statistical inference (Albers et al, 2018). Fornacon-Wood et al (2022) provided a guide to educate readers on the foundations of both statistical approaches, in addition to a case study in radiation therapy where both approaches are implemented. The differences between both approaches are shown in Table 2.2.

Table 2.2. Differences between the Frequentist and Bayesian Statistical Approaches according to Fornacon-Wood et al (2022)

Area	Frequentist	Bayesian
Hypotheses	Null hypothesis is true before data are collected (no effect of a particular treatment on dependent variable).	They are seen as probability distributions.
Probability	Assigned to the data, not to the hypothesis.	Assigned to the hypotheses.
Analysis	Analysis is driven by the data.	Incorporates prior information into the analysis, updating hypotheses probabilities as more data become available and based on experts beliefs.
Probability Computation Purpose	Obtaining p-value to calculate another dataset at least as extreme as the one collected	Determine that a particular hypothesis is true.
Interpretation	Tends to be misinterpreted	Intuitive
Estimation with Uncertainty	Use of Confidence Intervals (Albers et al, 2018)	Use of Credible Intervals (Albers et al, 2018)

It is important to notice that while the dataset is critical for the analysis from the frequentist perspective, with the Bayesian perspective, the focus is on the probability that the hypothesis could be credible based on the data available (or prior knowledge). In addition, Fornacon-Wood et al

(2022) stressed that data quality was still important for the analysis. In other words, following data mining standards, such as CRISP-DM (Schröer et al, 2021), is still advisable in both approaches.

In the upcoming subsections, methods using the Frequentist and Bayesian approaches will be introduced, including their advantages, disadvantages, and some applications reported in the literature. It is primordial to remark that the discussion will be focused on only those methods (or techniques) that can post signs on those variables that stand among the variables available on the dataset.

2.4.1. Frequentist approach: Machine Learning Feature Selection Techniques Current trend

Machine Learning (ML) Techniques go beyond mere prediction based on existing data; they can replicate tasks that typically demand weeks or even months of a researcher's time, accomplishing them in seconds or hours. Unlike traditional statistical methods, ML techniques excel at classifying objects based on specific features or groups of features and identifying the features that genuinely elucidate the response. Despite the introduction of numerous variants in the literature, ML techniques, by their very nature, continue to be employed, yielding valuable insights. However, sometimes, the considerable number of variables (or features) or high dimensionality, that a dataset may offer could threaten the ML modeling performance metrics, such as accuracy or overfitting. This is a constant problem that has been identified through the literature that should be considered in this work. For this literature review, the focus will be on those ML techniques that can perform Feature Selection, which refers to those algorithms that can select a subset of existing features (or variables) that contribute most significantly to the response variable. Rudin (2019) stated that an ML model capable of not only predicting, but also being interpretable and faithful to what the model computes, might ensure safety and trust in the model for high level decision-taking processes. Therefore, based on Lamberti's (2023) overview

regarding the explainability and interpretability of Artificial Intelligence (AI) algorithms analysis, the literature scope will be reduced to three regularization methods considered to possess high levels of model understanding in the Dimensionality reduction realm: Ridge, Lasso, and Elastic Net.

2.4.1.1. Ridge Regression

Ridge Regression was introduced by Hoerl & Kennard (1970) as a technique to reduce the high outlier sensitivity and overfitting caused by the implementation of the Ordinary Least Squares (OLS) method or Linear Regression. Its main purpose is to estimate the best coefficient explanatory variables set that is reasonable and it can explain the variability of the response variable.

To achieve this purpose, minimizing the Sum of Squared Residuals (SSR) between the original response variable and the model response variable is necessary. The following Ridge objective function is provided below, given a *nxm* dataset:

$$min_{\beta_0,\beta} \sum_{i=1}^{n} \left\{ y_i - \left[\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} \right] \right\} + \lambda_2 \sum_{j=1}^{m} \beta_j^2$$

Where

n	Dataset Row (or object) dimension
m	Dataset Column (or attribute) dimension
β_0	Model original intercept
β	Attribute or explanatory variable coefficients vector
β_j	Explanatory variable coefficient *j*
x_{ij}	Data entry of Object *i* for explanatory variable *j*
y_i	Response variable *i*
λ	Model Regularization Hyperparameter

$\lambda \geq 0$

As a regularization technique, the Model hyperparameter (λ) or *L2 regularization parameter*, has the capability of penalizing larger coefficients, playing an important role in balancing the model simplicity and predictive performance of the model. The hyperparameter multiplied by the coefficient-squared summation brings the *shrinking penalty*, which will penalize those models with higher explanatory variable coefficients (except β_0), assuring to bring a model that provides those features that really affect the final response variable. This Model hyperparameter can be obtained either empirically or through cross-validation, which can provide the best λ that could be used to obtain the best Ridge model.

Starting on iteration 0, where β is obtained from the OLS method, it is possible to start computing the SSR. To update the coefficients per iteration, it is recommended to use optimization techniques such as Gradient descent, Coordinate descent, or Stochastic gradient descent, depending on the available computational power and dataset dimensions. This process is repeated until the minimum SSR cost is achieved or the maximum number of iterations (previously defined) has been reached, depending on the approach taken.

The technique has been applied to the Education sector by determining those climate factors that had an importance on the school's overall academic performance (Quammie & Hosein, 2024), and extracting those variables that could determine Graduate School admission, being outstanding by having lower RMSE and MAE scores, and higher R^2 score than other regression models such as Decision Tree, Gradient Boosting, Random Forest or Support Vector Machine (Krishna Kireeti et al, 2023).

In addition, in the Healthcare sector Ridge exceeded expectations over the Recurrent Neural Network (RNN) by 5.36% (on average) in terms of accuracy in Brain tumor severity prediction (Bilal & Tamiselvan, 2024). In the same way, Ridge has been used to obtain the set of variables that could perform an accurate predictive stroke risk assessment model, using Bootstrap as a model validation method (Jeena & SukeshKumar, 2018).

Through the literature, multiple studies have found advantages when using the algorithm compared with other techniques. In addition to dealing with multicollinearity problems (Yang &

Wen, 2018), through the literature it has been considered that this deterministic approach costs less time to construct the models, allowing interpretability between the features and response variables (Yang & Wen, 2018) and a Bias-Variance tradeoff (James et al, 2023).

In the same way, the literature has also reported that this method possesses weaknesses that can affect feature selection. Bilal and Tamilselvan (2024) concluded that Ridge required a large dataset to explode its selection capabilities. Multiple authors (Muthukrishnan & Rohini, 2016; Grampurohit and Sunkad, 2020; Chumachenko et al, 2021; Shiomi et al, 2022) reported that Ridge could not perform variable selection well since all the variables would have a coefficient close to but not equal to zero, keeping all predictor variables in the final model, affecting the model interpretation (James et al.,2023; Melkumova & Shatskikh, 2017; Pereira et al.,2016).

Ridge Regression is considered as the first ML technique to perform feature selection subject to regularization. Overfitting prevention and interpretability are its main advantages, while poor variable selection capability and bias introduction are its main disadvantages. These are the characteristics that need to be taken in consideration when selecting a model that could provide the main features that impact the response variable.

2.4.1.2. Lasso Regression

Tibshirani (1996) introduced the Lasso (Least Absolute Shrinkage and Selection Operator) Regression to address not only the linear regression method limitations but also to provide a different solution perspective from the one provided by Hoerl & Kennard (1970). This technique allows to be shrinking of coefficient explanatory variables to zero, assuring a set of outstanding variables capable of explaining the response variable.

To achieve this purpose, it is necessary to minimize the Sum of Squared Residuals (SSR) between the original response variable and the model response variable. The following Ridge objective function is provided below, given an *nxm* dataset:

$$min_{\beta_0,\beta} \sum_{i=1}^{n} \left\{ y_i - \left[\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} \right] \right\} + \lambda_1 \sum_{j=1}^{m} |\beta_j|$$

Where

n	Dataset Row (or object) dimension
m	Dataset Column (or attribute) dimension
β_0	Model original intercept
β	Attribute or explanatory variable coefficients vector
β_j	Explanatory variable coefficient *j*
x_{ij}	Data entry of Object *i* for explanatory variable *j*
y_i	Response variable *i*
λ	Model Hyperparameter
$\lambda \geq 0$	

As a regularization technique, the Model hyperparameter (λ) or *L1 regularization parameter*, has the capability of penalizing larger coefficients, playing an important role in balancing the model simplicity and predictive performance of the model. The hyperparameter multiplied by the coefficient-squared summation brings the *shrinking penalty* or *L1 regularization parameter*, which will penalize those models with higher explanatory variable coefficients (except β_0), assuring to bring a model that provides those features that really affect the final response variable. This Model hyperparameter can be obtained either empirically or through cross-validation, which can provide the best λ that could be used to obtain the best Lasso model.

Starting on iteration 0, where β is obtained from a linear regression method (such as OLS), it is possible to start computing the SSR. To update the coefficients per iteration, it is recommended to use optimization techniques such as Gradient descent, Coordinate descent, or Stochastic gradient descent, depending on the available computational power and dataset dimensions. This

process is repeated until the minimum SSR cost is achieved or the maximum number of iterations (previously defined) has been reached, depending on the approach taken.

This technique has been applied to the Power generation aspect by predicting the power output of a solar PhotoVoltaic (PV) system on different year months, considering Lasso and XGBoost Regression techniques as the most suitable models to predict the power output (Sanewal & Khanna, 2023). Furthermore, Yeom and Choi (2018) derived 25 variables (from 232) based on the annual electric power produced in a manufacturing plant using Lasso, with a 79% accuracy.

In addition, this technique has been implemented in the Healthcare aspect by analyzing how Lasso could support the performance of other ML methods (Support Vector Machine, Decision Tree, Linear Regression, and Random Forests) by extracting those spatiotemporal features obtained from motion sensors during hand rotation tests to detect Parkinson's disease (Javed et al, 2018). Similarly, Wu et al (2017) discovered that by fusing Lasso and Logistic regression, it was possible to improve the accuracy on a Electronic Health record classification model.

Multiple articles have confirmed Lasso's advantages when comparing it not only with Ridge but also with other advanced Machine Learning methods. Lasso Regression can perform variable selection by shrinking non-outstanding variables and improving model interpretability (Muthukrishnan & Rohini, 2016; Melkumova & Shatskikh, 2017; Hassan et al, 2023), as reported in the literature, being capable of dealing with multicollinearity problems (Yang & Wen, 2018), preventing overfitting (Han et al, 2022), and allowing Bias-Variance trade-off (James et al.,2023).

However, it has been reported that when comparing with Ridge, Lasso tends to obtain higher bias, variances, and MSE (Mean Square Error) results (James et al.,2023). In addition, it depends on the lambda or tuning parameter which determines the number of outstanding variables,

if the predictor variables will have equal size when providing a response variable (James et al.,2023). Grampurohit and Sunkad (2020) reported that this method cannot select more variables than the number of data rows available. Moreover, in the case of having correlated predictor variables between each other, the model will select one of them. Fan et al (2015) stated that despite of that the Lasso variable selection capability was better than other methods (such as Elastic Net or Ridge), its prediction was not as accurate as with other methods. Finally, it is suggested to refrain from reporting p-values when evaluating the model (Pereira et al.,2016; Tomaschek et al., 2018).

Lasso Regression is considered as a successor of Ridge Regression to perform feature selection subject to regularization. Real variable selection, model simplicity and interpretability are its main advantages, while bias introduction and arbitrary selection of only one variable when two or more variables interact between each other are its main disadvantages. These are the characteristics that need to be taken into consideration when selecting a model that could provide the main features that impact the response variable.

2.4.1.3. Elastic Net Regression

Introduced by Zou and Hastie (2005) as the sum of efforts performed by Hoerl and Kennard (1970), and Tibshirani (1996), this technique combines the Lasso (L1) and Ridge (L2) regularization penalties into an objective function, given a *nxm* dataset, as shown below:

$$min_{\beta_0,\beta} \sum_{i=1}^{m} \left\{ y_i - \left[\beta_0 + \sum_{j=1}^{n} \beta_j x_{ij} \right] \right\} + \lambda_1 \sum_{j=1}^{n} |\beta_j| + \lambda_2 \sum_{j=1}^{n} \beta_j^2$$

Where

n Dataset Row (or object) dimension

m Dataset Column (or attribute) dimension

β_0 Model original intercept

β Attribute or explanatory variable coefficients vector

β_j Explanatory variable coefficient *j*

x_{ij} Data entry of Object *i* for explanatory variable *j*

y_i Response variable *i*

λ Model Hyperparameter

$\lambda \geq 0$

As a regularization technique, the Model hyperparameters (λ) or *L1* and *L2 regularization parameters*, have the capability of penalizing larger coefficients, playing an important role in balancing the model simplicity and predictive performance of the model. The hyperparameter multiplied by the coefficient-squared summation brings the *shrinking penalty*, which will penalize those models with higher explanatory variable coefficients (except β_0), assuring to bring a model that provides those features that really affect the final response variable. This Model hyperparameter can be obtained either empirically or through cross-validation, which can provide the best λ that could be used to obtain the best Ridge model.

On the stock market area, Elastic-Net was implemented by Szczygielski et al (2023) using not only stock market data but also COVID-19 information (i.e. growth in cases, growth in deaths, or growth in recoveries) to determine how aspects such as geographical proximity to the virus's outbreak and a country's development level may impact in the stock market trends. Additionally, it was also used by Szczygielski et al (2024) as a method to create a general stock market-related index based on Google search data to provide a better understanding of the narrative proposed by the Google Search Trends and its connections with stock markets.

In addition, Elastic Net has been used to analyze possible behavior causes in the population. With that technique, Shiomi et al (2022) were able to determine that the current legislation, infrastructure, and education could be indicators of potential traffic accidents or violations at the country level after they fused Country Fact Survey (CSF) data and international questionnaire survey data. In the same way, but at the county-level, Keeney et al (2023) examined those factors that were allowing the increment in the child abuse and neglect rates, such as having an Agriculture, Forestry, Fishing (AFF) occupation, educational opportunities, or regional biases, in selected USA states.

As a combination of the Ridge and Lasso Regression capabilities, It prefers the simplest model with the least necessary predictors and penalizes the inclusion of unnecessary predictors (Tomaschek et al., 2018). Before Lamberti (2023), Fan et al (2015) had confirmed that Elastic Net was considered as the most interpretative model when comparing with Lasso or its variant, Adaptive Lasso.

Elastic Net can also retain the group of predictors, even those that are correlated to each other, as Grampurohit and Sunkad (2020) reported. Srisa-An (2021) concluded that Elastic Net outperformed its model performance metrics [R-squared (using testing data), MAE, MSE] against other models such as Ridge, Random Forest, and Decision Tree. However, Bangroo et al (2023) considered that this technique could be computationally expensive and the tuning of hyperparameters was more complex before running the model, while Monteiro et al (2024) detected a bias generation in the training set when the model was being generated, obtaining lower performance metrics than Lasso.

Elastic Net Regression could be considered as a direct descendant of both, Ridge and Lasso, to perform feature selection subject to regularization. Real variable selection, model

simplicity and interpretability are its main advantages, while bias introduction and hyperparameters tunning complexity are its main disadvantages. These are the characteristics that need to be taken into consideration when selecting a model that could provide the main features that impact the response variable.

2.4.1.4. Sample Selection Bias: Imbalanced Dataset Impact on ML Performance

As has been shown in sections 2.5.1.1 through 2.5.1.3, Feature Selection ML techniques have obtained high-performance metrics when compared to other techniques if not between each other. However, all these results rely on one common factor: the dataset's nature and Sample selection bias. Bangroo et al (2023) warned that the context where the dataset and problem to be solved were important to consider when evaluating the ML Algorithm performance metrics. In the same way, Cho et al (2023) recognized that ML techniques, despite their virtues and capabilities, could not be able to reduce the impact of a poor study design that could have brought quality issues on the final dataset. As it was shown on section 2.2, Social Media data could be an alternative to data from sources such as survey or interview, however, it introduces new challenges and bias types to consider when analyzing its data. Hargittai (2018) demonstrated that when social media data was being used, there was a lack of representativeness of social media users that should be considered when reporting findings, driving to a selection bias. In other words, the number of social media posts (assuming each post was authored by a a different user) cannot be equal to the population itself since not everyone has access to a social media platform. Therefore, findings are suggested to be taken with caution, considering the dataset limitations. In addition, based on a personal email communication between Loni Hagen and David Garson (Garson, 2021), Hagen stated the following:

> *"It was observed that Machine Learning was good at learning biases and the majority rules of the world but as a consequence, minority rules can be easily ignored in the process of machine learning"*

In other words, the dataset nature, even with all the pre-processing data techniques applied, could compromise the real ML algorithm performance by showing wrong analysis and interpretation results but still acquiring high performance metrics. This concern was confirmed by Cho et al (2023) who considered that the decision-taking processes, where a ML technique was involved, should be supported by a fair or non-discriminatory dataset. They considered that the origins of data collection, such as biases on measurement, representation and sampling, could produce biased results and inaccurate prediction, bringing misinformed interpretations.

Finally, simulating data based on the data available, although tempting, brings also challenges. Mo et al (2024) determined that such action, in addition to high computational costs, could bring mismatch between the simulated and empirical data, leading to poor generalization when new data is used in the proposed model. In summary, the researchers informed that the lack of data must be treated seriously to obtain the closest possible model to the real-world scenario.

To understand how the Sample Selection Bias is presented, Enders (2010) presented the three missing data treatments that can be implemented depending on the dataset nature.

- **Missing Completely At Random (MCAR).** This mechanism considers that the probability of a missing value for a measured predictor variable is independent of other predictor variables and the response variable. In this case, the missing data is random, and there is no systematic difference between the missing data and the observed data.
- **Missing At Random (MAR)**. This mechanism considers that the probability of missing data one measured predictor variable could be dependent on other measured predictor variables but not to the response variable. In this case, the missingness can be related to the observed data, but not to the unobserved data.

- **Missing Not At Random (MNAR).** This mechanism considers the probability of missing data of one measured predictor variable due to unmeasured predictor variable, which could include the response variable.

A different way to represent missingness mechanisms is by using graphical representations, as proposed by Enders (2010). As shown on Figure 2.3, It is proposed that each mechanism will contain four components: predictor variable(s) (*x*), response variable (*y)*, missing data probability indicator (ranging from 0 to 1) (*R)* and unmeasured predictor variables (*Z*).

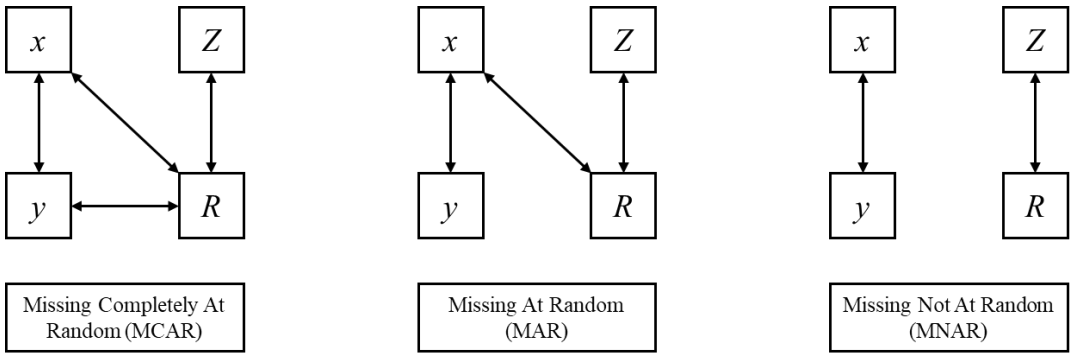


Figure 2.3. Graphical representation of the missingness mechanisms

At it is shown in Figure 2.3, the mechanisms consider two sides, the observed data (predictor and response variables) and the unobserved data (unmeasured variables and missing data probability indicator). First, the **MCAR mechanism** considers that the missing data probability indicator will be the one in charge of allowing the relationship between all the observed data entities and the unmeasured variables. In other words, the missingness probability is completely random and it not related to either the measured or the unmeasured variables. Second, the **MAR mechanism** considers only the measured predictor variables will connect with the probability indicator, still allowing the full connection of the panel, despite that the response

variable is not connected directly to the probability indicator. In other words, the missingness is connected to observed data but not to the missing data itself. Finally, the **MNAR mechanism** considers that there is no connection between the observed and unobserved data, implying that the missing data probability is related to the missing data itself.

To enhance these explanations, an example is provided, using Figure 2.4 as reference. It is assumed that all the opinions regarding a topic are produced by two sectors: social-media users and non-social media users. Inside each group, two age ranges emerge: the one of those who are between 18 and 40 years old and the one of those who are between 41 and 70 years old. The ideal dataset should be the one where all the population members are involved, as shown in Figure 2.4. However, it is known that in the real-world and not everything will be perfect as planned.

As an example of the **MCAR concept**, it happens when random people may not participate due to reasons independent from the ones considered on the study, such as being in an age group or an opinion sector, not being systematic in other people. On the Figure 2.4, it is observed that two groups are complete, as in the ideal dataset, while the other two lack at least one individual, either male or female.

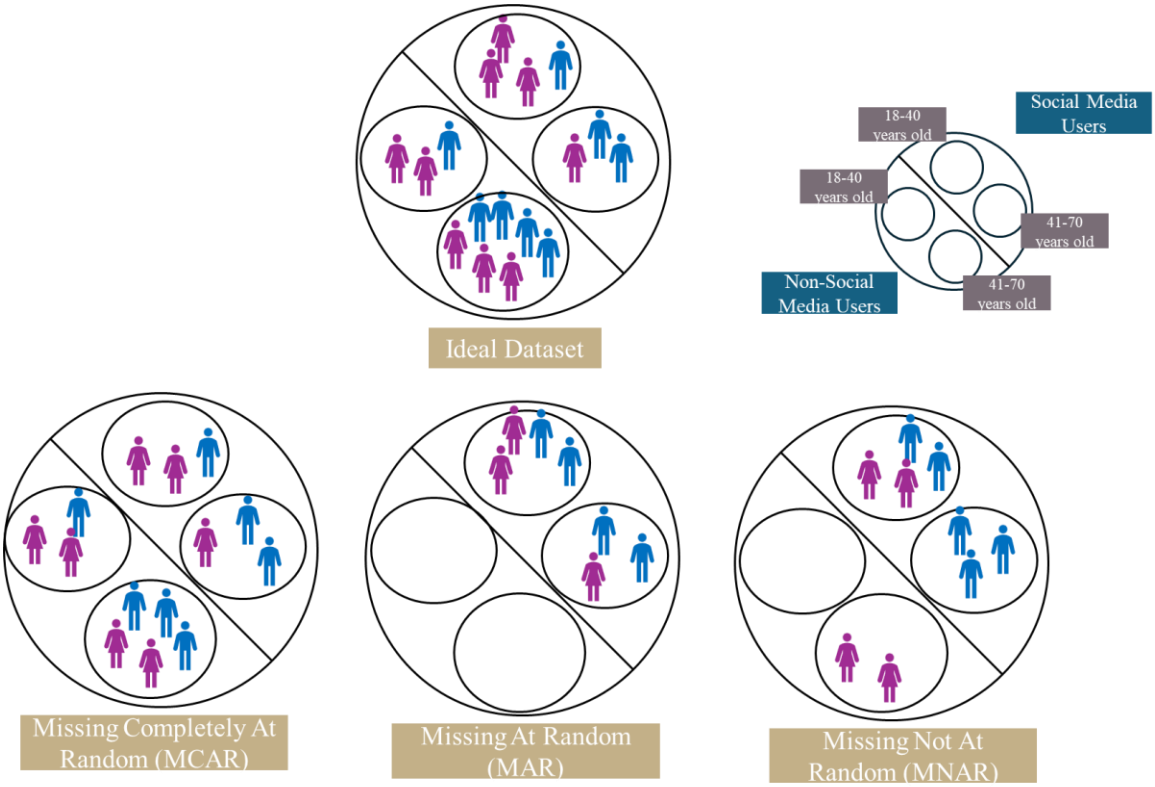


Figure 2.4. Representative explanation of the Missingness mechanisms. Own creation

The **MAR concept** example considers that the study may only focus on only the population that is a Social Media user, not considering the other sector (Non-Social media user), having bias since not all the population is being considered due to the study design. In addition, some individuals from the studied sector may not still being able to participate due to reasons independent from the variables considered in the study, as seen on the MCAR example.

Finally, the **MNAR concept** example would show certain patterns on the lack of participation in the study, as shown on Figure 2.4. One pattern could be that all the Non-Social media users members from the 18-40 years old group do not share their opinion since they fear to be embarrassed when showing it, whether positive or negative. Other pattern could be that all the males or all the females who belong to the 41–70-year-old group may prefer not to participate since they were not allowed to participate by their companies due to reasons connected to the topic.

In addition, through the literature it has been reported how the Sample Selection Bias (SSB) has been considered as an issue in the methodology when obtaining results, offering potential solutions as shown in Table 2.3.

Table 2.3. Sample Selection Bias Examples and proposed solutions

Article/Paper	Problem statement	Methodology Issue	Possible Solution
Marshall et al (2010)	Determine the variables that significantly influence the probability of default credit score based on historical data from a USA bank.	Single-stepwise procedure or *ad-hoc* bank expert judgment system, causing bias in the credit scoring cards and potentially excluding genuine explanatory variables	Explore the significant variables of loan performance and approval decision processes by using the bootstrap variable selection procedure
Kekkonen et al (2015)	Analyze adolescents' mental health status through a survey in Eastern Finland evaluating demographics, school performance, depression, and alcohol use.	It did not consider those students who were absent from school when the survey took place and poor information the students had in some survey sections (parents' professional status)	Intensify recruitment strategies (telephone or email contact, incentives, parental support) focusing on recruit adolescents belonging to the high drop-out risk group
Luna (2019)	Determine leading indicators that influence public perception in social media regarding evacuation due to natural disasters in the USA	The nature of social media data from by then Twitter was focused only on those users who posted content related to the disaster, instead of the whole population affected by the phenomenon.	A Missing-At-Random (MAR) mechanism was implemented in a node of the Bayesian Belief Network (BBN) model based on the social media post user's gender and location when posted.

Moua et al (2020)	Mapping of Species Distribution Models (SDMs) to predict species habitat distribution subject to environmental features.	Access to certain geographic areas to take sampling might not be available and the number of presence sites is limited, causing an environmental bias.	The use of *BGgeo* since it detected the highest number of presence sites compared with other methods. The technique is based on geographical criteria which assumes that the habitat characteristics are similar within the geographic space. It is suggested to supplement or confirm results with other methods, such as Ecological Niche Factor Analysis (ENFA) or Generalized Linear Model (GLM) and Generalized Additive Model (GAM)
Chen et al (2022)	Analyze the characteristics (such as demographic, health, social support, among others) of People Aging with Long-Term Physical Disability (PAwLTPD) in	Using only one survey format, such as web-based, may cause not all the PAwLTPD to participate in the survey due to accessibility issues.	A survey was offered in two formats: Phone and Web.
Xu and Lin (2023)	Evaluate residents' attitudes towards Garbage Incineration Power Plants (GIPP) in main Chinese cities by analyzing the Not In My BackYard (NIMBY) effect.	Not all the respondents were positive to Willingness To Pay (WTP), either underestimating or overestimating the NIMBY effect.	Design of a sample selection model that considers a selection and a elicitation function. In addition, two additional questions that allow seeing the respondents' willingness to pay more ro avoid GIPP

			construction near their residence.
Zhao et al (2023)	Forecast the movie demand by using historical sales data.	Defective design of the sample selection process, where a subset of the data is systematically excluded due to a particular attribute.	A new Model Averaging Optimal Correction (MAOC) was created based on the Heckman two-step estimator.

It is expected that these concepts and examples support the understanding of this sort of bias, which does not allow us to obtain the whole model as it is on real-life. These missing data mechanisms should be considered alongside the limitations that the frequentist approaches may still have even if the bias is solved.

2.4.1.5. Machine Learning Limitations

Through the literature, either on research journal articles or conference papers, authors have elevated the predictive capabilities of Machine Learning (ML) algorithms in multiple research fields. However, it is also important to consider their limitations due to their nature, in addition to the dataset structure provided (explored on section 2.5.).

On the healthcare sector, ML methodologies should be used carefully since they could be important decision-maker components of physicians when deciding the proper treatment to the patients based on their clinical records, analysis results and vital signs monitoring. Kelz et al (2021) provided critical feedback regarding the adoption of the Predictive Optimal Trees in Emergency Surgery Risk (POTTER) tool in emergency general surgery. They argued that the ML tools implementation tends to make difficult the interpretation of the performance metric results and their translation not only to the medical context but also to the context of the patient family. In addition, it was considered that the lack of knowledge area or patient recovery objectives

implementation on the ML models could underestimate the risks and final patient care results. In other context, Cho et al (2023) introduced guidelines of how to integrate ML techniques into the scientific method focusing on the health sciences. They considered that the scientific method gives importance of the prediction's interpretability need regarding a phenomenon. Such a need is difficult to achieve when the ML predictive model cannot explain how it reached the provided predictions, also known colloquially as "black box" model. Finally, in the Forensic sciences field, Barash et al (2024) introduced a literature review regarding the ML methods application in the context of forensic DNA analysis. They concluded that ML methods should be as transparent as possible to allow proper results interpretation and ensure their acceptance in legal context.

In the social sciences, researchers tend to use ML techniques on their methodologies to understand human behavior and social trends around a study object, maximizing the discovered findings. Suvorova (2022) analyzed the social sciences research state where ML techniques were being implemented on the research methodologies. Despite the existence of human validation and estimation testing of the implemented methodologies, the author concludes the need of detecting relationships between variables and the design of interpretable ML techniques (colloquially known as "white box" techniques) is critical when listing possible solutions and to allow interaction with researchers from other domain knowledge area. In the Economics field, Wu et al (2023) focused on demonstrating how the use of ML methods have had an impact on Economics research, focusing on two aspects: prediction and modeling. However, they admitted that there should be a cautious use of those methods, even the most recently created, as skepticism prevails regarding their "black box" nature, irreproducibility, and priority of results over-explanation.

2.4.2. Bayesian Approach: Causal Inference Current Trend

2.4.2.1. Bayesian Belief Network

Causal inference is not a new research field. Still, it has been growing in importance through multiple research areas in the last few years since its methodologies adapt to the available data, being one of its exponents the Bayesian Belief Network (BBN). This concept requires to remember a basic Probability concept: the joint probability, which represents the likelihood that two events occur at the same time given the probability of an event multiplied by the probability of the second even given that the first event happened. The joint probability mathematical expression is represented below:

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A)$$

Where

$P(A,B)$	Joint Probability of Event A and Event B
$P(A\|B)$	Conditional Probability of Event A, given Event B
$P(B\|A)$	Conditional Probability of Event B, given Event A
$P(A)$	Marginal Probability of Event A
$P(B)$	Marginal Probability of Event B

Following this first approach, Bayes (1763) proposed that it was possible to obtain the probability of an event about happen given prior recorded evidence of a different event, starting the inductive reasoning science (Pearl, 1982). In other words, instead of obtaining the probability that both events could happen, Bayes was attempting to demonstrate that if the user had the probability of the two events as independents from each other and the probability of the first event to happen given the second event happened, it was possible to obtain the other probability that the second event could happen given that the first event happened. The Bayes Theorem mathematical expression is provided below:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Where

$P(A\|B)$	Probability of Event A to happen, given event B has occurred
$P(B\|A)$	Probability of Event B to happen, given event A has occurred
$P(A)$	Probability of Event A to happen
$P(B)$	Probability of Event B to happen

These foundations allowed Pearl (1982) to introduce the BBN concept as a method to propagate the impact of new beliefs or evidence through a network. This method allows researchers to study a set of joint probability distributions of variables with a potential causal relationship, represented as Directed Acyclic Graphs (DAGs), a central focus of Causal Inference. Being a graphical model, a DAG contains nodes representing the random variables and edges representing relationships between random variables. In addition, each node possesses either a Conditional Probability Table (CPT), which shows all the probabilities of possible node events given all the possible parent node events, or a Marginal Probability Table (MPT), which only shows the possible node events when there are no events. Finally, there are three possible node configurations in a DAG: the chain (A→B→C), the fork (B←A→C), and the collider (B→A←C).

The main goal of BBN is to estimate the impact of a treatment effect implemented in a node (or variable) into another node, such as a response variable, capturing conditionally dependent and conditionally independent probability relationships between variables. A key characteristic of Bayesian networks is the conditional independence rule which states that each node is conditionally independent of its non-descendants given its parents (Pearl et al, 2016). In other words, a node A probability cannot be affected by a node B when node B is not a descendent

from node A, unless node B is a node A parent. A Joint Probability formula for BBN is provided, also called as the Rule of Product Decomposition (Pearl et al, 2016):

$$P(x_1, \dots, x_n) = \prod_i P(x_i | pa_i)$$

Where

x_i Variable node *i*

pa_i Parent variable(s) of variable node *i*

It supports researchers to find the probability in the whole DAG given the available attributes and their responses, usually as binary (True or False). However, it is preferred to have a single objective or target (either set to True or False) to analyze its reaction to the provided evidence from one or multiple nodes. To address this, Pearl (1982) proposed that the probability computation of two target scenarios (assuming the target data type node was binary), when the target is True and False, should be computed separately before obtaining the final probability as shown below.

$$P\left(x_{target} = True \middle| x_1 = e_{1,1}, \dots, x_n = e_{n,m}\right) = \frac{P\left(x_{target} = True, e_{1,1}, \dots, e_{n,m}\right)}{P\left(e_{1,1}, \dots, e_m\right)}$$

$$= \alpha P(x_{target} = True, e_{1,1}, \dots, e_{n,m})$$

$$P\left(x_{target} = False \middle| x_1 = e_{1,1}, \dots, x_n = e_{n,m}\right) = \frac{P\left(x_{target} = False, e_{1,1}, \dots, e_{n,m}\right)}{P\left(e_{1,1}, \dots, e_{n,m}\right)}$$

$$= \alpha P(x_{target} = False, e_{1,1}, \dots, e_{n,m})$$

$$\alpha = \frac{1}{P\left(x_{target} = True, e_{1,1}, \dots, e_{n,m}\right) + P\left(x_{target} = False, e_{1,1}, \dots, e_{n,m}\right)}$$

Where

x_{target} Target node or Target Variable

$e_{n,m}$ Evidence *m* found on variable *n*

α Ratio of all possible target variable scenario probabilities

Rouder and Morey (2019) considered that some of the benefits of Bayes' theorem, in addition to calculate conditional probability, was its capability of probability updating on a variable (or node) based on evidence. To determine whether model provides significant evidence, causal inference models rely on Bayes Factors which are the ratio between the observed and unobserved data, as provided below:

$$B.F. = \frac{P(Observed\ Data|H_1)}{P(Observed\ Data|H_0)} = \frac{P(x_1, \dots, x_n|H_1)}{P(x_1, \dots, x_n|H_0)}$$

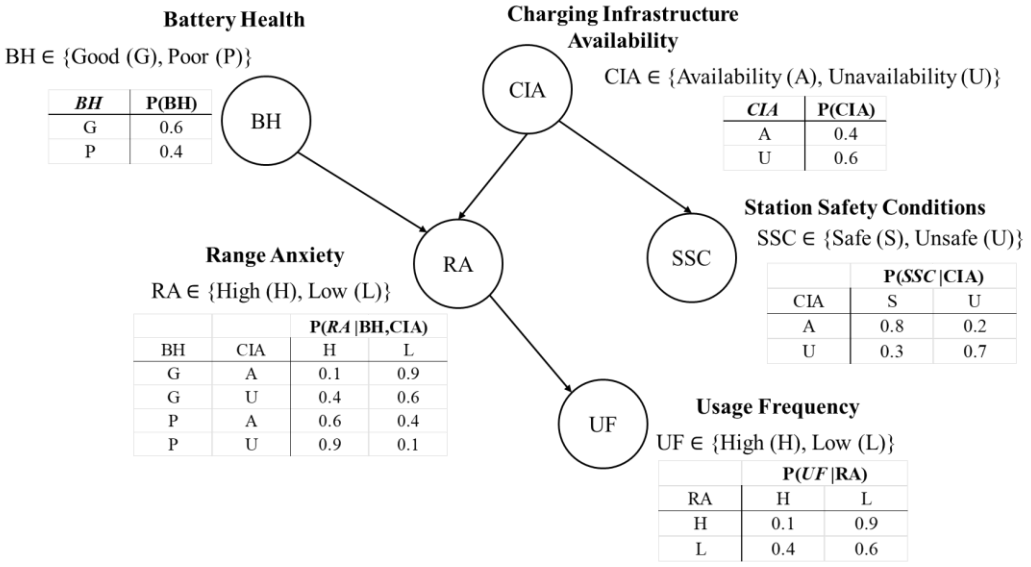
Table 2.4 shows a scheme used for the interpretation of Bayes Factors as suggested and adjusted by Lee and Wagenmakers (2014) from Jeffreys (1961)

Table 2.4. Classification scheme for the Bayes Factors interpretation according to Lee and Wagenmaker (2013)

Bayes Factor	Evidence Category
>100	Extreme Evidence for H_1
30-100	Very strong evidence for H_1
10-30	Strong evidence for H_1
3-10	Moderate evidence for H_1
1-3	Anecdotal evidence for H_1
1	No evidence
0.333-1	Anecdotal evidence for H_0
0.1-0.333	Moderate evidence for H_0
0.03-0.1	Strong evidence for H_0
0.01-0.03	Very strong evidence for H_0
<0.01	Extreme evidence for H_0

Inspired by the findings of Franke and Krems (2013) and Neubauer and Wood (2014), an illustrative example of a BBN (shown in Figure 2.5) is provided, showing the impact of the user context conditions on the EV usage frequency. As it is shown on Figure 2.5, factors such as Battery Health (BH) and Charging Infrastructure Availability (CIA) (both with their MPTs) impact on Range Anxiety (RA), turning it into a Collider node. Finally, RA impacts the EV Usage Frequency (UF), making the last one to turn into a chain node either for the BH→RA→UF or the CIA→RA→UF structure. CIA is an example of a fork node since it is the parent of RA and Station Safety Conditions (SSC), however, the last one does not impact the EV UF.

Each node possesses either an MPT (for CIA and BH nodes) or a CPT (for RA, UF, and SSC nodes), where the data is arbitrary. The way BBN computation works will be shown in the next section.



BH	P(BH)
G	0.6
P	0.4

CIA	P(CIA)
A	0.4
U	0.6

		P(*RA* |BH,CIA)	
BH	CIA	H	L
G	A	0.1	0.9
G	U	0.4	0.6
P	A	0.6	0.4
P	U	0.9	0.1

	P(*SSC* |CIA)	
CIA	S	U
A	0.8	0.2
U	0.3	0.7

	P(*UF* |RA)	
RA	H	L
H	0.1	0.9
L	0.4	0.6

Figure 2.5. A BBN that represents the impact of EV usage by multiple conditions.

In order to understand the BBN mechanics, an example is provided by using the network and data provided on Figure 2.5. It is wanted to know the probability that an EV driver will have a low level of Range anxiety, given the context that involves them. That context involves that the

Charging Infrastructure Availability (CIA) status (or level) is Unavailable (U), the user Usage Frequency (UF) level is High (H), and the charging Station Safety Conditions (SSC) level is Unsafe (U). In other words, our statistical Hypothesis are the following:

$$H_O: RA = H \mid CIA = U, UF = H, SSC = U$$

$$H_1: RA = L \mid CIA = U, UF = H, SSC = U$$

The data, represented by the CPTs and JPTs provided on Figure 2.5, is arbitrary, used only for demonstration purposes.

First, the statement is transformed into a mathematical expression:

$$P(RA = ra|CIA = U, UF = H, SSC = U)$$

It means that all the probabilities coming from the parent nodes in all their categories (or levels) shall be considered in the computation, as shown below

$$P(uf, ra, bh, cia, ssc)$$

The node variable names are in lower case in order to represent that all possible levels per variable node available. By following Decomposition Rule for this specific case and network structure, the equation obtained is the following:

$$\alpha \sum_{bh} P(UF = H|RA = ra) * P(RA = ra|BH = bh, CIA = U) * P(BH = bh) * P(CIA = U)$$

$$* P(SSC = U|CIA = U)$$

It is important to remember that it is necessary to calculate all the target variable levels available to normalize each scenario and obtain the real probability. The scenario when RA level is low.

When $RA = L$

$$P(UF = H|RA = L) * P(CIA = U) * P(SSC = U|CIA = U)$$

$$* \alpha \sum_{bh} P(RA = L|BH = bh, CIA = U) * P(BH = bh)$$

$$P(UF = H|RA = L) * P(CIA = U) * P(SSC = U|CIA = U) * \alpha$$

$$* [P(RA = L|BH = G, CIA = U) * P(BH = G) + P(RA = L|BH = P, CIA = U)$$

$$* P(BH = P)]$$

$$\alpha(0.4)(0.6)(0.7)[(0.6)(0.6) + (0.1)(0.4)]$$

$$\alpha(0.168)[0.36 + 0.04] = \alpha[0.0672]$$

Now, the probability of the RA High level scenario is computed, as shown below:

When $RA = H$

$$P(UF = H|RA = H) * P(CIA = U) * P(SSC = U|CIA = U)$$

$$* \alpha \sum_{bh} P(RA = H|BH = bh, CIA = U) * P(BH = bh)$$

$$P(UF = H|RA = H) * P(CIA = U) * P(SSC = U|CIA = U) * \alpha$$

$$* [P(RA = H|BH = G, CIA = U) * P(BH = G)$$

$$+ P(RA = H|BH = P, CIA = U) * P(BH = P)]$$

$$\alpha(0.1)(0.6)(0.7)[(0.4)(0.6) + (0.9)(0.4)]$$

$$\alpha(0.042)[0.24 + 0.36] = \alpha[0.0252]$$

Once the probabilities of both scenarios are obtained, it is necessary to normalize them, as shown below

$$\alpha = \frac{1}{P(RA = L|CIA = U, UF = H, SSC = U) + P(RA = H|CIA = U, UF = H, SSC = U)}$$

$$\alpha = \frac{1}{0.0672 + 0.0252}$$

$$\alpha = \frac{1}{0.0924}$$

$$\alpha = 10.8225$$

Now that the normalizing constant is obtained, it can be substituted in both probability scenarios, obtaining the following results:

$$P(RA = L|CIA = U, UF = H, SSC = U) = 0.7273$$

$$P(RA = H|CIA = U, UF = H, SSC = U) = 0.2727$$

Since there is only interested on the scenario when the RA Level is Low, the probability that event happens given the provided evidence is 0.7273 or 72.73%. Now, there might be interest in determining the strength of the problem hypothesis, situation that can be handled by the Bayes Factor and its classification table.

$$B.F. = \frac{P(Observed\ Data|RA = H)}{P(Observed\ Data|RA = L)}$$

$$B.F. = \frac{\frac{P(RA = H|Observed\ Data) * P(Observed\ Data)}{P(RA = H)}}{\frac{P(RA = L|Observed\ Data) * P(Observed\ Data)}{P(RA = L)}}$$

$$B.F. = \frac{P(RA = H|Observed\ Data) * P(RA = L)}{P(RA = L|Observed\ Data) * P(RA = H)}$$

$$B.F. = \frac{(0.7273) * (0.48)}{(0.2727) * (0.52)} = 2.4618$$

Based on Table _, it is determined that there is an Anecdotal Evidence for the Alternative Hypothesis (H_1). In other words, there is a weak or inconclusive support that there is a High Level in the Range Anxiety given the context that surrounds the EV driver based on the available data.

As was shown, it is possible to model probabilistic relationships between variables. However, as the number of nodes increases, the number of levels (or categories) increases as well, increasing the complexity of the network and its computation.

Although the number of papers using BBN cannot be compared to those using ML techniques, it still has been used on the methodology of multiple paper with the same goal, determine key factor nodes affecting the target node subject to data limitations.

In the Epidemiology field, Semakula et al (2016) investigated the household factors that were increasing the malaria parasitemia risk among children under five years old. By fusing data from the Malaria Indicator Survey, Demographic Health Survey, and the Acquired ImmunoDeficiency Syndrome (AIDS) indicator survey, the authors could create a BBN model to show the data variables' impact on the disease risk. With the support of the model, it was discovered that the use of boreholes in house as the main drinking water source was increasing the risk by 6.64%, while the use of piped water and rainwater had decreased the probability by 13.02%. With a model accuracy of 86.39%, the authors demonstrated that the BBN model could be used as a logical approach to understanding these interactions between household factors and malaria risk, pointing out the importance of monitoring as more data is available.

In the Natural Disaster Management field, Dlamini (2010) analyzed the contributing factors to the wildfires on Swaziland ecosystems by analyzing data from Terra and Aqua satellites' Moderate Resolution Imaging Spectroradiometer (MODIS) integrated in a Geographic Information System (GIS). After the data was modeled into a BBN, it was found that the climate (mean annual rainfall and wildfire), elevation, and land cover were strong predictors of wildfire occurrence. In addition, Luna (2019) attempted to analyze the public sentiment produced on Twitter toward natural disaster response, focusing on three hurricane scenarios in USA. After the author used Lasso regression with Bootstrapping, it was noticed that the significant tweets amount difference per scenario produced a selection bias. Based on the evidence, the author suggested that the use of a Bayesian causal inference approach, such as the BBN, could help to perform a bias

correction by assigning probabilities to the hypothesis, which stated that the sentiment was coming from dataset variables given that gender was given.

On the forensic anthropology field, Giles et al (2023) explored the use of BBN to determine those taphonomy variables that were affecting the Post-Mortem Interval (PMI) of the deceased people. By comparing both training datasets, one from USA and the other from United Kingdom (UK), it was determined that age, sex, clothing did not have effect in both BBN. In the case of UK, it was observed that BMI, temperature and season did not have effect on its network, while in the USA model those variables still had a limited influence on their model. Those models were validated using testing data and metrics such as accuracy, obtaining a mean posterior probability of 86% and 81% for the USA and UK cases respectively, even though there was a significant difference between both population means due to their sample size.

On the construction management field, Jitwasinkul et al (2016) determined that by extracting the data from a survey of workers in a Thai construction field and modeling it into a BBN, it was possible to determine those key factors that may affect the safe work behavior level, whether it was safe or at risk. Based on this Bayesian network with a accuracy of 79.33%, an improvement in safe work behavior can be obtained by controlling leadership, management commitment, participation, and the perceived behavioral control node, instead of controlling organizational or psychological factors.

The advantages of the BBN are that it can develop models whose nature produces uncertainty (limited knowledge expertise or understanding) (Aalders, 2008; McNair, 2018; Landuyt et al, 2013; Wooldridge, 2003), it considers qualitative and quantitative data on the modeling process (Landuyt et al, 2013), it can be updated as more data becomes available (Landuyt et al, 2013), it can get adapted from the data fusion coming from multiple sources even when there

is a data shortage (Landuyt et al, 2013; McNair, 2018; Wooldridge, 2003). However, disadvantages have been reported such as the overfitting risk due to the high dimension complexity (Wooldridge, 2003; Zhang et al, 2019), the loss of information due to the discretization hypothesis on the variable nodes (Landuyt et al, 2013), and multiple iterations needed to understand the causality between nodes (Zhang et al, 2019).

As it was shown, complexity is a disadvantage in both Frequentist and Bayesian approaches. In order to allow explainability and transparency in the BBN model, it is necessary to use a technique that can filter the nodes based on how they support the target node. That technique is called Naïve-Bayes, and it will be explained on the next section.

2.4.2.2. Naïve-Bayes

Introduced by Shannon (1948) alongside with the Information Theory, Mutual Information allows to quantify the amount of information obtained regarding a variable (Y) through another random variable (X). The formula is the one below:

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log_2 \frac{p(x|y)}{p(x)p(y)}$$

In the context of BBN, instead on focusing on feature selection ML techniques (correlation, p-value, Principal Component Analysis) to determine the connection between an explanatory variable and a response variable, Naïve-Bayes focuses on providing a Feature Importance when comparing with multiple variables to determine which ones really have an impact on the target variable (Bayesia S.A.S., 2024). In terms of accuracy and precision, Bangroo et al (2023) considered that this technique outperformed ML techniques such as Lasso and Elastic Net.

The advantages of this approach are that it is easy to implement and fast in training since it requires less training data and it performs well with categorical predictor variables instead of numerical variables (Sarang et al, 2023). However, the discretization is required to initiate the

algorithm, losing information in the process (Kotu & Deshpande, 2019), and it must hold the assumption that all the categorical variables are independent from each other, a phenomenon that is impossible on real world scenarios (Romadhon & Kurniawan, 2021; Kotu & Deshpande, 2019)

2.5 RESEARCH QUESTIONS

With this motivation and these concepts in mind, we identified the following research question emerges as the following:

- What insights could be obtained from the exploration of social media that could support our understanding of public perception towards Electric Vehicle Adoption (EVA)?

  Which can be answered by answering the following support questions:

  - What are the primary socio-economic variables that might influence public sentiment positively or negatively regarding EVs?
  - What sentiment do Twitter users in areas where EVs are less common (compared to U.S. hub cities) hold towards EVs?

Chapter 3: Methodology

3.1 NSF-ERC ASPIRE BACKGROUND

The National Science Foundation (NSF) Engineering Research Center (ERC) Advancing Sustainability through Powered Infrastructure for Roadway Electrification (ASPIRE) aims to eliminate business, technical, and social barriers that limit access to Electric Vehicles (EV) in the United States in collaboration with strategic partnerships from industry, government, universities and community (Utah State University, 2024). The center projects are five: Charging Stations of the Future, Electrified Roadways, Systems of Systems and Learning and Engagement, being the third one the focus on this work.

This project focuses on analyzing those factors that affect Electric Vehicles Adoption (EVA) rates, power grid operations and electric markets (NSF-ERC ASPIRE, 2023). The first point is the reason for this work: understanding what variables are really impacting on the public perception regarding Electric Vehicle Adoption (EVA) at a local level.

3.2 STUDY CITY PROFILE

The following USA cities were selected to implement this methodology based on the idea that they still have a young EVA and their connection to the NSF ERC ASPIRE. The selected cities were the following:

- **Salt Lake City, Utah** It is the home of Utah State University (USU), the main NSF ERC ASPIRE campus. At the EV Charging stations National ranking (U.S. Department of Energy, 2022), the city is located at the 41st position with 151 charging stations.
- **Indianapolis, Indiana.** Although the city does not possess an NSF ERC ASPIRE campus [being the closest one located in Purdue University (West Lafayette, IN)], it still being

important for the ERC. The collaboration between the Indiana State Government and Purdue University on a pilot program by designing and implementing multiple Wireless EV charging infrastructures throughout the city has been critical to see the reaction of the people regarding the arrival of these new technologies (Indiana Department of Transportation, n.d; Sullivan, 2022; Pierce, 2023). At the EV Charging stations National ranking (U.S. Department of Energy, 2022), the city is located at the 107th position with 60 charging stations.

- **El Paso, Texas.** It is the home of the University of Texas at El Paso (UTEP), main NSF ERC ASPIRE campus. At the EV Charging stations National ranking (U.S. Department of Energy, 2022), the city is located at the 41st position with 151 charging stations.

3.3. MODELING FRAMEWORK

To deliver the methodology of this work, a modeling framework is provided adapted from Gutierrez Araiza et al (2024). It was considered to divide it into two phases: Descriptive and Diagnostic Analytics (Stage 1 through 4) and Prescriptive Analytics (Stage 5).

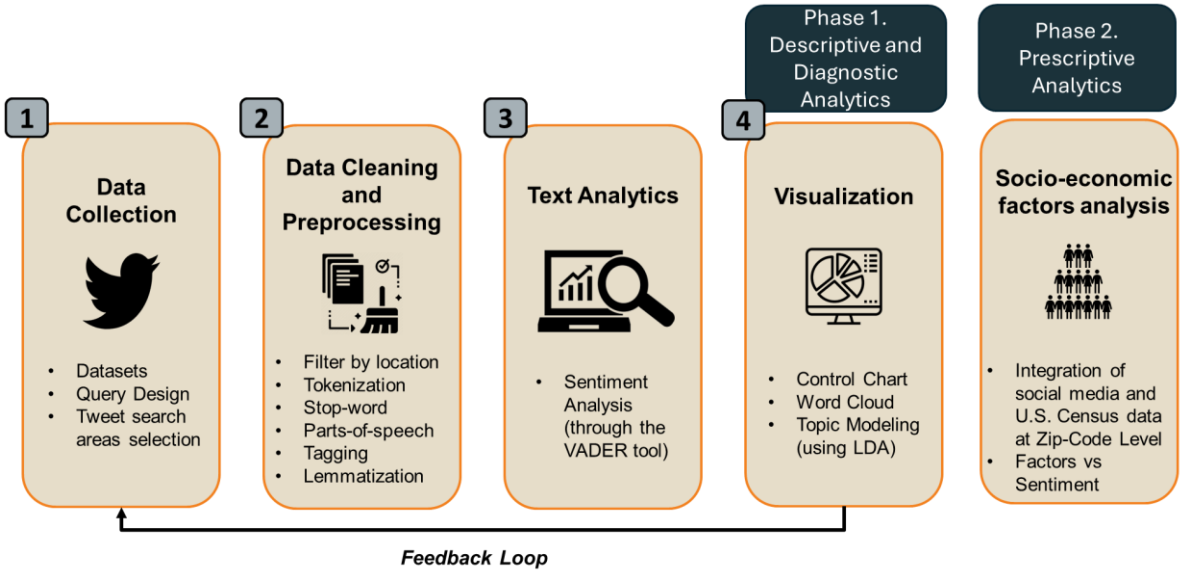


Figure 3.1. Main Modeling Framework

3.3.1. Phase 1, Stage 1: Data Collection

3.3.1.1 Datasets

- **Twitter Social Media Data.** Twitter stands out as a pivotal platform for accessing news and gaining insights into current events that significantly impact their respective contexts (Mitchell et al., 2022). Notably, it ranks among the most widely used social network services in the United States (Odabas, 2022). As the dataset extraction occurred in December 2022, the Twitter API Academic Research Access (Twitter, 2022) provided several crucial fields per social media post. These encompassed the tweet creation time, user-generated text, the source (whether from a mobile device or another method), geographic coordinates, full location name, place type, and username. The inclusion of these fields proves essential in situating the user's message within the broader context where the idea is developed. Once the sentiment score per post is obtained, it will be considered as the response variable in our dataset.
- **U.S. Charging Stations Dataset.** The strategic analysis of charging station locations across the United States serves as a crucial step in determining focal points for this study. It is assumed that areas with a higher concentration of electric charging stations correlate with a more significant number of Electric Vehicle (EV) users and, consequently, an increased volume of tweets on the topic. The data, sourced from the Alternative Fuels Data Center provided by the U.S. Department of Energy (U.S. Department of Energy, 2022), enables users to tailor their information searches based on criteria such as location (Canada, United States, or both) and fuel type (Biodiesel, Compressed Natural Gas (CNG), Electric (ELEC), Ethanol (E85), Hydrogen, Liquefied Natural Gas (LNG), and Propane (LPG). For this work, the focus was narrowed to U.S.-based charging stations and specifically targeted

the fuel type ELEC. The database presented the information in Comma Separated Value (CSV) format. As of November 10, 2022, the dataset recorded 51,486 electric charging stations across the U.S. The primary interest attributes within the dataset were City and state, as these would serve as crucial indicators for the locations of electric charging stations. In essence, the charging station dataset provides a comprehensive overview of the EV infrastructure expansion in the U.S., playing a pivotal role in guiding the tweet search areas selection. From there, a variable will be extracted to determine whether the number of charging stations at the city is higher or equal to the mean per city at the state level.

- **2016-2022 American Community Census (ACS) – 5 Year U.S. Census Bureau.** Since its foundation in 1902, the U.S. Census Bureau mission has been collecting economic, demographic and societal data from its country to support the federal government decision-making process in how to distribute funds throughout local communities as equitable as possible. The data collection, either for the Decennial Census or 5 Year American Community Surveys, used to be only either by mail, telephone or in-person visits. Nevertheless, the 2020 Decennial Census consolidated its digitization process by providing U.S. residents with the flexibility to respond to the questionnaire not only on paper but also through phone calls and various mobile devices, including computers, laptops, and smartphones. An additional advantage offered was the ability to switch the questionnaire language, providing added convenience for participants and encouraging broader participation (U.S. Census Bureau, 2018). The census contains a vast data to be analyzed, however it will be limited to the estimates that are updated every year, which come from the American Community Survey (ACS) 5-year estimates. By extracting data from the database, the following topics were considered:

Table 3.1. USA Census Bureau Explanatory Variables to be used on the methodology.

Area	Sub-Field	Total Variables
Education	• Educational Attainment (S1501)	3
Employment	• Employment Status (S2301) • Means of transportation to Work by Travel time to work (B08134), • Travel time to work (B08303).	18
Families and Living Arrangements	• Households and Families (S1101) • Marital Status (S1201)	4
Health	• Disability Characteristics (S1810)	1
Housing	• Financial Characteristics (S2503) • Financial Characteristics for Housing Units with a Mortgage (S2506) • Occupancy Characteristics (S2501) • Physical Housing Characteristics for occupied units (S2504).	28
Income and Poverty	• Income in the last 12 months (S1901)	10
Population and People	• Age and Sex (S0101) • Language spoken at home (S1601), • Sex by Age (B01001) • Type of Computer and Internet subscription (S2801)	13
Race and Ethnicity	• Racial group (B02001).	8

A full list of each variable used is provided on Appendix 1.

3.3.1.2 Tweets Collection

- **Query Design.** A precise query design is crucial for effectively limiting our search area and streamlining data preparation. Leveraging Academic Research credentials enables us to progressively narrow our scope, focusing on posts that align with our specific interests.

- **Language Specification:** Our initial filter ensures that the posts are in English, mitigating the inclusion of content from other languages. This is particularly relevant for cities like El Paso, TX, where the use of both English and Spanish is shared, either separately or within the same message due to its proximity to the border with Mexico.
- **Exclusion of Retweets:** To enhance the relevance of the collected data, retweets were excluded, as they often lack the personal experience conveyed by the original user (Haman, 2020).
- **Geographical Delimitation**: Our search areas are delineated as circular regions with a radius of 25 miles. This aligns with the maximum radius allowed by the Twitter API for each city. The center of each circle is chosen arbitrarily, ensuring comprehensive coverage of the respective city (Twitter, 2022).

In the final query design (Table 3.2.), keywords were strategically selected encompassing not only Electric Vehicles (EVs) and charging stations but also included terms associated with leading companies involved in either vehicle sales or the infrastructure manufacturing process, as shown in Table 3.2.

Table 3.2. Keywords list regarding to the Electric Vehicles and their infrastructure

Section	Words sample
Primary Words	*ev, electric charging, electricvehicles, electric vehicle, evcharging, ev charging, electriccar, electric car, charging station*
EV Brands	Tesla, TSLA, teslamodel, karmaautomotive, lucid, lcid, LucidMotors, LoveLucid, FaradayFuture, FFIE, rivian, RIVN, RideWithLordstown, nikola (but removing *Nikola Tesla), canoo*
EV Infrastructure Brands	*ChargePoint, SemaConnect, Electrify America, EV Connect, EvoCharge, EVSE (Electric Vehicle Supply Equipment), Blink,GreenLots*
EV Key Words	*ev, electric vehicle, electric, evcharging, ev charging, electriccar, electric car, charging station*

To synthetize the filtering process, a diagram is attached as it is shown on Figure 3.2.

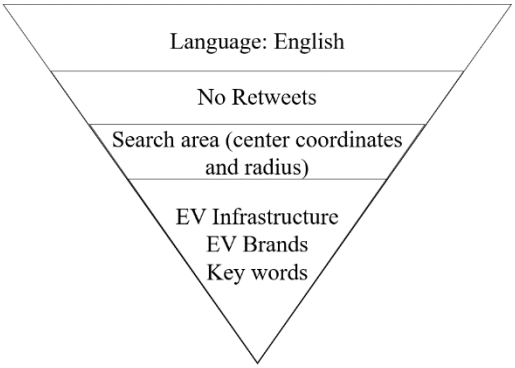


Figure 3.2. Twitter Query Design

In addition, it was decided that the study time to be analyzed was from January 1, 2016, to September 30, 2022.

- **Gathering Tweets.** Utilizing the Python 3.9 programming language, the capabilities of the Tweepy package were harnessed to establish a connection with the Twitter API, leveraging our credentials (Tweepy). The acquired data, presented in JavaScript Object Notation (JSON) format, is considered unstructured. To enhance usability, the code was designed to transform this data into a structured, tabular format systematically. Given that the API provides results in pages, each containing up to 500 social media posts, automation was implemented within the code to efficiently retrieve all relevant data within a short timeframe. The output is then formatted and saved in Comma Separated Value (CSV) format. In the final step all pages were consolidated into a unified file for each city.

3.3.2. Phase 2, Stage 2: Data Cleaning and Preprocessing

Following the collection of social media data for each city, our focus shifted to the crucial task of cleaning and preprocessing to ensure accuracy in subsequent analyses. The initial step involved filtering by location, allowing us to narrow our attention to the specific cities of interest. Some instances posed challenges, such as locations named after popular community places rather than the city itself or only the state where they were published (probably, those post were generated when the user was on a highway; it is impossible to determine if they were heading or leaving the city); these cases were meticulously analyzed and substituted with the correct city names.

With the aid of the Natural Language Toolkit (NLTK) Package, the process was streamlined, automating several steps to save time. The tokenization process, a critical step, involved breaking down each text into words, treating each word as a distinct token. Subsequently, tokens comprising punctuation, exclamation and admiration symbols, emoticons, and URLs, were eliminated

Once the updated token list was obtained, the next step involved the removal of stop-words—those lacking sentiment, such as connectors, pronouns, or prepositions. Following this, a Parts-of-Speech (POS) analysis categorized each token into its respective POS category, such as noun, adjective, or comparative. Finally, lemmatization was employed to remove prefixes from each word, transforming them into singular or infinitive forms (e.g., "books" -> "book" or "studies" -> "study").

In summary, this preprocessing stage plays a pivotal role in obtaining accurate results for Topic Modeling and sentiment analysis.

3.3.3. Phase 1, Stage 3: Text Analysis

With our prepared samples in hand, the next step involves conducting Sentiment Analysis. For this task, the Valence Aware Dictionary for sEntiment Reasoning (VADER) algorithm was implemented, seamlessly integrated into the NLTK package. Known for its sensitivity to opinion polarities, the VADER algorithm has demonstrated accuracy and viability in previous studies (Hutto & Gilbert, 2014; Ruan & Lv, 2022; Nandurkar et al, 2023).

Given the anticipated variations in the number of social media posts due to population differences among cities, the Bootstrapping methodology will be employed to ensure sample variability and significance.

The analysis will categorize each text (social media post) into a specific polarity: positive, negative, or neutral.

3.3.4. Phase 1, Stage 4: Visualization

Subsequently, this stage consists of visualizing the analysis and results of the texts through various strategies. To discern the most frequently used words related to EV adoption, the Word-cloud technique was employed using unigrams and bigrams. This preliminary analysis provides insights that pave the way for implementing the Topic Modeling technique, specifically through Latent Dirichlet Allocation (Blei et al, 2003). This technique clusters words into topics based on their probability index to belong to that topic and it was used by Banik (2023) to evaluate the maternal patient experience during COVID-19 after he processed the gathered social media posts referring to the topic.

Moving forward, control charts are utilized to depict how sentiment has evolved over time and how many Twitter users were talking about the topic, segmenting the data into months based on the dates of social media posts. To assess potential influences on public perception, the sentiment trends were compared with news releases on similar days. This strategy was adopted by

Luna et al (2022) as a method to explain the sentiment variation through the time, to explain the sentiment peaks with news reports.

Finally, since one objective of the study is to explore the sentiment of social media users by binary gender, the next step involves assigning gender attributes to each tweet. The R package "*gender*" (Mullen et al, 2018) was utilized to compare Twitter usernames with the U.S. Social Security Administration (SSA) names database. If a username matched a name in the government database, the corresponding gender registered by the SSA was assigned to the user. However, not all social media users use their real names, often opting for pseudonyms or nicknames. To address this challenge, the user's profile picture was observed, and gender was manually assigned based on the image. If a real picture was not available, the tweet was discarded.

3.3.5. Phase 2: Prescriptive Analytics - Socio-Economic Factors Analysis

In order to explain the details of this Phase 2 work, a secondary framework is provided on Figure 3.3.

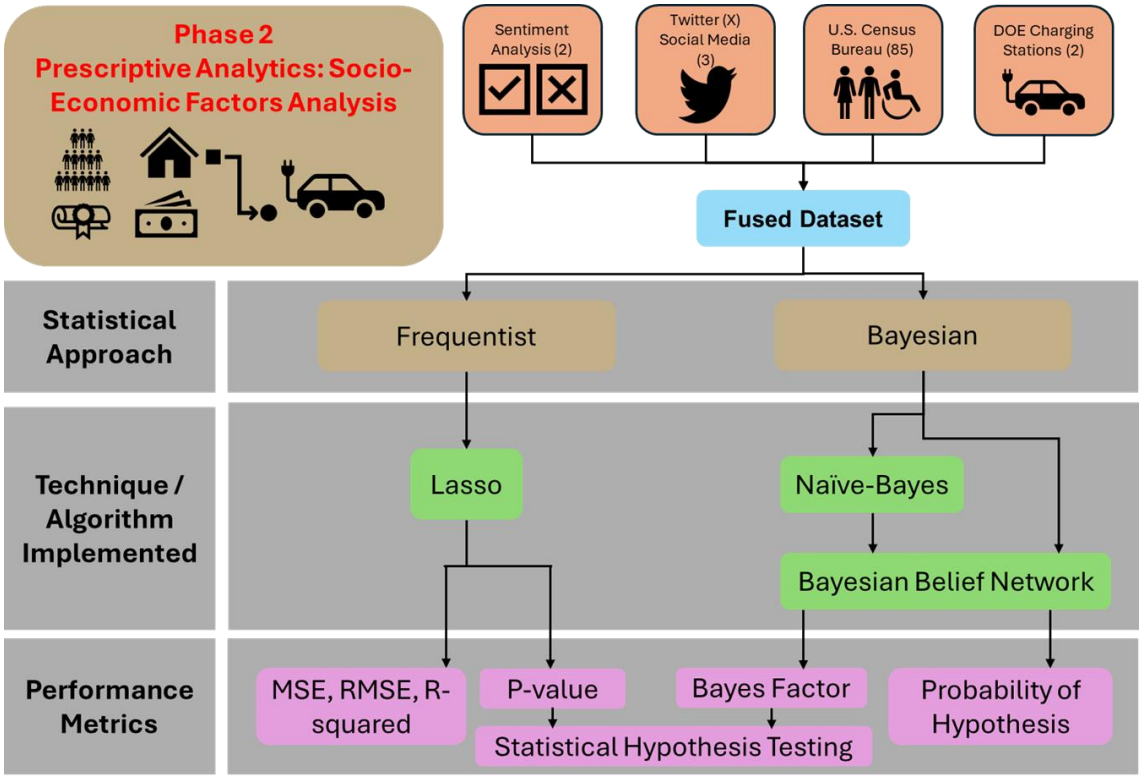


Figure 3.3. Main Modeling Framework focused on the Phase 2 Process.

Once the Phase 1 tasks are completed, Phase 2 starts by fusing all the datasets that were mentioned in Section 3.3.1, adding the sentiment score column obtained on Phase 1 Stage 3 into the fused dataset. Despite of that the US Census ACS-5 Year estimates were obtained by year by Zip Code, its data will be aggregated by obtaining the average of all the Zip Code areas per city per year under the assumption that the social media user was surrounded by a similar context throughout the city.

Once the dataset is consolidated, the dataset will be analyzed from the **Frequentist Approach** by using Lasso as the Feature Selection technique. Before the modeling process starts, it is necessary to decompose the data from categorical variable to binary variable, which will be the case of the *City* [which contains three categories (El Paso, Indianapolis and Salt Lake City)] and *Year* [which contains 7 categories (from 2016 to 2022)] predictor variables, obtaining as result 10 binary variables instead of two categorical variables. Once that step is completed, the dataset was divided into a matrix that contain all the predictor variables and a vector that contains the response variable (sentiment score).

It is important to note that multicollinearity can affect the performance metrics, as it was shown on the Literature review. For this reason, two pathways were considered as follows:

- **Pathway 1: Using all the predictor variables.**
- **Pathway 2: Removing variables with multicollinearity issues.** This pathway considers removing those predictor variables that have a correlation higher than 0.7 between each other, allowing to assess the effect of the remaining predictor variables on the response variable (Sentiment score).

The analysis levels will be the following: Three-city, Indianapolis, Salt Lake City and El Paso. With exception of the Three-city analysis, the *City* binary variables are removed from the dataset.

Since it is assumed that the data will be biased due to the number of social media posts per location, a Bootstrapping technique will be necessary when obtaining the outstanding variables per three-city analysis and per city analysis to reduce the sample selection bias in the final modeling. Therefore, a for loop is introduced, which consists of repeating the previous steps for 1000 iterations, where the 10% of the dataset rows will be randomly selected with replacement for the next iteration to obtain their outstanding variables and performance metrics are saved for analysis.

It is considered that during the iteration, the dataset is divided into the training set and testing set, having data size proportional to the 70% and 30% respectively, ensuring a proper modeling and testing of the iteration model. In addition, the dataset will be standardized to ensure that the impact of outliers is minimized on the performance metrics by using the Python *Standard Scaler* package.

Since cross validation will be used to determine the hyperparameters to be used on the models, the number of validations were kept by its default of 10. Once the hyperparameters are available, it will be possible to run the Frequentist Machine Learning Algorithm (Lasso), being used the Machine Learning Python package, *scikit-learn*, to run the models. Based on the model training and testing, the performance metrics to be obtained from the model are MSE, RMSE, R^2,

and the outstanding variables, being saved in a data frame that will be updated per iteration. With the previous step, a new iteration starts.

Once the iterations of the Bootstrapping method end, it is counted the total of occurrences that each variable was outstanding throughout the iterations, being sorted in a descending way and with its cumulative sum. The variables that sum the 70% of occurrences will be the ones to be modeled in an OLS analysis. Based on the model training and testing, the performance metrics to be obtained from the Ordinary Least Squares (OLS) method are MSE, RMSE, R^2, and the p-value of the outstanding variables in the model. For this work, it is considered that a p-value less or equal to 0.05 was statistically significant.

The flowchart that summarizes the process is given in Figure 3.4.

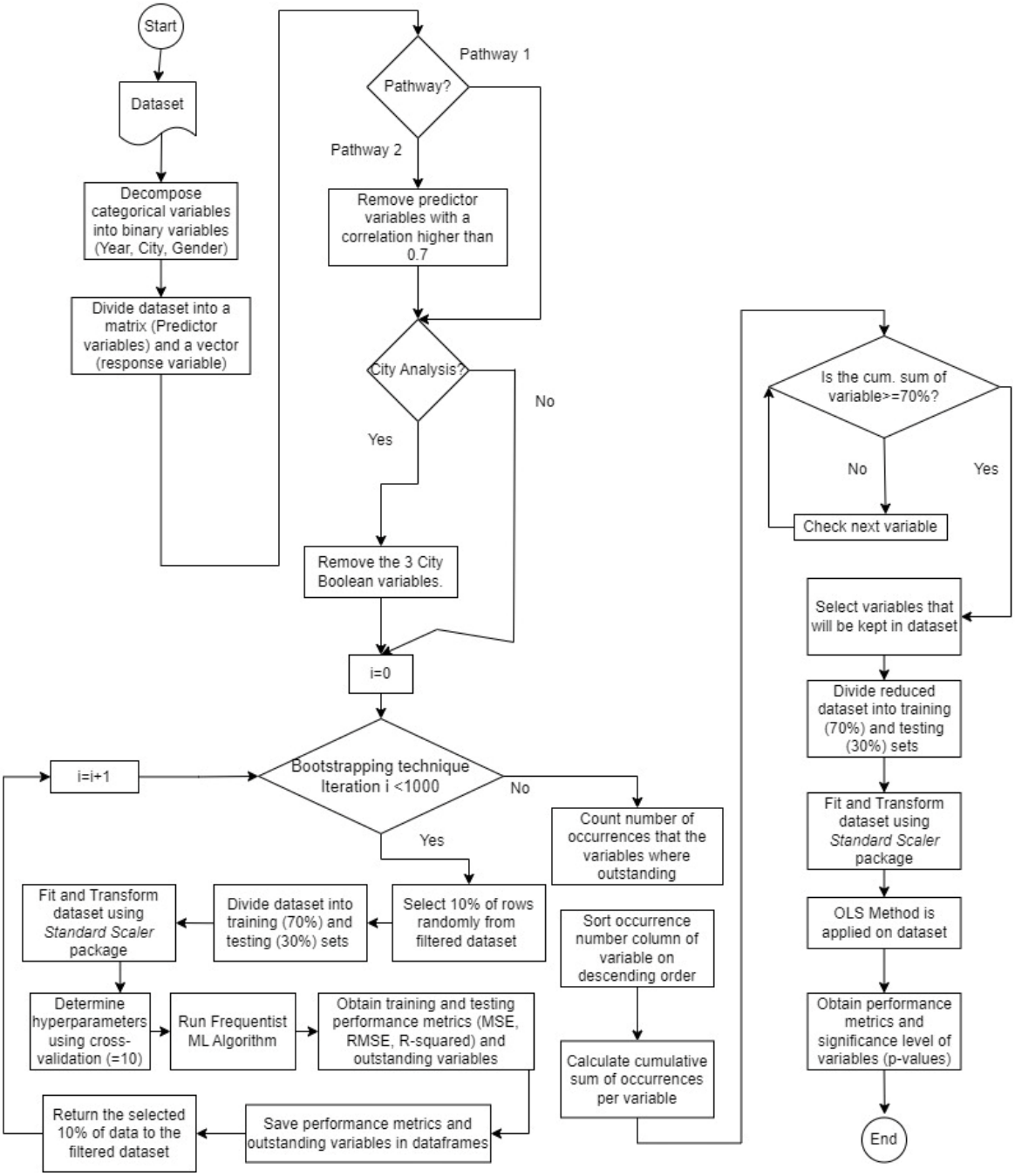


Figure 3.4. Steps to be followed to obtain a Frequentist Machine Learning model

Thereafter, the dataset from the **Bayesian Approach** by using Bayesian Belief Network to perform the causal inference will be analyzed, where our response variable (or target value) will be the Sentiment Category obtained from the Sentiment Analysis. It was considered that all the sentiment scores lower than 0 would be considered negative, while all of those with a sentiment score higher than zero would be considered as positive, otherwise It would be considered as

neutral. However, for this approach two paths will be analyzed, one where the Naïve-Bayes is implemented as Feature Importance tool before the BBN design, and other where the dataset is analyzed directly by BBN (as shown on Figure 3.3). The performance metrics to be implemented are the Bayes Factors and the Probability that the Hypothesis can happen given the current data and main variable nodes that can significantly improve the EVA perception. The statistical software *BayesiaLab version 14.1* will be used to process the networks.

However, it is expected that the Census variables might not have a strong connection with the sentiment since those attributes represent the context attributes rather than the individual user attributes, as reported by Luna (2019). Therefore, it is proposed the following BBN to represent those individual attributes, as shown on Figure 3.5.

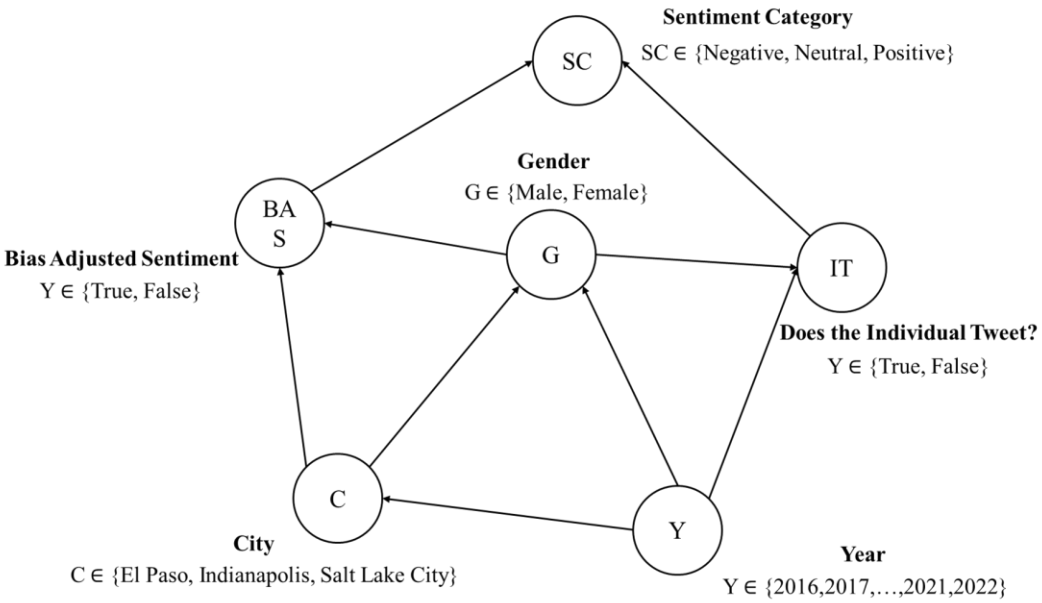


Figure 3.5. Bayesian Belief Network for Public Perception of EVA

The model is developed under the assumption that a random social media user is generated to broadcast an opinion regarding EVA on social media. The model is initialized on the node *Year*, where the individual has a probability distribution of being generated in a year from 2016 to 2022. The value taken on the node *Year* impacts, as a fork, on three nodes: *City*, *Gender* and the *Does the Individual Tweet?* While the first two nodes, *City* and *Gender*, come from the observable data,

the node *Does the Individual Tweet?* acts as a MAR mechanism since it connects the observable predictor data with the target node *Sentiment Category*, which relies also on the node *Bias Adjusted Sentiment*. The node *Bias Adjusted Sentiment* is a collider of the nodes *Gender* and *City* since it neutralizes the bias generated by the sample selection size per city and gender. With the nodes *Bias Adjusted Sentiment* and *Does the Individual Tweet?* it is possible to obtain the probability that a random social media user discussing about EVA will have a specific sentiment probability, weather positive, neutral or negative. The CPT of the last nodes will be obtained as a Parameter estimation, which is calculated by the *BayesiaLab* software. The Causal inference hypothesis tests will be the following:

$$H_o = P\left(S = \bar{h} \middle| Y = i, C = j, G = k\right)$$

$$H_1 = P(S = h | Y = i, C = j, G = k)$$

The alternative hypothesis H_1 considers the probability that the polarity *Sentiment* node value h (from the list provided in Figure 3.5) given that the *Year* node value i, the *City* node value j, and the user *Gender* node value k are evidence of it. The null hypothesis H_o considers that all the other polarity sentiment will occur given the same observed variable values. For example, if the *Sentiment* value for the alternative hypothesis is Positive, then the null hypothesis considers the probabilities where the sentiment was not positive (negative and neutral).

3.3.6. Software Implemented

- **Python.** The programming language is used to connect the user with the Application Programming Interface (API) of the data sources, Twitter and U.S. Census Bureau, to extract the datasets and save them in Comma-Separated Values (CSV) files for upcoming analysis. It is used also to run the VADER method located as function of the Natural Language Toolkit (NLTK) package (Bird et al, 2009). In addition, the *scikit-learn* package (Pedregosa et al, 2011) will be required to run the Machine Learning Bayesian statistical

methods: Ridge, Lasso, and Elastic Net. The *pandas* package is also used for data manipulation and standardization (McKinney, 2010).

- **R.** This second programming language will be used to run the *gender* package (Mullen et all, 2022)
- **Microsoft Excel.** The use of its Visual Basic Application (VBA) allows to minimize the processing time on integrating the US Census Bureau datasets into a consolidated dataset.
- **Minitab® Statistical Software.** It is used to perform the Bootstrapping resampling method in the city samples and to obtain the average monthly sentiment monitoring charts (or control charts) per city (Minitab LLC, 2023)
- **BeyesiaLab.** The software is used to run the Bayesian statistical methods: Naïve-Bayes and the modeling of the Bayesian Belief Network and its calculations (Bayesia S.A.S. 2024).

Chapter 4: Results and Discussion

4.1 Descriptive Analytics

This section focuses on providing a clear and comprehensive overview of historical data by identifying patterns and presenting key metrics. On this analysis,

4.1.1 Social Media Posts Filtration

The social media posts were extracted and stored on multiple CSV files and joined per city. They were passed through three phases, as shown on Table 4.1.

Table 4.1. Social media post filtration process results

Stage/ City	**Indianapolis, IN**	**Salt Lake City, UT**	**El Paso, TX**	**Total**
Tweets collection from the Twitter Application Programming Interface (API)	3444	1399	384	5227
Tweets filtered by location	2682	410	371	3463
Tweets filtered by the VADER (Valance Aware Dictionary for Sentiment Reasoning) tool	1820	156	307	2283

As it is shown on Table 4.1, two thousand two hundred eighty-three (2283) social media posts received a numerical value, considered the sentiment score, because of the filtration process. In other words, 66.29% of the social media posts collected from the Twitter API were kept for the data analysis in Phase 1, as shown in the Modeling Framework of Figure 3.1.

4.1.2. Sentiment Polarity Distribution

A brief description of how the EVA polarity is distributed throughout the cities helps diagnose the three cities first.

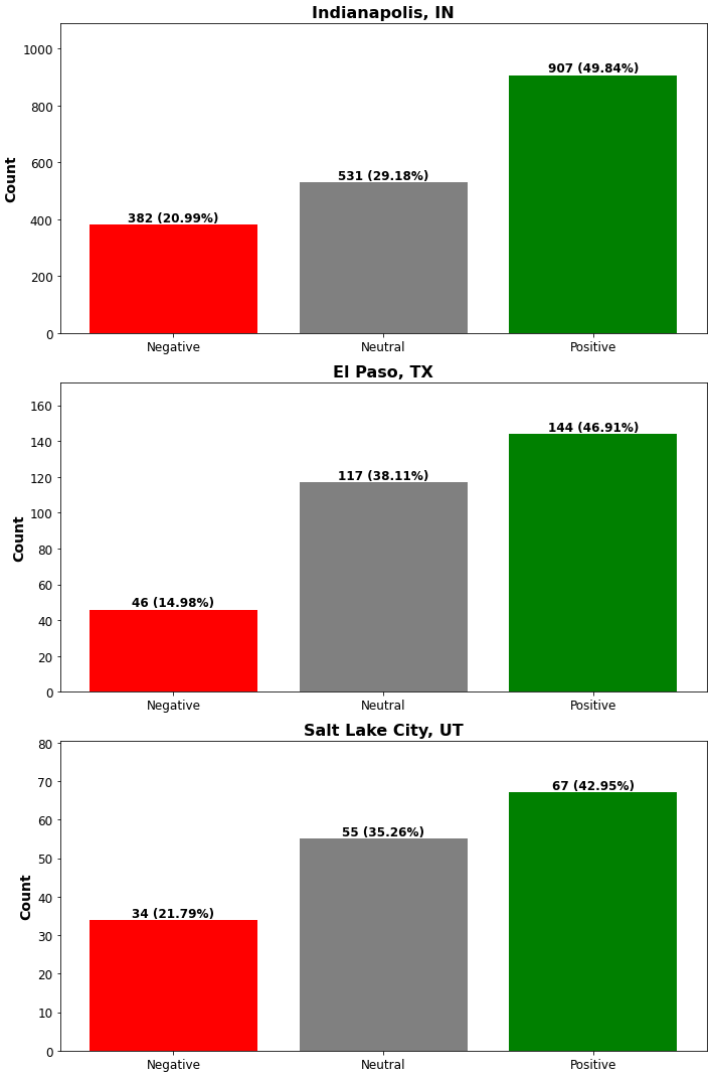


Figure 4.1. Perception Polarity distribution from January 2016 to September 2022 in the study cities

As it was shown on Figure 4.1, it was revealed that despite the sample differences between the cities, the polarity perception trend towards EVA remained the same. The social media posts in the three cities indicated that at least 42% of them were positive regarding EVA, while at least 29% were neutral in each city. Lozada-Medellin et al (2023) noted on their study that the EVA perception was positive; however, some doubts and concerns regarding the EVA implementation were still present. These plots reveal that most users in the three cities were either unconcerned or interested in EVA, expressing either a neutral or positive sentiment.

4.1.3 Sentiment Average per city

Since the sample social media post size differed significantly throughout the three cities, the mean stabilization was implemented to obtain the actual sentiment average per city. The Bootstrap method was implemented by using Minitab® Statistical Software, having 1000 iterations as a reference, to normalize the mean per city. The results are shown in Table 4.2.

Table 4.2. Statistical mean comparison between the original and the one with Bootstrap Method applied for the three cities.

		Original		**After Bootstrap Method was applied**	
City	**Sample Size**	Mean	Std. Deviation	Mean	Std. Deviation
Indianapolis (IN)	1820	0.1306	0.3677	0.1343	0.0186
El Paso (TX)	307	0.1476	0.3262	0.1248	0.0302
Salt Lake City (UT)	156	0.1339	0.3395	0.1611	0.0186

As shown on Table 4.2., after Bootstrap was applied, it was confirmed that the people's sentiment regarding EVA was positive on average, as shown in Figure 3.1. However, it cannot be stated that all the time the sentiment was positive on similar levels; therefore, statistical monitoring is needed.

4.1.4 Unigram word cloud

Word clouds are essential visualization tools since they help us to have a general context of the standard terms (or unigrams) used in the discussion. From the largest to the smallest word sizes, it becomes evident how frequently particular words appear in the analyzed texts. This visualization technique is especially useful in our focus on social media posts, as it highlights the key themes and trends present in the content.

The first case shows the Indianapolis unigram word cloud, which is depicted in Figure 4.2

Figure 4.2 Unigram Word cloud of the 2016-2022 Social Media posts on Indianapolis, IN

The second case shows the El Paso unigram word cloud, which is depicted in Figure 4.3

Figure 4.3 Unigram Word cloud of the 2016-2022 Social Media posts in El Paso, TX

The third case shows the Salt Lake City unigram word cloud, which is depicted in Figure 4.4

Figure 4.4 Unigram Word cloud of the 2016-2022 Social Media posts in Salt Lake City, UT

To improve the unigram visibility, a top 10 words (unigrams) used on the posts per city is provided, as shown in Table 4.3.

Table 4.3. Top 10 Unigram words of social media posts

#	Indianapolis, IN		El Paso, TX		Salt Lake City, UT	
	Unigram	**Frequency**	**Unigram**	**Frequency**	**Unigram**	**Frequency**
1	EV	570	Car	51	EV	233
2	Car	410	EV	40	Car	178
3	Will	266	One	27	One	107
4	TSLA	224	Now	26	Will	94
5	One	211	Want	24	New	81
6	New	204	El	21	Electric	76
7	M	171	Got	20	Make	74
8	Buy	204	New	19	TSLA	68
9	Electric	149	Electric	19	Thank	67
10	Want	138	love	18	year	66

As it was shown on Table 4.3, the unigrams, *Electric, EV, New,* and *One* prevailed on the posts of the three cities. Meanwhile, the words *TSLA* and *Want* were in the discussion of two cities only, Indianapolis and Salt Lake City. The word *TSLA* is recalled since it is the Nasdaq Stock Market acronym for the since it is the Nasdaq Stock Market acronym for the EV automotive company, Tesla, Inc. Using unigrams helps us to a general perspective; however, it might limit the dataset visualization. Therefore, the use of bigram word clouds was critical to have a closer look at the post content.

4.1.5 Bigram Word Cloud

The first case shows the Indianapolis bigram word cloud, which is depicted in Figure 4.5

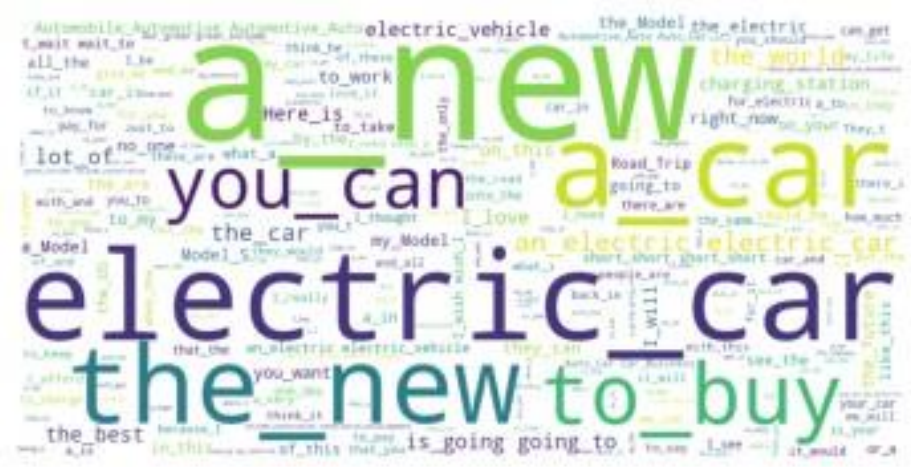


Figure 4.5 2016-2022 Bigram Wordcloud of the 2016-2022 Social Media posts on Indianapolis, IN

The second case shows the El Paso bigram wordcloud, which is depicted in Figure 4.6

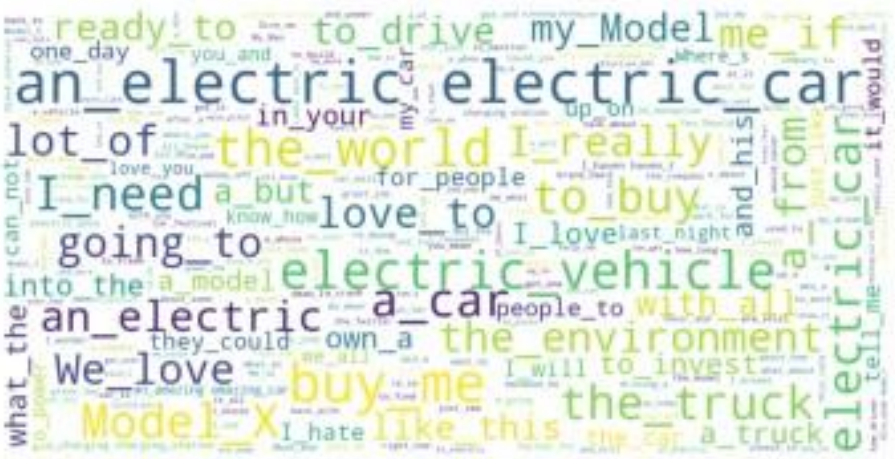


Figure 4.6. 2016-2022 Bigram Wordcloud of the 2016-2022 Social Media posts in El Paso, TX

The third case shows the Salt Lake City bigram word cloud, which is depicted in Figure 4.6.

Figure 4.6. 2016-2022 Social Media posts Word cloud on Indianapolis, IN

To improve the unigram visibility, the top 10 pair words (bigrams) is provided, as shown in Table 4.4.

Table 4.4. Top 10 Bigram words of social media posts

#	Indianapolis, IN		El Paso, TX		Salt Lake City, UT	
	Bigram	Frequency	Bigram	Frequency	Bigram	Frequency
1	A_new	35	To_see	8	An_electric	13
2	Electric_car	34	An_electric	8	Electric_car	10
3	A_car	31	Is_a	7	All_the	9
4	The_new	30	When_I	7	Going_to	9
5	To_buy	29	Buy_me	7	Salt_Lake	8
6	An_electric	28	I_just	7	Electric_vehicle	7
7	You_can	28	This_Is	7	To_see	7
8	The_world	25	I_want	6	To_charge	7
9	The_cart	22	The_world	6	The_new	6
10	Lot_of	21	Want_to	6	You_can	6

As it is shown in Table 4.4, the bigram *An_Electric* was the only one that prevailed on the posts of the three cities. Meanwhile, the bigrams *Electric_car*, *the_new*, *the world*, and *you_can* once again prevailed on the discussion of two cities only, Indianapolis and Salt Lake City. Word clouds are a useful tool to visualize quickly the terms being discussed on social media; however, analyzing topics inside each city might confirm how evolved the EVA discussion is, as it is presented in the following section.

4.1.6 Social Media Topic Modeling

Using the LDA probabilistic model to uncover latent topics within our text corpus could support seeing the current conversation topics. The number of topics was limited to only 3 due to the sample size per city, obtaining the results that are shown in Figure 4.7.

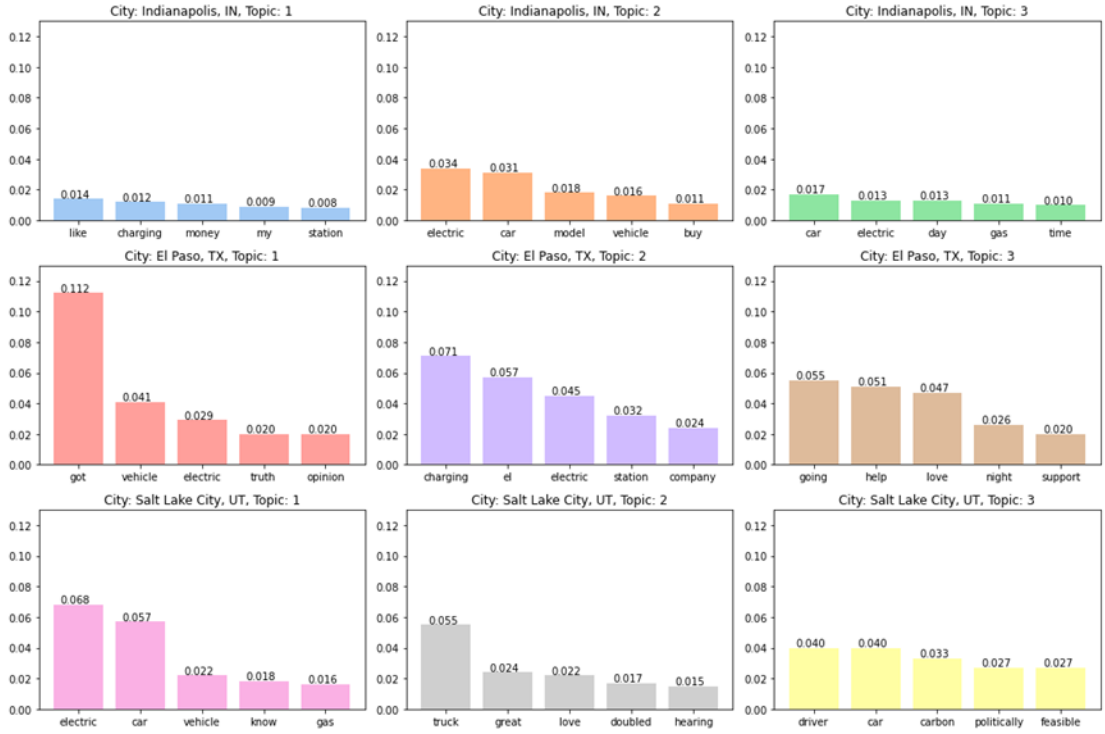


Figure 4.7. Top three Topics discussed in social media related to EVA by city

As it was shown on Figure 4.7, a set of words may have a chance to be on the conversation. For Indianapolis' topics, it was detected the probability of each word in the three topics is lower than 0.03 or 3% but higher than 0.008 or 0.8%. For El Paso's topics, the probability of each word in the three topics is lower than 0.112 or 11.2% but higher than 0.02 or 2%, having the highest probabilities from the three cities. Finally, for SLC's topics, the probability of each word in the three topics is lower than 0.068 or 6.8% but higher than 0.015 or 1.5%. It is possible to notice that the words allowed to search on the corpus, based on the test from the social media posts, how those words were used and interpret them. Table 4.5 provides an explanation of the possible topics that the LDA provided.

Table 4.5. Explained Topics obtained from LDA Algorithm application to social media data

City	Topic	Post Explanation
Indianapolis, IN	1: Charging Stations	• The posts were doing emphasis on the lack of charging station infrastructure
	2: EV Equity Awareness	• The posts were focusing on equity concerns regarding State EV Charging plan and claiming that African American and low - income communities should have access to opportunities on the EV and manufacturing industry.
	3: Gas and EV connection	• The posts were complaining since as EV drive they had to pay a EV tax and some irony regarding how the fossil fuels were still needed to support the EV transition.
El Paso, TX	1: EV Adoption Costs	• The posts were claiming about the high EV registration fees and the house adaptation to have a in-home charging station.
	2: Boasting EV on City	• The posts were boasting that they saw an EV on the street or that their EV arrive faster after it was ordered, showing satisfaction for it.
	3: EV Showcase	• The posts were promoting a Electric Car Festival that happened on September 2019.
Salt Lake City, UT	1: EV Charging Infrastructure	• The posts were talking about the eager of the new charging stations opening but also making awareness about the need to adapt the power requirements needed by the grid.
	2: Electric Trucks on the Road	• The posts were only about their users had spotted Electric trucks from multiple EV manufacturing companies on the SLC roads
	3: Carbon on EV manufacturing	• The posts were indicating that the EV manufacturing process still required carbon on their assembly

The social media posts that allowed to determine the topics are provided on Table 4.6.

Table 4.6. Social media posts that support the topic interpretation given in Table 4.5

City	Topic	Social Media posts
Indianapolis, IN	1: Charging Stations	• Since there's a high probability of rolling blackouts this summer due to energy crisis, how exactly would we be charging all these electric cars if we all had them? Serious question. • @JamesEBriggs @IndyMayorJoe We could be the first city with adequate EV parking/charging • Tesla and other EVs block gas station in protest against charging station Icing - Electrek

		• @Chyanne1107 In the Midwest that would be 100 kwh x $0.11 = $11. Adjust for your electric rate which could be as high as $0.45 per kWh. The biggest issue is the lack of infrastructure to support charging.
	2: EV Equity Awareness	• Ohio Congressman Tim Ryan sees a bright economic future in Electric Vehicle and battery manufacturing and charging infrastructure. He wants the Black and low income communities to directly benefit. @ Indiana • Despite equity concerns, federal government approves Indiana electric vehicle charging plan | news - Indiana Public Media • @jonnyktweets @Kaleidoscape2 @TheTweetOfRhea Honestly it would be both. The price of Electric Vehicle is unaffordable to most and even if you did have one. Most communities don't have sufficient number of charging stations. I live in Indianapolis and very few places here are equipped with charging stations for EVs.
	3: Gas and EV connection	• @MattDun10435175 @EileenDiana @offtheraiIs Well maybe Biden plans to buy you a new $100,000 electric car and pay for charging stations at each house. Oops, I think we need fossil fuels to support his transformation. But since he doesn't drive and never pays for anything, I doubt he cares • Indiana had us pay $150 EV tax this year because we own a @Tesla and won't be using gas..... really!? Heaven forbid we be more efficient. • These gas prices are gonna make everyone buy an electric car
El Paso, TX	1:EV Adoption Costs	• Electric, hybrid vehicle owners to see change in registration fees • States Hit Electric Vehicle Owners With High Fees - Consumer Reports • Thanks a lot @JoeBiden ! This REALLY makes me want to spend minimum $40-$120k on an electric vehicle. Not to mention the costs of upgrading my home electrical breakers and finding charging stations.
	2:Boasting EV on City	• I just saw a Tesla in El Paso! • Ordered my new #TeslaModel3 using my phone on Sept 16 before midnight deadline and was delivered in El Paso, Texas today oct 18. Thank you #tesla team. Happy customer. Happy to be part of Tesla family.
	3:EV Showcase	• Electric car festival in El Paso was a great idea without the pressure of a car salesperson. I enjoyed talking to the owners and hearing their experiences and stories. #ElectricVehicles

		• Is an electric bike in your future? Not sure about mine but they are cool and can take me to work. Checked them out today at the El Paso electric car festival. #ElectricVehicles
Salt Lake City, UT	1:EV Charging Infrastructure	• Supporting Leaders for Clean Air and @Packsize today, at the ribbon cutting for 52 new EV charging stations-the largest workplace charging infrastructure in the state. #cleanair • "@Electric_PC @_hypx @UndecidedMF EV charging infrastructure is not ready for the power requirements of long haul class 8 fast fill trucks Companies perusing hydrogen know it’s not economical but if/when the carbon tax hits, they need a working solution • Watch today's clean energy legislation celebration and unveiling of eight new EV charging stations in #SLC right here. #EarthDay @SLCgreen #utpol
	2:Electric Trucks on the Road	• Spotted a @Tesla truck in downtown SLC • First @Rivian #ev truck I've seen in the wild. • @TeslaMotors Unveiling a Electric Semi Truck @elonmusk #ILikeBigBatteries #ICanDriveForMiles #LudacrisMode
	3:Carbon on EV manufacturing	• @jtolds Your also missing the carbon foot print of manufacture... much higher for a new electric then for a used car • @ManMadeMoon With a fuel cell any large electric vehicle (bus, train, bus) can trade 90-95% of its batteries for a carbon fiber hydrogen tank. The vehicles will be lighter and have more range â€” all the benefits of an electric drivetrain remain (regenerative braking and instant torque)

This Descriptive analysis allowed the author to appreciate what is the status of EVA on social media per city based on the available data from social media. However, it is important to understand what could have been the observable causes that may have had an impact on the Descriptive analysis, which only can be explained by a diagnostic analysis, as shown in the following section.

4.2. DIAGNOSTIC ANALYTICS

Diagnostic analytics focuses on identifying the possible causes or reasons behind observed patterns or trends in data. Building on the insights provided by descriptive analytics, diagnostic

analytics delves deeper to uncover the factors that could have contributed to the sentiment scores obtained in the previous section. This type of analysis is essential for understanding the 'why' behind the data patterns observed.

4.2.1 Average Sentiment and Social Media Users through Control Charts

Statistical control charts help monitor the dataset's behavior through time. In this context, it is possible to determine what moments the sentiment went beyond its usual limits and the number of users talking about the topic. The control charts were divided by city by month, each with the sentiment average.

The first case corresponds to Indianapolis, where it is possible to appreciate that through the time study period, at least one social media post related to EVA was uploaded by a user being in the Indiana state capital as shown in Figure 4.8 and Figure 4.9.

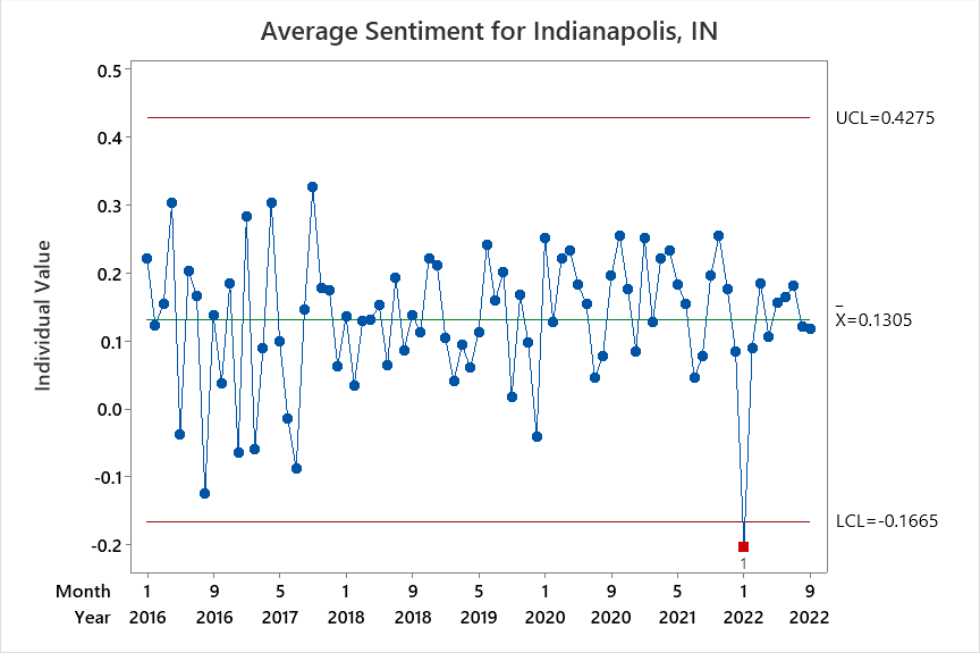


Figure 4.8. Monthly Average Sentiment control chart on Indianapolis, IN from January 2016 to September 2022

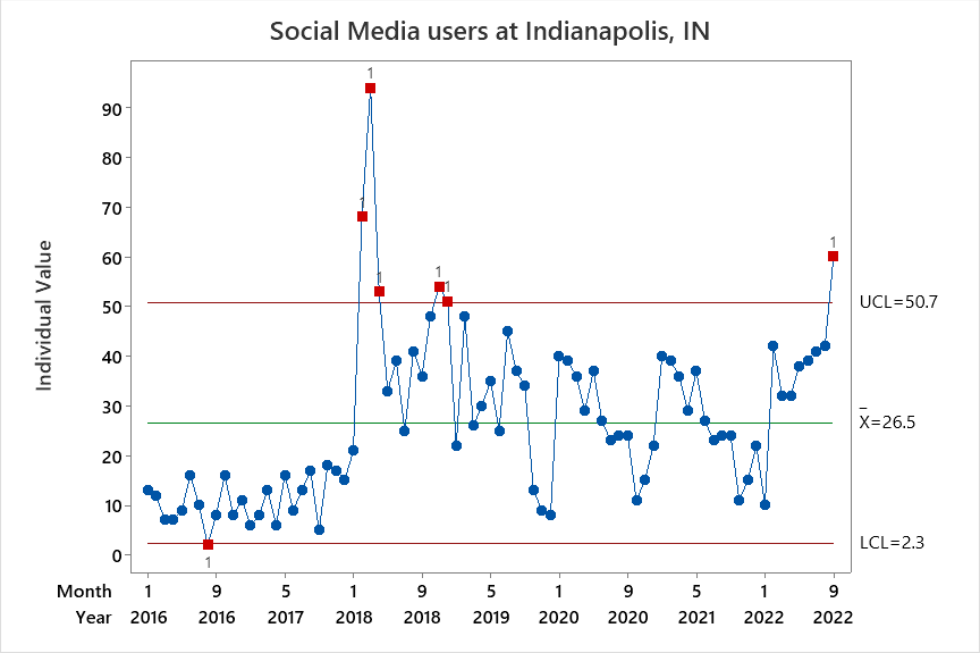


Figure 4.9. Monthly Social Media users control chart on Indianapolis, IN from January 2016 to September 2022

As shown on the previous charts (Figure 4.8 and Figure 4.9), the EVA topic was constant throughout the study period on Indianapolis. The Upper Control Limits (UCL) were 0.4275 and 50.7, while the Lower Control Limits (LCL) were -0.1665 and 2.3 on the Sentiment (Figure 4.8) and Social Media Users (Figure 4.9) charts, respectively. Regarding the average sentiment, it went below the LCL with -0.2033 in January 2022. In addition, before January 2018, the topic sentiment constantly fluctuated, with the monthly social user number at most 18. The extraordinary increment of users in the February-April and November-December periods in 2018, up to 98 users, shows that the average sentiment was close to the average sentiment throughout the time. On his article, Stall (2017) pointed out on November 2017 that Indiana had the potential to return as battery development and manufacturing leader on North America by exploring other type of batteries different from the lithium-ion battery to power electric buses and microgrid state systems. This fact could have boosted the interest on EVA on the February-April period. In addition, it is

possible to appreciate that in the May-July 2022 period there was a constant increase in the aggregated average sentiment and the number of social media users discussing about EVA. This phenomenon may be attributed by the involvement of the electrified roadways pilot program promoted by the state government and Purdue University research group, as reported by Sullivan (2022).

The second case corresponds to El Paso, where some social media posts had been uploaded in October and November 2016, but it was not until September 2017 when more people started to post or comment regarding EVs in the *El Paso del Norte* Borderland Area through the remaining study time, as shown on Figure 4.10 and Figure 4.11.

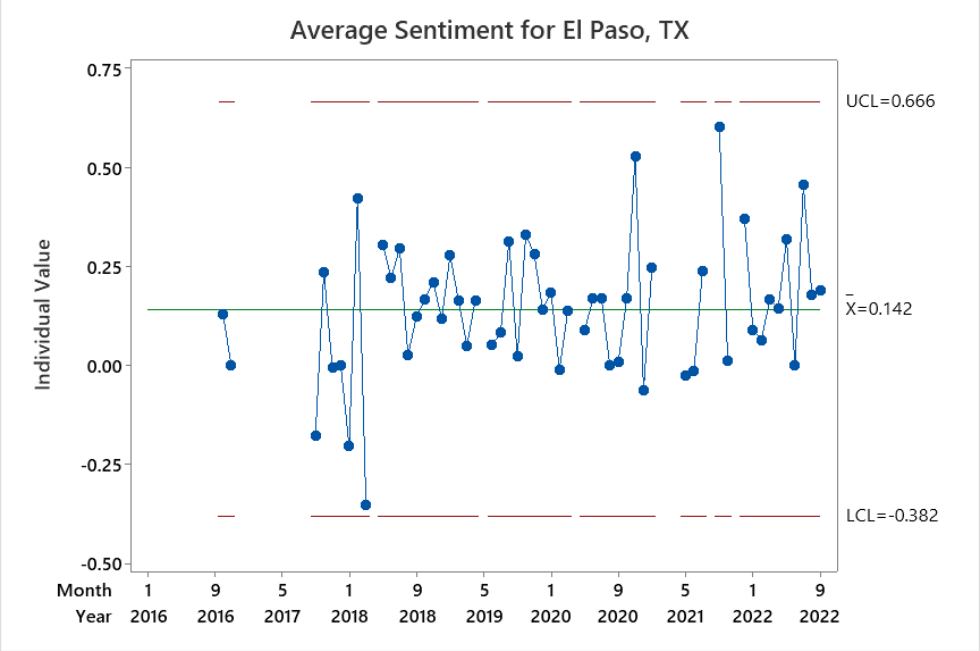


Figure 4.10. Monthly Average Sentiment control chart in El Paso, TX from January 2016 to September 2022

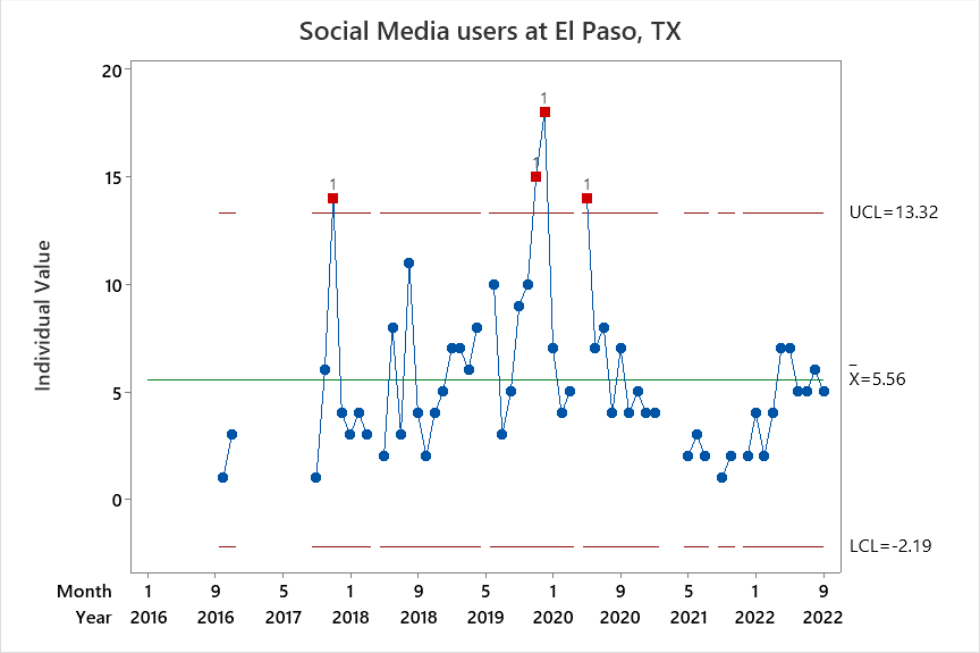


Figure 4.11. Monthly Social Media users control chart in El Paso, TX from January 2016 to September 2022

As shown on the previous charts (Figure 4.10 and Figure 4.11), the UCL were 0.666 and 13.32, while the LCL were -0.382 and 0 on the Sentiment (Figure 4.10) and Social media users (Figure 4.11) charts. The average sentiment remained inside control limits during the period. However, the social media users chart showed that on November 2017, November-December 2019, and May 2020, the number of users talking about EVA was higher than 13, reaching 18 as the maximum. It is possible to appreciate that on the August-November 2020 period, the sentiment was having an increment despite the constant change of the social media users talking about the topic. This event could exist since on August 2020 the UTEP announced its partnership with ASPIRE-ERC, which main objective is to develop new infrastructure that facilitates widespread adoption of EV, as reported UTEP Communications (2020). In addition, it was detected an increment in the aggregated average sentiment from January to July 2022, going from 0.0640 to 0.4565 during the period. Plans to develop state EV Charging infrastructure (Oxner, June 2022) including regulations (Torres, July 2022) and transforming local school bus fleet (Pskowski, March 2022) could have boosted the conversation regarding EVA in a positive way.

The third case corresponds to Salt Lake City, the Utah state capital, where it is possible to appreciate that through the time study period, most of the social media posts related to EVA were submitted only in two periods: January 2018-November 2020, and February 2022-September 2022, as shown on Figure 4.12 and Figure 4.13.

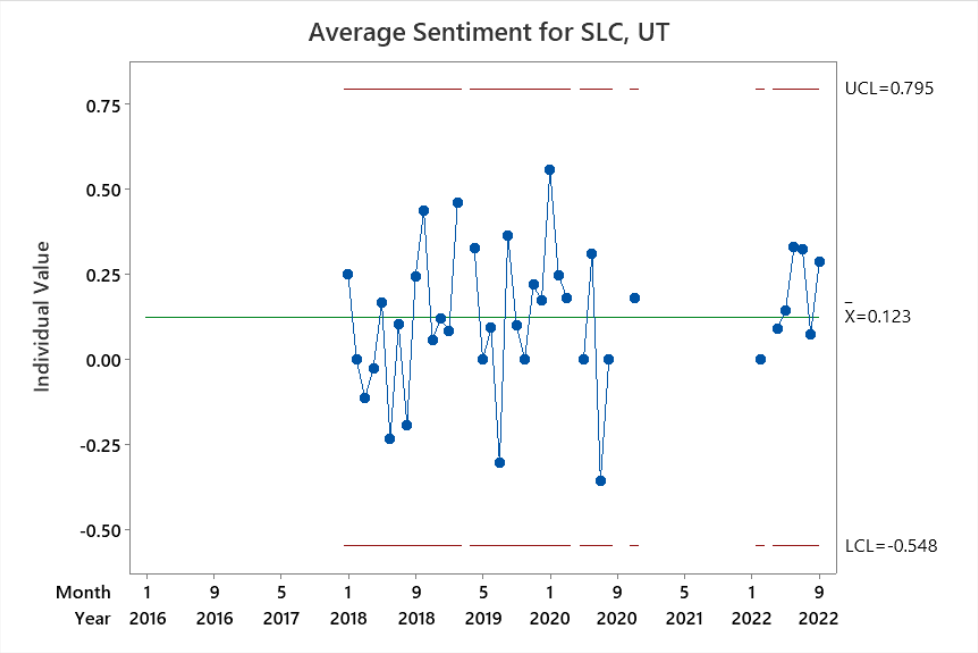


Figure 4.12. Monthly Average Sentiment control chart in Salt Lake City, UT from January 2016 to September 2022

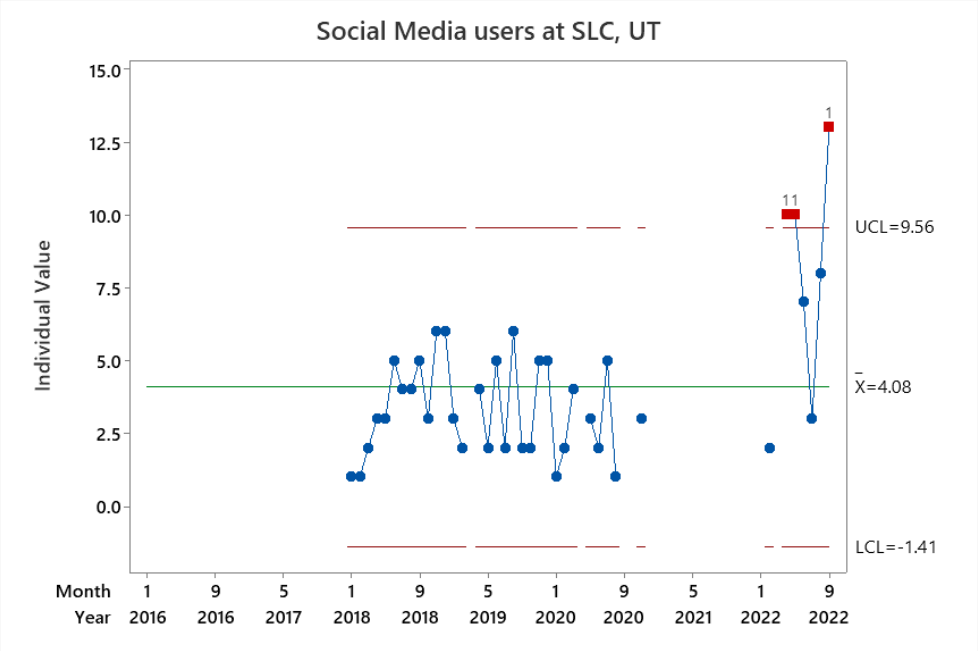


Figure 4.13. Monthly Social Media users control chart in Salt Lake City, UT from January 2016 to September 2022

As shown on the previous charts (Figure 4.12 and Figure 4.13), the UCL were 0.795 and 9.56, while the LCL were -0.548 and 0 on the Sentiment (Figure 4.12) and social media users (Figure 4.13) charts. The average sentiment remained inside control limits during the period. However, the social media users chart showed that on April-May 2022, and September 2022, the number of users talking about EVA was higher than 9, reaching 13 as the maximum. Even though the number of Twitter social media users was less than nine people, there were events that could have motivated the sentiment on some of those months. The completion of the I-15 Electric corridor (Lee, June 2018), new EV chargers available (McNaughton, January 2018; The Salt Lake City Tribune, July 2018) and future plans (Strong, February 2020) could be also a boost on the EVA public perception. On the other hand, accidents where an EV was involved (Hattem, May 2018; Holley, May 2018) may have a negative impact on the average sentiment, as seen on the May-June 2018 period. It is important to mention that on the time period from January 2018 to January 2020, in a continuous pattern, did not have average sentiment or social media user peaks. This behavior had been reported by Ruan and Lv (2023) when it was documented that despite the

increment number of social media users posting regarding EV on Reddit, it had not affected the average sentiment by causing a significant peak

A detailed table with the average sentiment, and the count of social media users per month is provided in Appendix 2.

4.3. PRESCRIPTIVE ANALYTICS

Finally, based on the available data, prescriptive analytics recommends specific actions that could optimize future outcomes through advanced techniques from knowledge fields such as Machine Learning and Causal Inference. The main goal is to offer strategies and interventions that could enhance performance and achieve desired outcomes.

For this analysis stage, each user's social media post was classified by gender with the support of the *gender* package, as stated on Section 3.3.6. The results are shown on Table 4.7.

Table 4.7. Decomposition of social media posts by gender and city

City	Female	Male	Total
El Paso, TX	76	192	268
Indianapolis, IN	1165	461	1626
Salt Lake City, UT	36	97	133
Total	1277	750	2027

When Table 4.1 and Table 4.7 are compared, it is possible to appreciate the fact that despite the sample size reduction based on the last filtration process step, the proportions between cities remain. The El Paso sample size was reduced from 307 social media posts to 268, a sample size reduction of 12.7%. The Indianapolis sample size was reduced from 1820 social media posts to 1626, a sample size reduction of 10.7%. Finally, the Salt Lake City sample size was reduced from 156 social media posts to 133 social media posts, a sample size reduction of 14.7%. In other words, only 38.78% of all the social media posts gathered from the Twitter API remained until the last analysis stage. It is not uncommon for data to have less than half of the social media posts prevails during the analysis step. Luna (2019) analysis only had the equivalent to the 7.195% (6504 out of

903,583) of social media posts to test statistical methods such as Lasso with Bootstrapping and Causal Inference when it was analyzed the public response to natural disasters in social media. In addition, Banik (2023) used only the 0.112% (31,438 out of 28 million) of the post's dataset after the preprocessing step before the patient experience during COVID-19.

4.3.1 Phase 1: Frequentist Approach Results

Analyzing these results involves seeing them from two perspectives to understand how the deletion of the multicollinearity problem aids in determining the statistical significance of the predictor variables. For the variable nomenclature meaning, please check Appendix 1.

By using **Pathway 1**, it was considered to keep all the original variables. The results for the *Three Cities* analysis showed that 87.88% (or 88 variables out of 99) of all the variables were considered outstanding in the Bootstrapping analysis at least in one iteration. A Pareto analysis was performed, focusing on those variables whose occurrences were adding a cumulative sum of 70%, obtaining only 22 variables (25%) with a cumulative occurrence sum percentage of 70.88%, as shown in Figure 4.14.

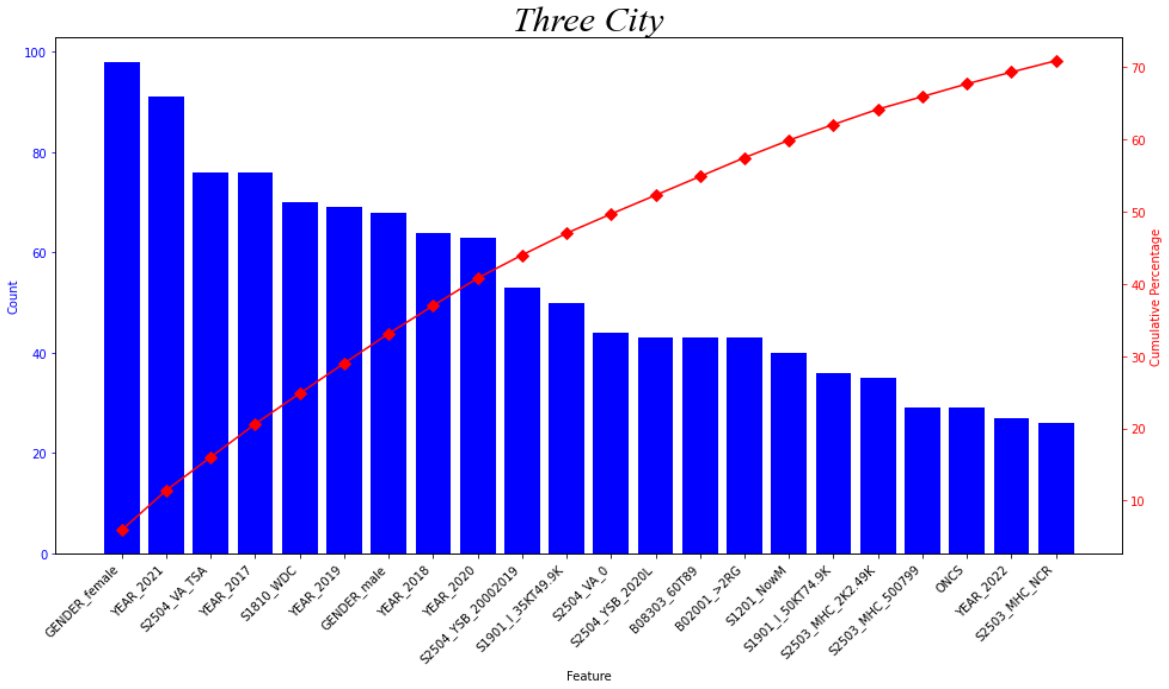


Figure 4.14. Three Cities Pareto analysis, where the Pathway 1 method was implemented

Once the OLS model was applied, it was reported that the lowest P-value was 0.126, corresponding to the *S1201_NowM* (Marital Status: Now Married (except separated)) variable, as shown on Table 4.8. In other words, despite those variables being considered outstanding through the iterations, they were not statistically significant for the response variable in the OLS Model. In addition, the coefficients shown were high, with a range of values between -6.45E+12 and 7.87E+12, except for the model constant -2.76E-17), as shown in Table 4.8. Finally, the performance metrics showed that there is a significant prediction error with Testing MSE of 1.0522, Testing RMSE of 1.0258 and a Testing R^2 of 0.0017, as shown in Table 4.17.

Table 4.8. OLS Model for the *Three Cities*, where the Pathway 1 method was implemented.

Variable	Coefficient	Std. Deviation	P-value
S1201_NowM	5.87E+12	3.83E+12	0.126
YEAR_2022	7.87E+12	5.59E+12	0.159
S2503_MHC_2K2.49K	-6.45E+12	4.71E+12	0.171
S1901_I_35KT49.9K	2.27E+12	1.89E+12	0.229
GENDER_female	-5.67E+12	5.37E+12	0.291
GENDER_male	-5.67E+12	5.37E+12	0.291
B08303_60T89	-4.83E+12	4.6E+12	0.294
ONCS	-1.97E+12	1.92E+12	0.306
S2503_MHC_500799	-4.53E+12	4.56E+12	0.32
S1810_WDC	-3.85E+12	4.32E+12	0.374
YEAR_2017	9.57E+11	1.15E+12	0.405
YEAR_2020	-3.04E+12	4.26E+12	0.475
S2504_YSB_20002019	-1.34E+12	2.11E+12	0.525
S1901_I_50KT74.9K	2.19E+12	3.86E+12	0.571
S2504_VA_TSA	1.84E+12	3.63E+12	0.612
S2504_YSB_2020L	1.43E+12	3.21E+12	0.657
B02001_>2RG	9.52E+11	2.21E+12	0.667
S2504_VA_0	1.05E+12	3.37E+12	0.754
S2503_MHC_NCR	5.65E+11	2.11E+12	0.789
YEAR_2019	5.88E+11	4.33E+12	0.892
YEAR_2021	4.03E+11	4.75E+12	0.932
YEAR_2018	1.82E+11	2.21E+12	0.934
Const	-2.76E-17	0.027	1

The results for the *Indianapolis* analysis showed that 77.08% (or 74 variables out of 96) of all the variables were considered outstanding in the Bootstrapping analysis at least in one iteration. The Pareto analysis was performed, reporting that only 23 variables (31.08%) gathered cumulative occurrences sum percentage of 70.20%, as shown in Figure 4.15.

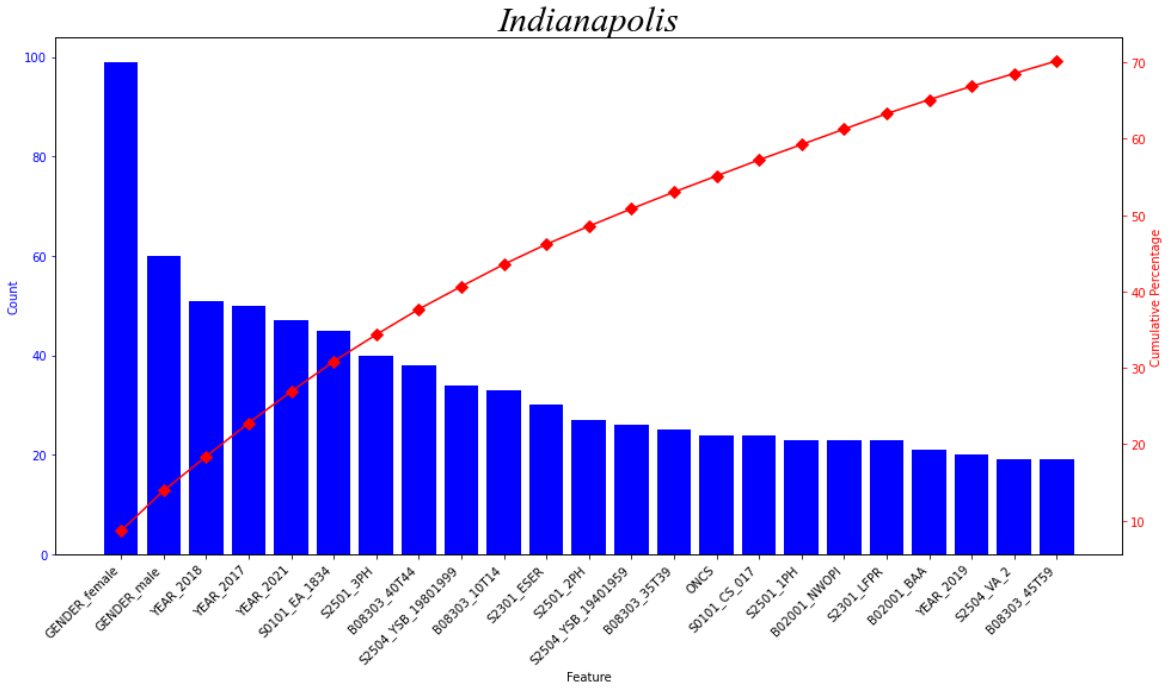


Figure 4.15. *Indianapolis* Pareto analysis, where the Pathway 1 method was implemented

The OLS model report showed that only three variables were statistically significant by having a P-value lower than 0.05. The variables were *GENDER_female* (the person who posted on social media was female), *GENDER_male* (the person who posted on social media was male), and *B02001_BAA* (the racial group was Black or African American alone). However, the coefficients showed were high, with a range of values between -7.95E+11 and 4.39E+12, except for the model constant (-0.2463), as shown in Table 4.9. Finally, the performance metrics showed that there is a significant prediction error with Testing MSE of 1.0522, Testing RMSE of 1.0258 and a Testing R^2 of 0.0017, as shown in Table 4.17.

Table 4.9. OLS Model for Indianapolis, where the Pathway 1 method was implemented.

Variable	Coefficient	Std. Deviation	P-value
GENDER_female	4.39E+12	2.05E+12	0.032
GENDER_male	4.39E+12	2.05E+12	0.032
B02001_BAA	-7.47E+12	3.69E+12	0.043
S0101_EA_1834	4.11E+12	2.32E+12	0.077
B08303_40T44	3.72E+12	2.53E+12	0.142
YEAR_2021	3.70E+12	2.89E+12	0.201
S2501_3PH	1.84E+12	1.50E+12	0.221
S2501_1PH	3.76E+12	3.38E+12	0.266
B08303_45T59	-3.87E+12	3.88E+12	0.318
S2504_YSB_19801999	-2.70E+12	2.89E+12	0.351
S0101_CS_017	2.38E+12	2.66E+12	0.371
S2504_VA_2	-1.38E+12	1.65E+12	0.405
B08303_35T39	-2.34E+12	2.90E+12	0.42
S2504_YSB_19401959	9.14E+11	1.17E+12	0.436
S2501_2PH	1.76E+12	2.35E+12	0.455
B02001_NWOPI	-7.95E+11	1.44E+12	0.582
S2301_LFPR	-1.64E+12	3.55E+12	0.645
S2301_ESER	-7.99E+11	1.82E+12	0.661
Const	-0.2463	0.8	0.758
YEAR_2017	-8.15E+11	3.62E+12	0.822
YEAR_2018	4.65E+11	2.22E+12	0.835
ONCS	1.85E+11	9.13E+11	0.84
YEAR_2019	1.74E+11	2.74E+12	0.949
B08303_10T14	2.48E+10	2.67E+12	0.993

The results for the *Salt Lake City* analysis showed that 96.88% (or 88 variables out of 96) of all the variables were considered outstanding in the Bootstrapping analysis at least in one iteration. It is important to mention that due to the training sample size was only nine, the Cross-validations were reduced to nine as well, allowing the model to run under those conditions. A Pareto analysis showed that only 30 variables (31.25%) possessed a cumulative occurrence sum percentage of 70.57%, as shown in Figure 4.16.

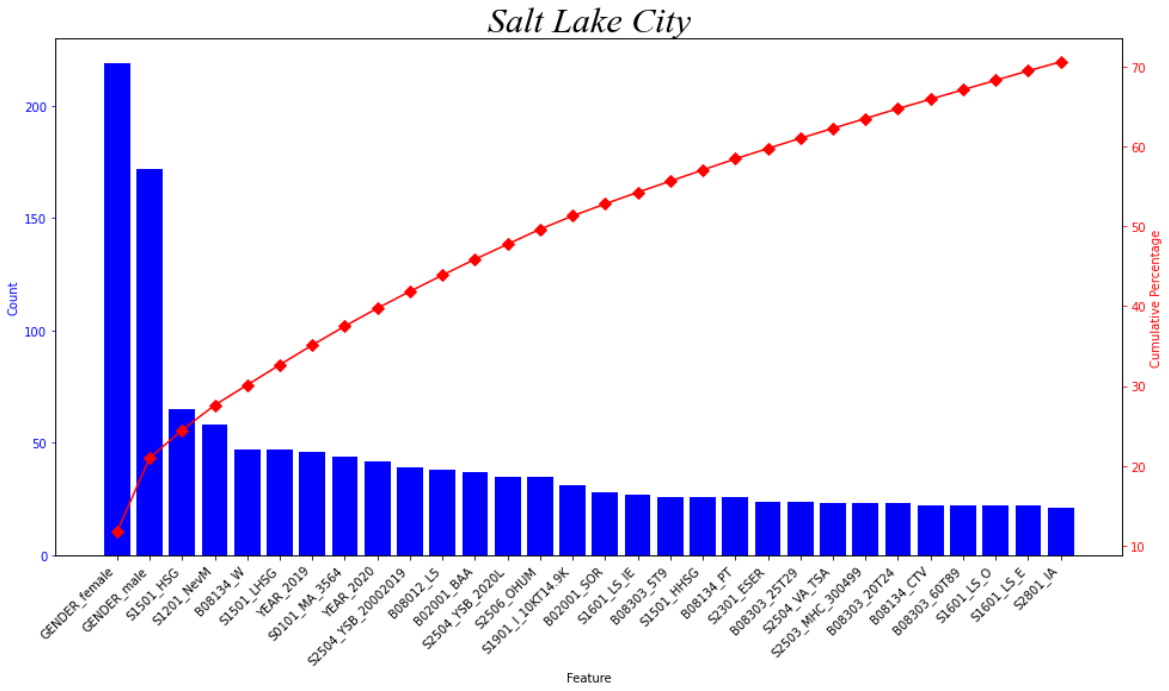


Figure 4.16. *Salt Lake City* Pareto analysis, where the Pathway 1 method was implemented

The OLS model report showed that only seven variables were statistically significant by having a P-value lower than 0.05. The variables were *S1501_HHSG* (Educational Attainment was Higher than HS Degree), *B08134_CTV* (Means of transportation to Work is either Car, Truck or Van), *B08134_PT* (Means of transportation to Work is Public Transportation), *S1501_LHSG* (Educational Attainment was Lower than HS Degree), *B08303_20T24* (Travel Time to work is from 20 to 24 minutes), *S2301_ESER* (Employment Rate), and *S2504_VA_TSA* (Telephone Service Availability in Housing Unit), as shown in Table 4.10. Finally, the performance metrics showed that there is a significant prediction error with Testing MSE of 1.9887, Testing RMSE of 1.4102 and a Testing R^2 of -0.1901, as shown in Table 4.17.

Table 4.10. OLS Model for Salt Lake City, where the Pathway 1 method was implemented.

Variable	Coefficient	Std. Deviation	P-value
S1501_HHSG	0.0212	0.007	0.005
B08134_CTV	-0.0198	0.007	0.007
B08134_PT	-0.018	0.007	0.013
S1501_LHSG	-0.0269	0.011	0.014
B08303_20T24	-0.0183	0.008	0.023
S2301_ESER	-0.0238	0.011	0.029
S2504_VA_TSA	-0.0247	0.011	0.031
S2506_OHUM	-0.0258	0.013	0.054
B02001_SOR	-0.0149	0.008	0.074
S1601_LS_O	-0.0193	0.011	0.082
B08303_25T29	-0.0185	0.011	0.093
B08134_W	-0.0212	0.014	0.122
S2504_YSB_20002019	-0.0221	0.014	0.125
S0101_MA_3564	0.0167	0.011	0.149
S2504_YSB_2020L	0.0204	0.015	0.169
S1901_I_10KT14.9K	-0.0187	0.014	0.183
B08303_60T89	0.011	0.008	0.186
GENDER_female	-0.0582	0.052	0.264
GENDER_male	0.0582	0.052	0.264
YEAR_2020	0.0156	0.015	0.29
S2503_MHC_300499	-0.0092	0.009	0.326
B08012_L5	-0.0117	0.014	0.399
B08303_5T9	0.0064	0.01	0.525
S2801_IA	0.0046	0.01	0.655
S1601_LS_E	0.0054	0.012	0.659
YEAR_2019	-0.0055	0.014	0.699
S1201_NevM	-0.0036	0.014	0.797
B02001_BAA	0.0027	0.012	0.824
S1601_LS_IE	0.0023	0.013	0.858
S1501_HSG	0.0007	0.011	0.949
Const	-4.12E-15	0.1	1

The results for the *El Paso* analysis showed that 92.71% (or 89 variables out of 96) of all the variables were considered outstanding in the Bootstrapping analysis at least in one iteration. A Pareto analysis showed that only 25 variables (22.09%) possessed a cumulative occurrence sum percentage of 70.20%, as shown in Figure 4.17.

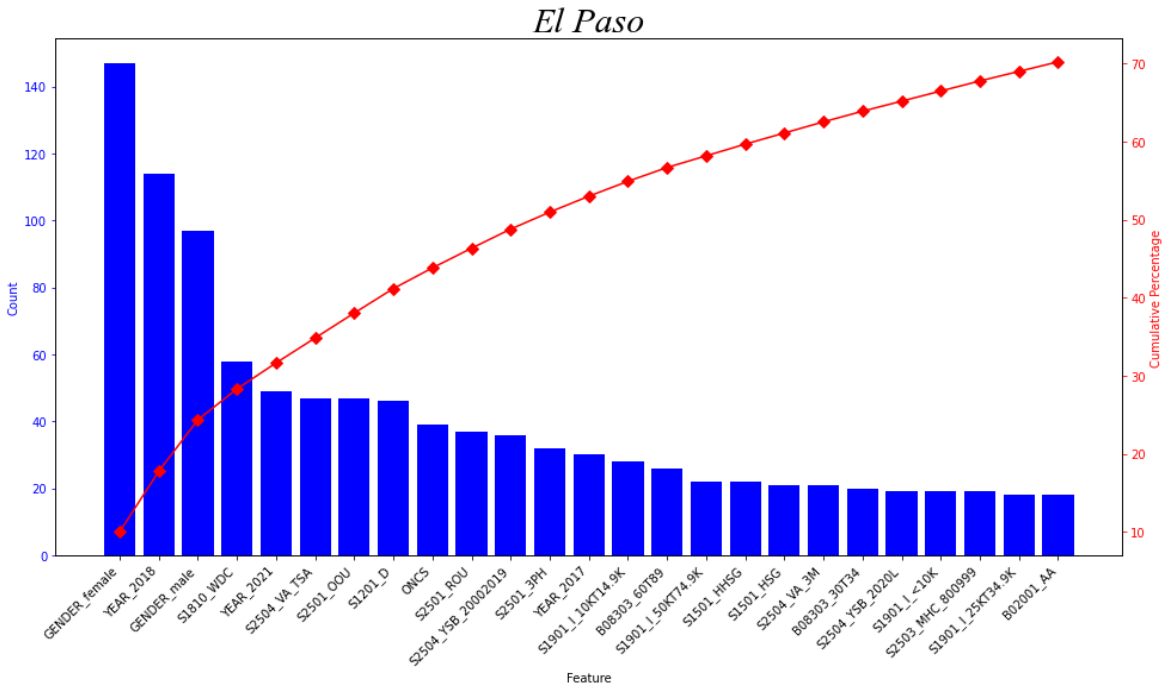


Figure 4.17. *El Paso* Pareto analysis, where the Pathway 1 method was implemented

Once the OLS model was applied, it was reported that the lowest P-value was 0.188, corresponding to the *B02001_AA* (Racial Group was Asian alone) variable, as shown on Table 4.11. In other words, despite those variables being considered outstanding through the iterations, they were not statistically significant for the response variable in the OLS Model. Finally, the performance metrics showed that there is a significant prediction error with Testing MSE of 0.9542, Testing RMSE of 0.9768 and a Testing R^2 of -0.0039, as shown in Table 4.17.

Table 4.11. OLS Model for El Paso, where the Pathway 1 method was implemented.

Variable	Coefficient	Std. Deviation	P-value
B02001_AA	-0.0151	0.011	0.188
S1901_I_<10K	-0.0198	0.015	0.19
S2501_ROU	-0.0188	0.015	0.209
S1901_I_25KT34.9K	-0.0267	0.021	0.211
S2501_3PH	-0.0162	0.015	0.27
S2501_OOU	-0.0168	0.015	0.273
YEAR_2021	-0.0166	0.021	0.437
S2504_YSB_20002019	-0.0129	0.017	0.442
ONCS	-0.0137	0.019	0.472
S2504_VA_TSA	0.0148	0.021	0.477
S1810_WDC	0.0148	0.021	0.478

S2503_MHC_800999	-0.013	0.019	0.503
S1901_I_50KT74.9K	0.0087	0.015	0.549
YEAR_2017	-0.0222	0.038	0.563
B08303_60T89	-0.0146	0.028	0.606
S2504_VA_3M	0.0048	0.014	0.727
S1201_D	-0.0093	0.027	0.732
GENDER_female	-0.0088	0.038	0.817
GENDER_male	0.0088	0.038	0.817
YEAR_2018	-0.0064	0.04	0.874
S1901_I_10KT14.9K	-0.0026	0.018	0.885
S2504_YSB_2020L	-0.0018	0.017	0.914
B08303_30T34	0.0013	0.013	0.917
S1501_HHSG	0.0014	0.017	0.935
S1501_HSG	0.0048	0.07	0.945
Const	-2.5E-16	0.074	1

Meanwhile, by using **Pathway 2**, it was considered to remove those predictor variables that correlated higher than 0.7 with other predictor variables, obtaining the following results. The results for the *Three Cities* analysis showed that 15.15% (or 15 variables out of 99) of all the variables were considered outstanding in the Bootstrapping analysis at least in one iteration, as shown in Figure 4.18.

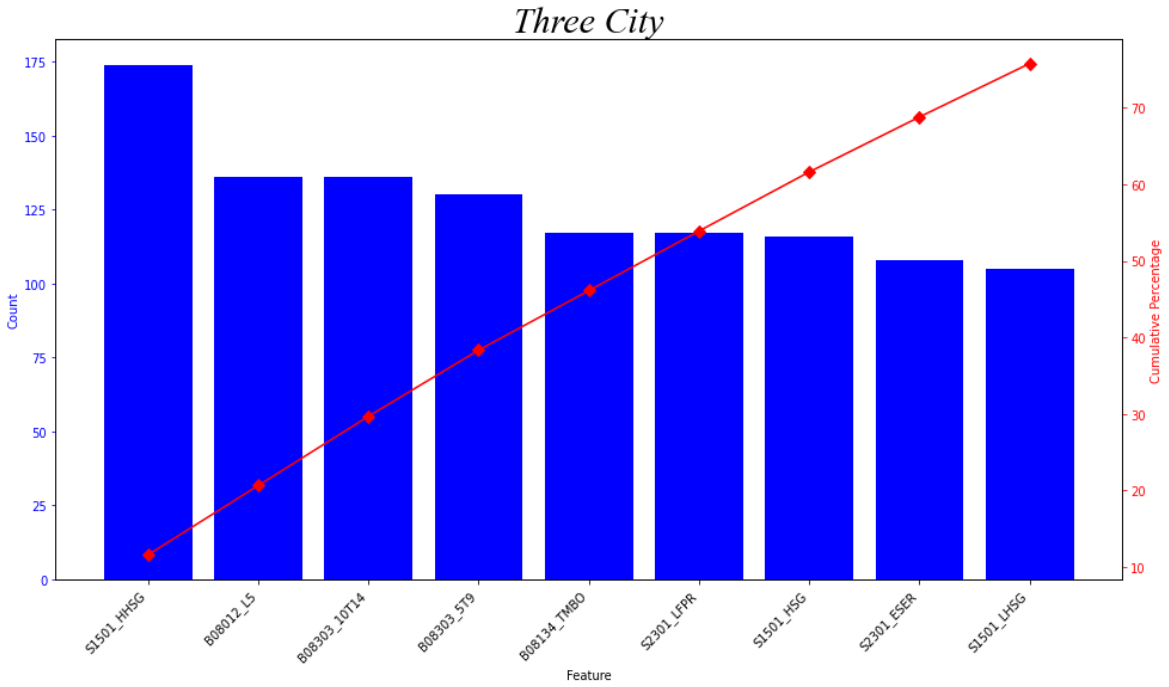


Figure 4.18. *Three Cities* Pareto analysis, where the Pathway 2 method was implemented

Once the OLS model was applied, it was reported that the lowest P-value was 0.204, corresponding to the *S1501_HSG* (Educational Attainment was High School graduate) variable, as shown on Table 4.12. In other words, despite those variables being considered outstanding through the iterations, they were not statistically significant for the response variable in the OLS Model. Finally, the performance metrics showed that there is a significant prediction error with Testing MSE of 1.0528, Testing RMSE of 1.0261 and a Testing R^2 of -0.0024, as shown in Table 4.17.

Table 4.12. OLS Model for the *Three Cities*, where the Pathway 2 method was implemented.

Variable	Coefficient	Std. Deviation	P-value
S1501_HSG	0.1682	0.132	0.204
B08303_5T9	0.0384	0.055	0.482
B08012_L5	0.1158	0.209	0.58
B08134_TMBO	-0.1538	0.278	0.581
B08303_10T14	0.2246	0.447	0.616
S1501_LHSG	0.0883	0.215	0.681
S2301_ESER	0.1092	0.402	0.786
S2301_LFPR	0.033	0.214	0.877
S1501_HHSG	0.0098	0.314	0.975
Const	-2.76E-17	0.027	1

The results for the *Indianapolis* analysis showed that 9.38% (or 9 variables out of 96) of all the variables were considered outstanding in the Bootstrapping analysis at least in one iteration. A Pareto analysis showed that only 5 variables (55.56%) possessed a cumulative occurrence sum percentage of 80.98%, as shown in Figure 4.19.

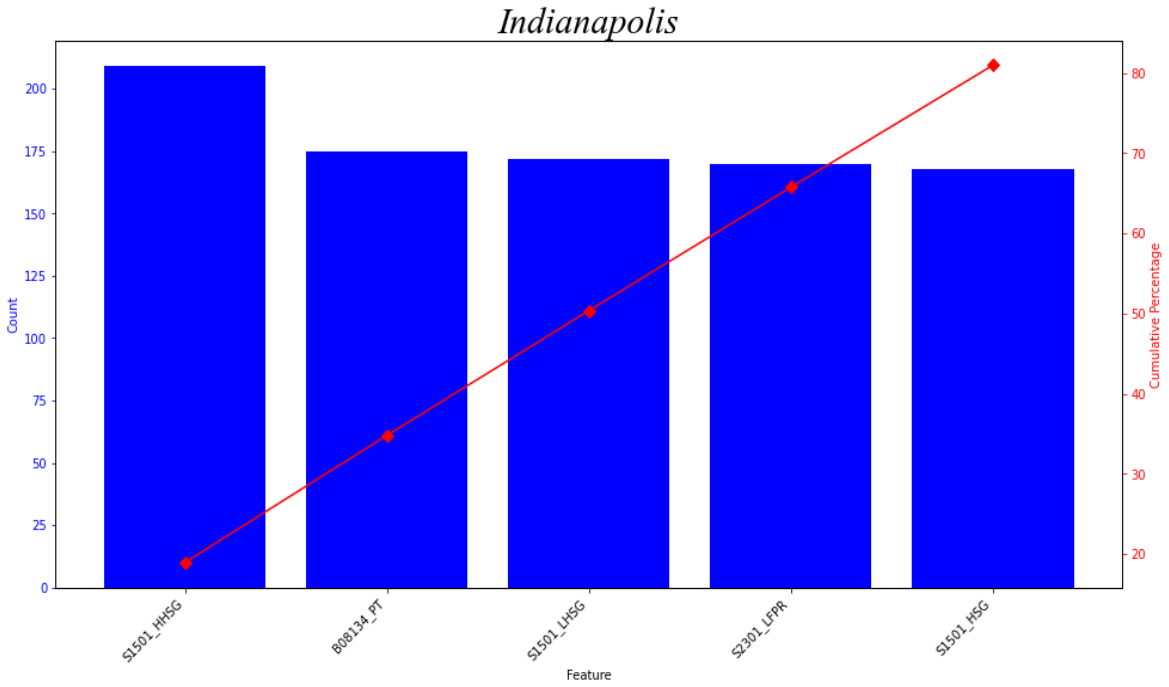


Figure 4.19. *Indianapolis* Pareto analysis, where the Pathway 2 method was implemented

Once the OLS model was applied, it was reported that the lowest P-value was 0.276, corresponding to the *S1501_HHSG* (Educational Attainment was Higher than High School graduate) variable, as shown in Table 4.13. In other words, despite those variables being considered outstanding through the iterations, they were not statistically significant for the response variable in the OLS Model. Finally, the performance metrics showed that there is a significant prediction error with Testing MSE of 1.0528, Testing RMSE of 1.0295 and a Testing R^2 of -0.0056, as shown in Table 4.17.

Table 4.13. OLS Model for the *Indianapolis*, where the Pathway 2 method was implemented.

Variable	Coefficient	Std. Deviation	P-value
S1501_HHSG	-44900	41200	0.276
S1501_LHSG	-37800	34700	0.276
S1501_HSG	-11300	10300	0.276
S2301_LFPR	-0.1621	0.214	0.448
B08134_PT	0.1133	0.371	0.76
Const	-2.5E-10	0.03	1

The results for the *Salt Lake City* analysis showed that 6.25% (or 6 variables out of 96) of all the variables were considered outstanding in the Bootstrapping analysis at least in one iteration. It is important to mention that due to the training sample size was only nine, the Cross-validations were reduced to nine as well, allowing the model to run under those conditions. A Pareto analysis showed that only 3 variables (50%) possessed a cumulative occurrence sum percentage of 90.35%, as shown in Figure 4.20.

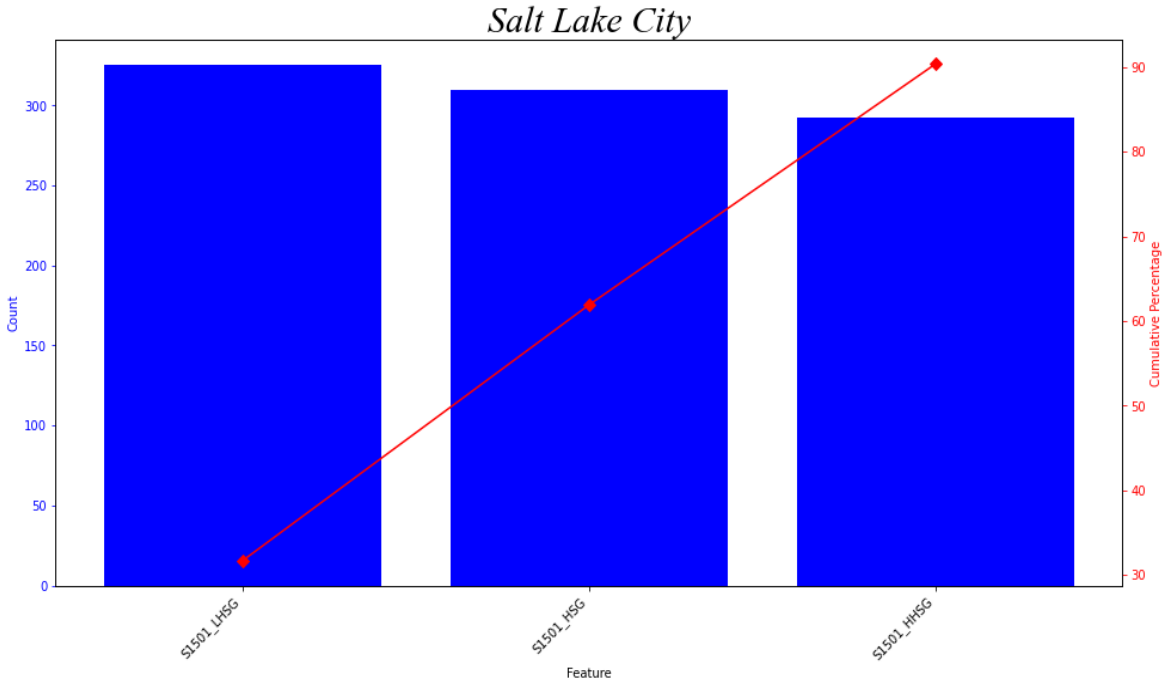


Figure 4.20. *Salt Lake City* Pareto analysis, where the Pathway 2 method was implemented

Once the OLS model was applied, it was reported that the lowest P-value was 0.558, corresponding to the *S1501_HHSG* (Educational Attainment was Higher than High School graduate) variable, as shown in Table 4.14. In other words, despite those variables being

considered outstanding through the iterations, they were not statistically significant for the response variable in the OLS Model. Finally, the performance metrics showed that there is a significant prediction error with Testing MSE of 2.0706, Testing RMSE of 1.4389 and a Testing R^2 of -0.2391, as shown in Table 4.17.

Table 4.14. OLS Model for the *Salt Lake City*, where the Pathway 2 method was implemented.

Variable	Coefficient	Std. Deviation	P-value
S1501_HHSG	0.283	0.481	0.558
S1501_HSG	-0.0985	0.257	0.702
S1501_LHSG	-0.0885	0.382	0.817
Const	-2.8E-17	0.1	1

The results for the *El Paso* analysis showed that 9.38% (or 9 variables out of 96) of all the variables were considered outstanding in the Bootstrapping analysis at least in one iteration. A Pareto analysis showed that only 4 variables (44.44%) possessed a cumulative occurrence sum percentage of 72.66%, as shown in Figure 4.21.

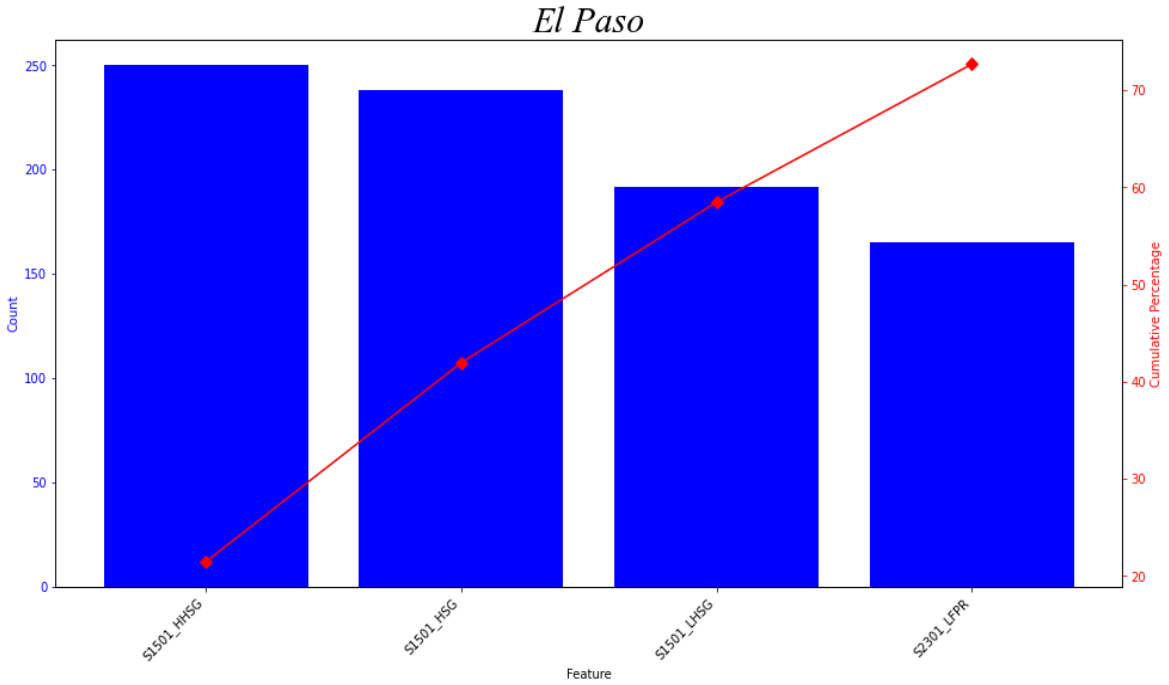


Figure 4.21. *El Paso* Pareto analysis, where the Pathway 2 method was implemented

Once the OLS model was applied, it was reported that the lowest P-value was 0.515, corresponding to the *S1501_HHSG* (Educational Attainment was Higher than High School graduate), *S1501_HSG* (Educational Attainment was High School graduate), and *S1501_HHSG* (Educational Attainment was Higher than High School graduate) variables, as shown in Table 4.15. In other words, despite those variables being considered outstanding through the iterations, they were not statistically significant for the response variable in the OLS Model. Finally, the performance metrics showed that there is a significant prediction error with Testing MSE of 0.9440, Testing RMSE of 0.9716 and a Testing R^2 of 0.0068, as shown in Table 4.17.

Table 4.15. OLS Model for *El Paso*, where the Pathway 2 method was implemented.

Variable	Coefficient	Std. Deviation	P-value
S1501_HHSG	4215.83	6459.98	0.515
S1501_HSG	4120.08	6313.11	0.515
S1501_LHSG	6936.18	10600	0.515
S2301_LFPR	0.0396	0.095	0.678
Const	1.3E-11	0.074	1

It is important to note that when Pathway 1 method was implemented, only in the Indianapolis and Salt Lake City scenarios showed two or three topics whose variables were at least once statistically significant, as it is shown on Table 4.16. In the case of Indianapolis, the social media observed variables topic, in addition to Race and Ethnicity, were the only topics that at least once their variables were significant in the models for that case.

Referring to the **Social media observed variables**, it was detected that the impact of gender, either male or female, was positive and statistically significant. A recent research paper of Plananska et al (2023) considered that the femininity had a positive correlation with the EV market share, suggesting that policies and marketing strategies should consider cultural characteristics and gender association when promoting EVA. This connection supports two ideas: the importance of understanding the user as individual whose attributes impact on their EVA perception and the

importance of studying what are the cultural impulse on the user's location that impact, directly or indirectly, on their EVA perception.

Referring to the **Race and Ethnicity**, it was detected that the impact of the Black African American racial group was negative to the EVA perception. Unfortunately, there have been scenarios where inequity affects the access of underrepresented racial groups to EV technology, therefore affecting their intentions to transition to a new transportation means. Khan (2022) discovered that the distribution of EV charging infrastructure on New York city was mostly focused on zip code areas where there were highways, not giving priority to populations where the habitants had a low-income, self-identified as Black and they were living in high population density zones. Other example was reported by Ermagun and Tian (2024), where it was discovered that areas predominantly inhabited by African Americans were experiencing low access to EV charging stations, causing to have a negative effect (statistically significant) in the probability of EV infrastructure presence and its density in the USA.

Meanwhile, for the Salt Lake City scenario, Education, Employment and Housing were the only topics that at least once their variables were significant in the models for that scenario were the only topics that at least once their variables were significant in the models for that scenario.

Referring to the **Education** topic, it was found through the coefficient models, the S1501_HHSG had a positive coefficient while the S1501_HSG and the S1501_LHSG had negative coefficients, despite they were not statistically significant (only for the model where the Pathway 1 method was implemented, and its p-value was 0.014). Kamis and Susan (2024) had concluded that Education could be an important predictor since those who had higher education would consider adopting an EVA while those with lower education attainment could be invited to pursue a technical career path in the EV industry or provide tools to improve their EVA tradeoff decision-making.

Referring to the **Employment** topic, the Transportation means as Public Transportation had a negative effect in the EVA perception sentiment, and it was statistically significant. After multiple articles were examined to study EVA patterns, Singh et al (2023) considered that this

transportation means could reduce the intention towards EV adoption due to the high level of transportation methods available for the users. This fact represents the need to analyze the current transportation services that are provided and users profile to focus on main locations where the EVA process should take place.

In addition, the Travel time to work from 20 to 24 minutes had a negative coefficient and it was statistically significant. This fact may connect with the findings from Mpoi et al (2023) where it was found that the charging time, as a statistically significant variable, was affecting negatively to the EV Payment intention. It should be considered that charging time could require the user to give more time on loading their EV in addition to the time needed to travel to their work destination.

Referring to the **Housing** topic, it was reported that the Telephone service availability could affect negatively the EVA perception. There is no study that talks about how the Telephone service affects EV Adoption, therefore future efforts should be towards understanding the function of this variable in the perception.

Finally, the Charging Infrastructure and Families and Living Arrangements topics were the ones with the lowest level of outstanding occurrences through the model, appearing only when the Pathway 1 method was being implemented in the dataset.

Referring to the **Charging Infrastructure** topic, it was detected that that its variable ONCS (Average number of new Charging Stations openings on the Year per Zip Code) was having either a positive (for Indianapolis model) or negative impact (for the 3 Cities and El Paso models) in the Sentiment but still not being statistically significant. This behavior is similar to the one reported by Brückmann et al (2021) where it was discovered that despite the charging availability showing positive to EVA, its effect was small and not statistically significant.

Referring to the **Families and Living Arrangements** topic, the Marital Status as Divorced and Never Married were considered to have negative effects in El Paso and Salt Lake City models respectively, while when the Marital Status was Now Married had positive effect in the Three

Cities model. Lin and Wu (2018) had reported that the Marital Status, when the respondent was married, had positive effects in the proposed model, being statistically significant.

A summary of the outstanding variables obtained on each model, divided by dataset topics is provided in Table 4.16.

Table 4.16. Summary of the number of variables that appeared in the Lasso models

	Pathway 1				**Pathway 2**			
Dataset Topic	Three Cities	Indiana polis	Salt Lake City	El Paso	Three Cities	Indiana polis	Salt Lake City	El Paso
Education			3*	2	3	3	3	3
Employment	2	6	8*	2	6	2		1
Families and Living Arrangements	1		1	1				
Health	1			1				
Housing	6	6	6*	8				
Income and Poverty	2		1	4				
Population and People		2	5					
Race and Ethnicity	1	2*	2	1				
Charging Infrastructure	1	1		1				
Social media Observed Variables	8	6*	4	5				
Total	**22**	**23**	**30**	**25**	**9**	**5**	**3**	**4**

*At least one time, one of its variables was statistically significant with a p-value<=0.05

As shown on Table 4.17, the performance metrics were similar either using Pathway method 1 or 2, showing how the dataset could be fitted into an OLS model. However, it is important to note that either the Training or Testing R-squared are negative, meaning that the models got overfitted when the testing sets were used. In other words, the models required specific dataset values, such as the training set values, losing the model generalization and having poor performance in its application. The only exception to the previous statement was for *El Paso* when the Pathway 2 method was applied, however, in both R-squared the results were close to 0, not being able to explain any variability of the response variable.

Table 4.17. Performance metrics of the six Lasso models divided by Pathway method and Analysis level

Pathway	Analysis Level	Training MSE	Training RMSE	Testing MSE	Testing RMSE	Training R^2	Testing R^2
1	3-City	0.9994	0.9997	1.0595	1.0293	0.0006	-0.0088
	Indianapolis	1.0053	1.0027	1.0522	1.0258	-0.0053	0.0017
	Salt Lake City	0.8848	0.9407	1.9887	1.4102	0.1152	-0.1901
	El Paso	0.9801	0.9899	0.9542	0.9768	0.0199	-0.0039
2	3-City	0.9968	0.9984	1.0528	1.0261	0.0032	-0.0024
	Indianapolis	0.9946	0.9973	1.0598	1.0295	0.0054	-0.0056
	Salt Lake City	0.8975	0.9474	2.0706	1.4389	0.1024	-0.2391
	El Paso	0.9933	0.9967	0.9440	0.9716	0.0067	0.0068

4.3.2 Phase 2: Bayesian Approach Results

First, Naïve-Bayes was implemented to remove those predictor variables that were not relevant to the target node *Sentiment_Category*. The boxplot (Figure 4.22) and Descriptive statistics (Table 4.18) are provided below.

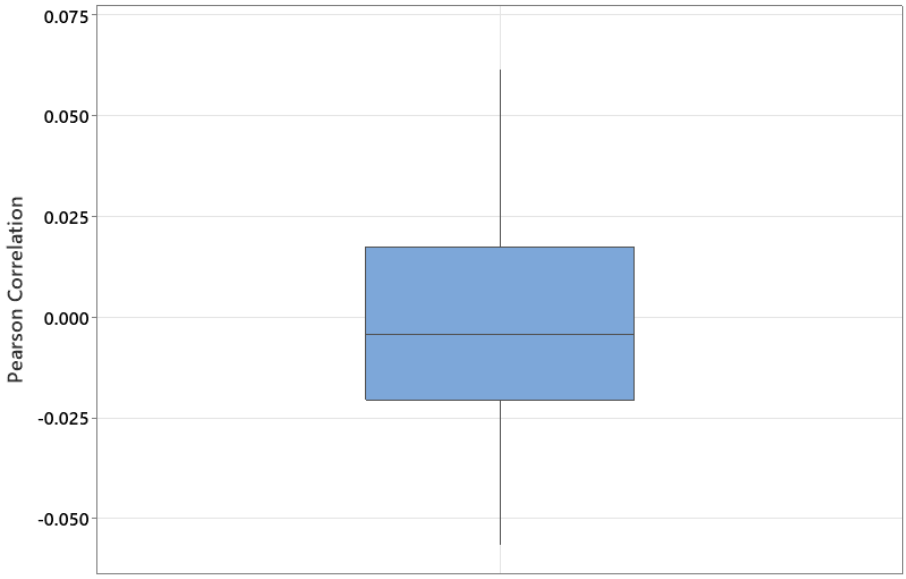


Figure 4.22. Boxplot distribution of the correlation between predictor variables and the target node (*Sentiment_Category*)

As it was shown on Figure 4.22 and Table 4.18, the Pearson correlation coefficients between the predictor variables and response variables varied from -0.0567 to 0.0615. As

published by LaMorte (2024), the correlation coefficients between -0.2 and 0.2 are considered either as weak association between variables or as no association between them.

Table 4.18. Descriptive Statistics of Pearson Correlation after Naïve-Bayes was applied in dataset.

Variable	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Pearson Correlation	90	-0.00136	0.00261	0.02480	-0.05670	-0.02063	-0.0043	0.01735	0.06150

In addition, the overall contribution also showed low levels of association between the response and predictor variables. Since the overall contribution from the node *Sentiment_Category* to the predictor nodes was less than 2%, having multiple ties between predictor nodes and keeping the same complexity level, it was considered not to continue with the BBN modeling once the *Naïve Bayes* was implemented. The detailed table where the Pearson correlation coefficients, overall contribution between variables, and p-values can be found in Appendix 3.

Since the first method did not release the outstanding number of nodes that could reduce the BBN complexity, the second method results were considered. The BBN shown in Figure 3.5 was implemented, showing the alternative hypothesis probabilities (Figure 4.23) and Bayes Factors (Figure 4.24) in Boxplots, decomposed by Sentiment and City respectively.

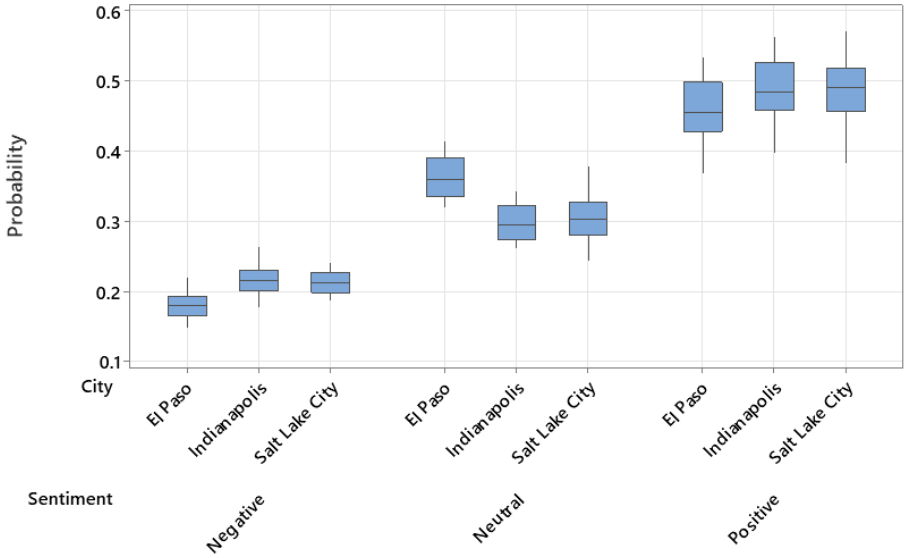


Figure 4.23. Sentiment category probability distributed by sentiment and city

As shown in Figure 4.23, the trend of each sentiment polarity per city remained similar to the polarity bar charts shown in Figure 4.1. In the cases of the Negative polarity, Indianapolis had the highest median of 0.4829, while El Paso had the lowest median of 0.4545. However, for the Neutral polarity, El Paso had the highest median of 0.3591, while Indianapolis had the lowest median of 0.2946. Finally, in the case of the Positive polarity, it is shown that the three cities showed the highest probabilities compared to the previous polarity sentiments. However, on this section Salt Lake City possess the highest median of 0.4892, being followed by Indianapolis (0.4829) and finally by El Paso (0.4545). Additional statistical metrics are found in Appendix 4.

In order to analyze the Bayes Factor when the sentiment polarity was positive based on the behavior given on Figure 4.24, the dataset was filtered by that attribute, obtaining boxplots divided by year by city, as shown on Figure 4.24.

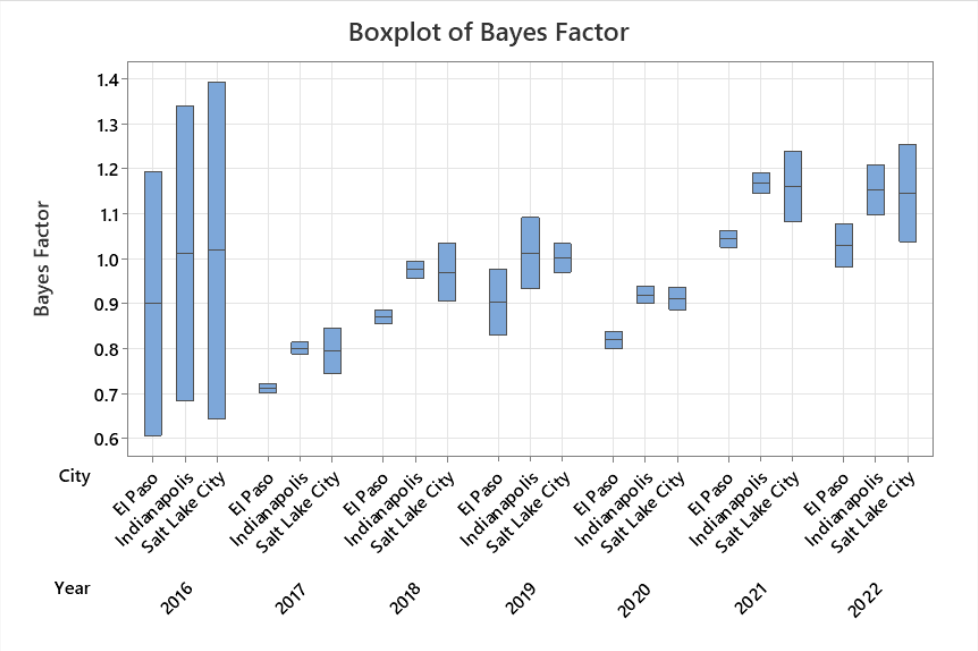


Figure 4.24. Boxplots of Bayes Factors when the sentiment polarity is Positive by year by city

As it is shown on Figure 4.24, there is an increment on the median trend through the years from 2017 through 2019 and from 2020 through 2022. The Bayes Factors found in 2016 showed an important variation, by having standard deviations between 0.415 and 0.527 in the three cities, while in the other years, the variation remained low, except for Salt Lake City in 2021 and 2022,

with standard deviations of 0.1091 and 0.153 respectively. Based on Table 4.19, from 2017 to 2020, the Bayes Factor medians were lower than or equal to 1(for the case of Indianapolis and Salt Lake City in 2019), indicating that there was either anecdotal evidence for the null hypothesis or no evidence to sustain none of the hypotheses provided respectively. Referring to the years 2021 and 2022, where the Bayes Factor medians were higher than 1, it was discovered that there was anecdotal evidence for the alternative hypothesis, as shown on Table 4.19.

Table 4.19. Hypothesis Results from the Causal Inference Model

Hypothesis Testing	Probability	Bayes Factor
$H_1 = P(S = Positive \mid Y = 2021, C = Indianapolis, G = Male)$	0.5319	1.3397
$H_1 = P(S = Positive \mid Y = 2022, C = SLC, G = Male)$	0.5455	1.2544
$H_1 = P(S = Positive \mid Y = 2021, C = SLC, G = Male)$	0.5421	1.2376
$H_1 = P(S = Positive \mid Y = 2022, C = Indianapolis, G = Male)$	0.5361	1.2081
$H_1 = P(S = Positive \mid Y = 2021, C = Indianapolis, G = Female)$	0.5226	1.1442
$H_1 = P(S = Positive \mid Y = 2022, C = Indianapolis, G = Female)$	0.5119	1.0964
$H_1 = P(S = Positive \mid Y = 2021, C = SLC, G = Female)$	0.5089	1.0833
$H_1 = P(S = Positive \mid Y = 2022, C = EP, G = Male)$	0.5075	1.0773
$H_1 = P(S = Positive \mid Y = 2021, C = EP, G = Male)$	0.5041	1.0625
$H_1 = P(S = Positive \mid Y = 2022, C = SLC, G = Female)$	0.4982	1.0379
$H_1 = P(S = Positive \mid Y = 2021, C = EP, G = Female)$	0.4951	1.0251
$H_1 = P(S = Positive \mid Y = 2022, C = EP, G = Female)$	0.4843	0.9818

It is important to mention that Salt Lake City Bayes factors hypotheses were bigger than El Paso Bayes Factors, despite the sample size was bigger for El Paso than for Salt Lake City. A detailed table with all the possible combinations of observed variables is found on Appendix 5.

Chapter 5: Conclusion

The fusion of multiple datasets, coupled with innovative applications of Natural Language Processing, Machine Learning algorithms, and Causal Inference techniques, has empowered us to glean invaluable insights into Electric Vehicle Adoption (EVA) from the customer perspective. Analyzing social media data alongside Census information unveiled public sentiments toward EVs and unearthed potential driving factors behind these sentiments in three cities transitioning to EVs. The descriptive analytics allowed to determine the EVA perception, determining that it was mainly positive, and the common EVA topics by city based on the sample size obtained. The diagnostic analysis allowed to monitor how was the average perception sentiment and social media users' evolution through time, attempting to connect their peaks with document evidence, such as news. This type of analysis helped to demonstrate how the sample selection bias was affecting the results provided, therefore, it was a decisive factor when running the prescriptive analysis by implementing Bootstrapping (for the Frequentist statistical approach) and Bayesian Belief Network (for the Bayesian statistical approach) as ways to reduce the sample size limitation effects found per city. That strategy allowed to determine that to achieve a positive perception towards, EVA, researchers should analyze the profile of male population in the cities of Indianapolis and Salt Lake City during the years 2021 and 2022.

From these insights, practical recommendations were drawn for organizations, EV manufacturers, and policymakers. Implementing these recommendations could lead to the formulation of better policies, the adoption of improved practices, and the development of targeted marketing strategies.

Chapter 6: Future Work

It is expected that the proposed methodology and outcomes of this work will help future researchers to improve their understanding regarding how to understand public perception towards EVA. First, future research efforts should focus on **studying the social media usage** in USA cities, since it was detected that cities such El Paso do not have a high number of Twitter users to post their opinions, as in bigger cities such as Indianapolis.

Second, a new impending challenge the researchers should be able **to adapt is to the new X API policies** (Developers, 2023; Jingnan, 2023) implemented on February 9, 2023, restricting the extraction of posts from their databases. Therefore, the utility of the *Tweepy* package may be contingent on updates from its developers.

Third, the fusion of Census Data and sentiment score from social media demonstrated that it was challenging task, especially on the analysis stage. It is important to **test new observed variables** that could be extracted **from social media posts** to create a more accurate individual profile of the user rather than the context that surrounds them.

Fourth, the **sample selection bias** needs to be considered on future research efforts where social media data is used. On this work, it could only be corrected by the *Gender* and *City* observed attributes, however, it is essential to explore other observed attributes on social media that could also act as bias reduction agents.

References

Aalders, I. (2008). Modeling Land-Use Decision Behavior with Bayesian Belief Networks. *Ecology and Society, 13*(1) Retrieved from http://www.jstor.org/stable/26267920

Albers, C. J., Kiers, H. A. L., & van Ravenzwaaij, D. (2018). Credible Confidence: A Pragmatic View on the Frequentist vs Bayesian Debate. *Collabra: Psychology, 4*(1), 31. doi:10.1525/collabra.149

Austmann, L. M., & Vigne, S. A. (2021). Does environmental awareness fuel the Electric Vehicle market? A Twitter keyword analysis. *Energy Economics, 101*, 105337. doi:10.1016/j.eneco.2021.105337

Babvey, P., Capela, F., Cappa, C., Lipizzi, C., Petrowski, N., & Ramirez-Marquez, J. (2021). Using social media data for assessing children's exposure to violence during the COVID-19 pandemic. *Child Abuse & Neglect, 116*, 104747. doi:10.1016/j.chiabu.2020.104747

Bangroo, R., Verma, S. R., Shivangi, & Shakuntala. (2023). Comparative study of Elastic Net Regression, Naive Bayes & Lasso regression. Paper presented at the *2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM)*, doi:10.1109/ELEXCOM58812.2023.10370433

Banik, D. (2023). Evaluation of maternal patient experience during covid-19 using natural language processing (Master's Degree). *Open Access Theses & Dissertations*. 3764. Retrieved from scholarworks.utep.edu/open_etd/3764

Barash, M., McNevin, D., Fedorenko, V., & Giverts, P. (2024). Machine learning applications in forensic DNA profiling: A critical review. *Forensic Science International: Genetics, 69*, 102994. doi:10.1016/j.fsigen.2023.102994

Bayes, T. (1763). An Essay towards solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London*, 53(53), 370-418. doi:10.1098/rstl.1763.0053

Bayesia S.A.S. 2024. Bayesialab software version 11.4. Change (FR). http://www.bayesia.com/

Bilal, D. M., & Tamilselvan, S. (2024). Prediction of brain tumor severity with magnetic resonance imaging using Ridge Regression over Recurrent Neural Network. Paper presented at the *2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE),* doi:10.1109/IITCEE59897.2024.10468020

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python* (First ed.) O'Reilly Media, Inc.

Birjali, M., Beni-Hssane, A., & Erritali, M. (2017). Machine learning and semantic Sentiment Analysis-based algorithms for suicide sentiment prediction in social networks. *Procedia Computer Science, 113*, 65-72. doi:10.1016/j.procs.2017.08.290

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research, 3*, 993-1022. doi:10.5555/944919.944937

Borg, A., & Boldt, M. (2020). Using VADER sentiment and SVM for predicting customer response sentiment. *Expert Systems with Applications, 162*, 113746. doi:10.1016/j.eswa.2020.113746

Borah, P. S., Iqbal, S., & Akhtar, S. (2022). Linking social media usage and SME's sustainable performance: The role of digital leadership and innovation capabilities. *Technology in Society, 68,* 101900. doi:10.1016/j.techsoc.2022.101900

Boushey, H. (2023). Full charge: The economics of building a national EV charging network. Retrieved from https://www.whitehouse.gov/briefing-room/blog/2023/12/11/full-charge-the-economics-of-building-a-national-ev-charging-network/

Brückmann, G., Willibald, F., & Blanco, V. (2021). Battery Electric Vehicle adoption in regions without strong policies. *Transportation Research Part D: Transport and Environment, 90,* 102615. doi:10.1016/j.trd.2020.102615

Carley, S., Krause, R. M., Lane, B. W., & Graham, J. D. (2013). Intent to purchase a plug-in electric vehicle: A survey of early impressions in large US cites. *Transportation Research Part D: Transport and Environment, 18,* 39-45. doi:10.1016/j.trd.2012.09.007

Chen, S., Keglovits, M., Devine, M., & Stark, S. (2022). Sociodemographic Differences in Respondent Preferences for Survey Formats: Sampling bias and Potential Threats to External Validity. *Archives of Rehabilitation Research and Clinical Translation, 4*(1), 100175. doi:10.1016/j.arrct.2021.100175

Chinnasamy, P., Suresh, V., Ramprathap, K., Jebamani, B. J. A., Srinivas Rao, K., & Shiva Kranthi, M. (2022). COVID-19 vaccine Sentiment Analysis using public opinions on Twitter. *Materials Today: Proceedings, 64,* 448-451. doi:10.1016/j.matpr.2022.04.809

Cho, H., She, J., De Marchi, D., Ek-Zaatari, H., Barnes, E. L., Kahkoska, A. R., . . . Virkud, A. V. (2023). Machine learning and health science research: Tutorial. *Journal of Medical Internet Research,* doi:10.2196/50890

Choi, J., Yoon, J., Chung, J., Coh, B., & Lee, J. (2020). Social media analytics and Business Intelligence research: A systematic review. *Information Processing & Management, 57*(6), 102279. doi:10.1016/j.ipm.2020.102279

Chumachenko, D., Bazilevych, K., Meniailov, I., Yakovlev, S., & Chumachenko, T. (2021). Simulation of COVID-19 Dynamics using Ridge Regression. Paper presented at the *2021 IEEE 4th International Conference on Advanced Information and Communication Technologies (AICT),* 163-166. doi:10.1109/AICT52120.2021.9628991

De Rosis, S., Lopreite, M., Puliga, M., & Vainieri, M. (2021). The early weeks of the Italian Covid-19 outbreak: Sentiment insights from a Twitter analysis. *Health Policy, 125*(8), 987-994. doi:10.1016/j.healthpol.2021.06.006

Developers [@XDevelopers]. (2023, February 1). Starting February 9, we will no longer support free access to the Twitter API, both v2 and v1.1. A paid basic tier will be available instead [Tweet]. Twitter. https://twitter.com/XDevelopers/status/1621026986784337922

Dlamini, W. M. (2010). A Bayesian belief network analysis of factors influencing wildfire occurrence in Swaziland. *Environmental Modelling & Software, 25*(2), 199-208. doi:10.1016/j.envsoft.2009.08.002

Du, W., Wu, X., & Tong, H. (2023). Fair regression under sample selection bias. Paper presented at the *2022 IEEE International Conference on Big Data (Big Data),* doi:10.1109/BigData55660.2022.10021107

Enders, C. K. (2010). *Applied missing data analysis* (First ed.) The Guilford Press.

Ermagun, A., & Tian, J. (2024). Charging into inequality: A national study of social, economic, and environment correlates of electric vehicle charging stations. *Energy Research & Social Science, 115*, 103622. doi:10.1016/j.erss.2024.103622

Fan, L., Chen, S., Li, Q., & Zhu, Z. (2016). Variable selection and model prediction based on Lasso, adaptive Lasso, and Elastic Net. Paper presented at the *2015 4th International Conference on Computer Science and Network Technology (ICCSNT)*, doi:10.1109/ICCSNT.2015.7490813

Fan, Z., Che, Y., & Chen, Z. (2017). Product sales forecasting using online reviews and historical sales data: A method combining the bass model and Sentiment Analysis. *Journal of Business Research, 74*, 90-100. doi:10.1016/j.jbusres.2017.01.010

Feldman, R. (2013). Techniques and applications for Sentiment Analysis. *Commun.ACM, 56*(4), 82–89. doi:10.1145/2436256.2436274

Fornacon-Wood, I., Mistry, H., Johnson-Hart, C., Faivre-Finn, C., O'Connor, J. P. B., & Price, G. J. (2022). Understanding the Differences Between Bayesian and Frequentist Statistics. *International Journal of Radiation Oncology*Biology*Physics, 112*(5), 1076-1082. doi:10.1016/j.ijrobp.2021.12.011

Franke, T., & Krems, J. F. (2013). Understanding charging behaviour of electric vehicle users. *Transportation Research Part F: Traffic Psychology and Behaviour, 21*, 75-89. doi:10.1016/j.trf.2013.09.002

Frenkel, S., Alba, D., & Zhong, R. (2020, March 8,). Surge of virus misinformation stumps Facebook and Twitter. *The New York Times* Retrieved from https://www.nytimes.com/2020/03/08/technology/coronavirus-misinformation-social-media.html

Garson, G. D. (2021). *Data analytics for the social sciences: Applications in R* (First ed.). London, England: Routledge. doi:10.4324/9781003109396

Gehrke, S. R., & Reardon, T. G. (2022). Patterns and Predictors of early Electric Vehicle Adoption in Massachusetts. *International Journal of Sustainable Transportation*, *16*(6), 514-525. doi:10.1080/15568318.2021.1912223

Giles, S., Errickson, D., Harrison, K., & Márquez-Grant, N. (2023). Solving the inverse problem of post-mortem interval estimation using Bayesian Belief Networks. *Forensic Science International, 342*, 111536. doi:10.1016/j.forsciint.2022.111536

Grampurohit, S., & Sunkad, S. Hospital Length of Stay Prediction using Regression Models. Paper presented at the *2020 IEEE International Conference for Innovation in Technology (INOCON)*, doi:10.1109/INOCON50539.2020.9298294

Gutierrez Araiza, J. A., Luna, S., Santiago, I., & Akundi, A. (2024). Perceptions of Electric Vehicle Adoption Through Natural Language Processing. Paper presented at the *2024 IEEE International Systems Conference (SysCon)*. doi:10.1109/SysCon61195.2024.10553625

Haman, M. (2020). The use of twitter by state leaders and its impact on the public during the COVID-19 pandemic. *Heliyon, 6*(11), e05540. doi:10.1016/j.heliyon.2020.e05540

Han, L., Pinson, P., & Kazempour, J. (2022). Trading data for wind power forecasting: A regression market with lasso regularization. *Electric Power Systems Research, 212*, 108442. doi:10.1016/j.epsr.2022.108442

Hassan, M. M., Hassan, M. M., Yasmin, F., Khan, M. A. R., Zaman, S., Galibuzzaman, . . . Bairagi, A. K. (2023). A comparative assessment of machine learning algorithms with the least absolute shrinkage and selection operator for breast cancer detection and prediction. *Decision Analytics Journal, 7*, 100245. doi:10.1016/j.dajour.2023.100245

Hattem, J. (2018). Tesla in autopilot mode sped up before crashing. Retrieved from https://phys.org/news/2018-05-apnewsbreak-tesla-autopilot-sped-utah.html

Hardman, S., Chandan, A., Tal, G., & Turrentine, T. (2017). The Effectiveness of Financial Purchase Incentives for Battery Electric Vehicles – A review of the evidence. *Renewable and Sustainable Energy Reviews, 80*, 1100-1111. doi:10.1016/j.rser.2017.05.255

Hargittai, E. (2018). Potential biases in big data: Omitted voices on social media. *Social Science Computer Review, 38*(1) doi:10.1177/0894439318788322

Higueras-Castillo, E., Guillén, A., Herrera, L., & Liébana-Cabanillas, F. (2021). Adoption of electric vehicles: Which factors are really important? *International Journal of Sustainable Transportation, 15*(10), 799-813. doi:10.1080/15568318.2020.1818330

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased estimation for nonorthogonal problems. *Technometrics, 42*(1), 80-86. doi:10.1080/00401706.2000.10485983

Holley, P. (2018, May 19,). Federal investigators are looking into tesla's autopilot crash in utah. *The Salt Lake Tribune*. Retrieved from https://www.sltrib.com/news/2018/05/19/federal-investigators-are-looking-into-teslas-autopilot-crash-in-utah/

Hou, Q., Han, M., & Cai, Z. (2020). Survey on Data Analysis in Social Media: A practical application aspect. *Big Data Mining and Analytics, 3*(4), 259-279. doi:10.26599/BDMA.2020.9020006

Huang, Y., & Qian, L. (2018). Consumer preferences for Electric Vehicles in lower tier cities of China: Evidences from south Jiangsu region. *Transportation Research Part D: Transport and Environment, 63*, 482-497. doi:10.1016/j.trd.2018.06.017

Huang, X., Wang, S., Zhang, M., Hu, T., Hohl, A., She, B., . . . Li, Z. (2022). Social media mining under the COVID-19 context: Progress, challenges, and opportunities. International *Journal of Applied Earth Observation and Geoinformation, 113*, 102967. doi:10.1016/j.jag.2022.102967

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Paper presented at the *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), doi: /10.1609/icwsm.v8i1.14550

Indiana Department of Transportation (n.d.). *Wireless electric vehicle charging solution for highway infrastructure*. Retrieved from https://www.in.gov/indot/current-programs/innovative-programs/wireless-electric-vehicle-charging-solution-for-highway-infrastructure/#:~:text=INDOT%2C%20Purdue%20to%20Develop%20Wireless,charging%20concrete%20pavement%20highway%20segment.

International Energy Agency. (2024). Electric vehicles. Retrieved from https://www.iea.org/energy-system/transport/electric-vehicles

Jagini, A., Mahajan, K., Aluvathingal, N., Mohan, V., & TR, P. (2023). Twitter sentiment analysis for bitcoin price prediction. Paper presented at the *2023 3rd International Conference on Smart Data Intelligence (ICSMDI)*, 32-37. doi:10.1109/ICSMDI57622.2023.00015

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning: With Applications in Python* (First ed.) Springer Cham. doi:10.1007/978-3-031-38747-0

Javed, F., Thomas, I., & Memedi, M. (2018). A comparison of feature selection methods when using motion sensors data: A case study in Parkinson's disease. Paper presented at the *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, doi:10.1109/EMBC.2018.8513683

Jeena, R. S., & SukeshKumar, A. (2018). Stroke Risk Assessment using Ridge Regression Model. Paper presented at the *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, doi:10.1109/ICOEI.2018.8553960

Jeffreys, H. (1961). The theory of probability. Oxford University Press.

Jingnan, H. (2023, Feb 9,). Twitter's new data access rules will make social media research harder. Retrieved from https://www.npr.org/2023/02/09/1155543369/twitters-new-data-access-rules-will-make-social-media-research-harder

Jitwasinkul, B., Hadikusumo, B. H. W., & Memon, A. Q. (2016). A Bayesian Belief Network model of organizational factors for improving safe work behaviors in Thai construction industry. *Safety Science, 82*, 264-273. doi:10.1016/j.ssci.2015.09.027

Kamis, A., & Susan Abraham, P. (2024). Predictive models of electric vehicle adoption in the United States: Charging ahead with renewable energy. *Transportation Research Interdisciplinary Perspectives, 24*, 101041. doi:10.1016/j.trip.2024.101041

Karmugilan, K., & Pachayappan, M. (2020). Sustainable manufacturing with Green Environment: An evidence from Social Media. *Materials Today: Proceedings*, *22*, 1878-1884. doi:10.1016/j.matpr.2020.03.087

Kaur, S., Kaul, P., & Zadeh, P. M. (2020). Monitoring the dynamics of emotions during COVID-19 using Twitter data. *Procedia Computer Science*, *177*, 423-430. doi:10.1016/j.procs.2020.10.056

Keeney, A. J., Beseler, C. L., & Ingold, S. S. (2023). County-level analysis on occupation and ecological determinants of child abuse and neglect rates employing elastic net regression. *Child Abuse & Neglect, 137*, 106029. doi:10.1016/j.chiabu.2023.106029

Kekkonen, V., Kivimäki, P., Valtonen, H., Hintikka, J., Tolmunen, T., Lehto, S. M., & Laukkanen, E. (2015). Sample selection may bias the outcome of an adolescent mental health survey: Results from a five-year follow-up of 4171 adolescents. *Public Health, 129*(2), 162-172. doi:10.1016/j.puhe.2014.11.015

Kelz, R. R., Airoldi, E. M., & Keele, L. (2021). Strengths and limitations of machine learning in surgical care. *Journal of the American College of Surgeons, 232*(6), 919-920. doi:10.1016/j.jamcollsurg.2021.03.001

Khan, H. A. U., Price, S., Avraam, C., & Dvorkin, Y. (2022). Inequitable access to EV charging infrastructure. *The Electricity Journal*, *35*(3), 107096. doi:10.1016/j.tej.2022.107096

Kotu, V., & Deshpande, B. (2019). Chapter 4 - classification. In V. Kotu, & B. Deshpande (Eds.), *Data science (second edition)* (pp. 65-163) Morgan Kaufmann. doi:10.1016/B978-0-12-814761-0.00004-6

Krishna Kireeti, G. S., Prithvi, J., Divya, M., & Kumari, C. U. (2023). Predicting employability and admission for MS students using ML regression models. Paper presented at the *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, doi:10.1109/I2CT57861.2023.10126208

Kwon, J., Grady, C., Feliciano, J. T., & Fodeh, S. J. (2020). Defining facets of Social distancing during the COVID-19 pandemic: Twitter analysis. *Journal of Biomedical Informatics*, *111*, 103601. doi:10.1016/j.jbi.2020.103601

Lamberti, W. F. (2023). Chapter 3 - An overview of explainable and interpretable AI. In F. A. Batarseh, & L. J. Freeman (Eds.), *AI assurance* (pp. 55-123) Academic Press. doi:10.1016/B978-0-32-391919-7.00015-9

LaMorte, W. W. (2024). The Correlation Coefficient (R). Retrieved from https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module9-Correlation-Regression/PH717-Module9-Correlation-Regression4.html

Landuyt, D., Broekx, S., D'hondt, R., Engelen, G., Aertsens, J., & Goethals, P. L. M. (2013). A review of Bayesian belief networks in ecosystem service modeling. *Environmental Modelling & Software, 46*, 1-11. doi:10.1016/j.envsoft.2013.03.011

Lee, J. (2018, Jun 30). Electric corridor along Utah's I-15 now fully charged. Retrieved from https://www.ksl.com/article/46352587/electric-corridor-along-utahs-i-15-now-fully-charged

Lee, M. D., & Wagenmakers, E. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139087759

Liu, S., & Liu, J. (2021). Public attitudes toward COVID-19 vaccines on English-language Twitter: A Sentiment Analysis. *Vaccine, 39*(39), 5499-5505. doi:10.1016/j.vaccine.2021.08.058

Lin, B., & Wu, W. (2018). Why people want to buy electric vehicle: An empirical study in first-tier cities of China. *Energy Policy, 112*, 233-241. doi:10.1016/j.enpol.2017.10.026

Long, S., Lucey, B., Xie, Y., & Yarovaya, L. (2023). “I just like the stock”: The role of Reddit sentiment in the GameStop share rally. *Financial Review*, 58(1), 19-37. doi:10.1111/fire.12328

Lozada-Medellin, L., Santiago, I., & Sang, Y (2023). Increasing equity in access to Electric Vehicles and Electrified Infrastructure through perceptions, opinions and knowledge of underrepresented communities in the Paso del Norte region. Paper presented at the *2023 ASEE Annual Conference & Exposition*; URL: https://peer.asee.org/43664

Luna, S. (2019). *Exploration of public sentiment as an indicator of public response to natural disasters: An analysis of hurricane scenarios* (Order No. 13882484). Available from ProQuest Dissertations & Theses Global. (2248624288). Retrieved from

https://utep.idm.oclc.org/login?url=https://www.proquest.com/dissertations-theses/exploration-public-sentiment-as-indicator/docview/2248624288/se-2

Luna, S., Guerrero, A., Gonzalez, K., & Akundi, A. (2022). Social media and Pandemic Events: Challenges for alert-warning systems. Paper presented at the *2022 17th Annual System of Systems Engineering Conference (SOSE)*. doi:10.1109/SOSE55472.2022.9812684

Mansour, S. (2018). Social media analysis of user's responses to terrorism using Sentiment Analysis and text mining. *Procedia Computer Science, 140,* 95-103. doi:10.1016/j.procs.2018.10.297

Marshall, A., Tang, L., & Milne, A. (2010). Variable reduction, sample selection bias and bank retail credit scoring. *Journal of Empirical Finance, 17*(3), 501-512. doi:10.1016/j.jempfin.2009.12.003

McKinney, W. (2010). Data structures for statistical computing in python. Paper presented at the *Proceedings of the 9th Python in Science Conference*; 56-61. doi:10.25080/Majora-92bf1922-00a Retrieved from http://conference.scipy.org.s3-website-us-east-1.amazonaws.com/proceedings/scipy2010/mckinney.html

McNair, D. S. (2018). Introductory Chapter: Timeliness of Advantages of Bayesian Networks. IntechOpen. doi: 10.5772/intechopen.83607

McNaughton, A. (2018, Jan 17). Electric vehicle 'fast chargers' installed in kimball junction. From *Park Record.* Retrieved from https://www.parkrecord.com/news/summit-county/electric-vehicle-fast-chargers-installed-in-kimball-junction/

Melkumova, L. E., & Shatskikh, S. Y. (2017). Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering, 201,* 746-755. doi:10.1016/j.proeng.2017.09.615

Min, Y., Lee, H. W., & Hurvitz, P. M. (2023). Clean Energy Justice: Different adoption characteristics of underserved communities in rooftop solar and electric vehicle chargers in Seattle. *Energy Research & Social Science, 96,* 102931. doi:10.1016/j.erss.2022.102931

Minitab, LLC. (2023). MINITAB (Version 21.1) [Computer software]. Minitab, LLC. https://www.minitab.com

Mitchell, A., Shearer, E., & Stocking, G. R. (2021, November 15). News on Twitter: Consumed by Most Users and Trusted by Many. *Pew Research Center*. https://www.pewresearch.org/journalism/2021/11/15/news-on-twitter-consumed-by-most-users-and-trusted-by-many/

Mo, Y. K., Hahn, M. W., & Smith, M. L. (2024). Applications of machine learning in phylogenetics. *Molecular Phylogenetics and Evolution, 196,* 108066. doi:10.1016/j.ympev.2024.108066

Monselise, M., Chang, C., Ferreira, G., Yang, R., & Yang, C. C. (2021). Topics and Sentiments of Public Concerns Regarding COVID-19 Vaccines: Social media Trend Analysis. *J Med Internet Res, 23*(10), e30765. doi:10.2196/30765

Monteiro, S. d. N., Pereira, A. A., Freitas, C. S., Serrão, G. X., de Sousa, M. A. P., Lima, A. C. S., . . . Lourenco-Junior, J. d. B. (2024). Machine learning regression algorithms for predicting muscle, bone, carcass fat and commercial cuts in hairless lambs. *Small Ruminant Research, 236,* 107290. doi:10.1016/j.smallrumres.2024.107290

Morgan, K. (2023, November 8). Three big reasons Americans haven't rapidly adopted EVs. Retrieved from https://www.bbc.com/worklife/article/20231108-three-big-reasons-americans-havent-rapidly-adopted-evs

Moua, Y., Roux, E., Seyler, F., & Briolant, S. (2020). Correcting the effect of sampling bias in species distribution modeling – A new method in the case of a low number of presence data. *Ecological Informatics, 57*, 101086. doi:10.1016/j.ecoinf.2020.101086

Xu, M., & Lin, B. (2023). Accessing people's attitudes towards

Mpoi, G., Milioti, C., & Mitropoulos, L. (2023). Factors and incentives that affect electric vehicle adoption in Greece. *International Journal of Transportation Science and Technology, 12*(4), 1064-1079. doi:10.1016/j.ijtst.2023.01.002

Mullen, L., Blevins, C. & Schmidt, B. (2022). Gender: Predict Gender from Names Using Historical Data. R package version 0.6.0. Retrieved from https://cran.r-project.org/web/packages/gender/gender.pdf

Muthukrishnan, R., & Rohini, R. (2016). LASSO: A feature selection technique in predictive modeling for machine learning. Paper presented at the *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, doi:10.1109/ICACA.2016.7887916

Nandurkar, T., Nagare, S., Hake, S., & Chinnaiah, K. (2023). Sentiment Analysis towards Russia - Ukrainian conflict: Analysis of comments on Reddit. Paper presented at the *2023 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP)*. doi:10.1109/ICETET-SIP58143.2023.10151571

Neppalli, V. K., Caragea, C., Squicciarini, A., Tapia, A., & Stehle, S. (2017). Sentiment Analysis during Hurricane Sandy in emergency response. *International Journal of Disaster Risk Reduction, 21*, 213-222. doi:10.1016/j.ijdrr.2016.12.011

Neubauer, J., & Wood, E. (2014). The impact of range anxiety and home, workplace, and public charging infrastructure on simulated battery electric vehicle lifetime utility. *Journal of Power Sources, 257*, 12-20. doi:10.1016/j.jpowsour.2014.01.075

NSF-ERC ASPIRE. (2023). ASPIRE: 2023 annual report. Retrieved from https://app.box.com/file/1325735913046?s=d76bpuicp08k4e43j0tsz8wt6s5vb73h

Nugroho, D. K. (2021). US Presidential Election 2020 prediction based on Twitter data using lexicon-based Sentiment Analysis. Paper presented at the *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. doi:10.1109/Confluence51648.2021.9377201

Odabas, M. (2022, May 5). 10 facts about Americans and Twitter. *Pew Research Center*. Retrieved on November 25, 2022, from https://www.pewresearch.org/fact-tank/2022/05/05/10-facts-about- americans-and-twitter/

Oxner, R. (2022). Texas plans to place charging stations for electric cars every 50 miles on most interstates. From *The Texas Tribune*. Retrieved from https://www.texastribune.org/2022/06/20/texas-electric-vehicle-charging-stations/

Pamidimukkala, A., Kermanshachi, S., Rosenberger, J. M., & Hladik, G. (2024). Barriers and motivators to the Adoption of Electric Vehicles: A global review. *Green Energy and Intelligent Transportation, 3*(2), 100153. doi:10.1016/j.geits.2024.100153

Park, C. W., & Seo, D. R. (2018). Sentiment analysis of twitter corpus related to artificial intelligence assistants. Paper presented at the *2018 5th International Conference on Industrial Engineering and Applications (ICIEA)*, 495-498. doi:10.1109/IEA.2018.8387151

Parra Aramburo, R. F., Lellis Moreira, M. Â, Lopes Fávero, L. P., De Araújo Costa, I. P., & dos Santos, M. (2022). Data Science in Social Politics with particular emphasis on Sentiment Analysis. *Procedia Computer Science, 214*, 420-427. doi:10.1016/j.procs.2022.11.194

Pearl, J. (1982). Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach. Probabilistic and Causal Inference, 129-138. doi:10.1145/3501714.3501727

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal Inference in Statistics: A primer*. West Sussex, United Kingdom: Wiley.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research, 12*, 2825-2830. doi:10.5555/1953048.2078195

Peng, R., Tang, J. H. C. G., Yang, X., Meng, M., Zhang, J., & Zhuge, C. (2024). Investigating the factors influencing the electric vehicle market share: A comparative study of the European union and United States. *Applied Energy, 355*, 122327. doi:10.1016/j.apenergy.2023.122327

Pereira, J. M., Basto, M., & Silva, A. F. d. (2016). The Logistic Lasso and Ridge Regression in Predicting Corporate Failure. *Procedia Economics and Finance, 39*, 634-641. doi:10.1016/S2212-5671(16)30310-0

Pew Research Center. (2021). Demographics of social media users and adoption in the United States. Retrieved from https://www.pewresearch.org/internet/fact-sheet/social-media/#panel-b14b718d-7ab6-46f4-b447-0abd510f4180

Pierce, B. (2023). Wireless EV charging is the perfect solution for electric public transportation bus fleets. Retrieved from https://www.inductev.com/resource-archive/wireless-ev-charging-is-the-perfect-solution-for-electric-public-transportation-bus-fleets

Plananska, J., Wüstenhagen, R., & de Bellis, E. (2023). Perceived lack of masculinity as a barrier to adoption of electric cars? an empirical investigation of gender associations with low-carbon vehicles. *Travel Behaviour and Society, 32*, 100593. doi:10.1016/j.tbs.2023.100593

Popovich, N. (2024, March 6). Where Electric Vehicles are (and aren't) taking off across the U.S. *The New York Times* .Retrieved from https://www.nytimes.com/interactive/2024/03/06/climate/hybrid-electric-vehicle-popular.html

Pskowski, M. (2022, Mar 22,). Electric school buses coming to el paso? EPA briefs districts on federal funding. From *El Paso Times.* Retrieved from https://www.elpasotimes.com/story/news/2022/03/22/electric-school-buses-el-paso-epa-episd-yisd-sisd/9454059002/

Pulido, C. M., Ruiz-Eugenio, L., Redondo-Sama, G., & Villarejo-Carballido, B. (2020). A new application of social impact in social media for overcoming fake news in Health. *International Journal of Environmental Research and Public Health, 17*(7) doi:10.3390/ijerph17072430

Quammie, M., & Hosein, P. (2024). School climate factors as predictors of school performance: A machine learning approach. Paper presented at the *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*, doi:10.1109/ICETSIS61505.2024.10459425

Rahman, M. S., & Reza, H. (2022). A systematic review towards Big Data Analytics in social media. *Big Data Mining and Analytics, 5*(3), 228-244. doi:10.26599/BDMA.2022.9020009

Reboredo, J. C., & Ugolini, A. (2018). The impact of Twitter sentiment on renewable energy stocks. *Energy Economics, 76*, 153-169. doi:10.1016/j.eneco.2018.10.014

Romadhon, M. R., & Kurniawan, F. (2021). A Comparison of Naive Bayes Methods, Logistic Regression and KNN for Predicting Healing of Covid-19 Patients in Indonesia. Paper presented at the *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, 41-44. doi:10.1109/EIConCIT50028.2021.9431845

Rouder, J. N., & Morey, R. D. (2019). Teaching Bayes' Theorem: Strength of Evidence as Predictive Accuracy. *The American Statistician, 73*(2), 186-190. doi:10.1080/00031305.2017.1341334

Ruan, T., & Lv, Q. (2022). Public perception of electric vehicles on Reddit over the past decade. *Communications in Transportation Research, 2*, 100070. doi:10.1016/j.commtr.2022.100070

Ruan, T., & Lv, Q. (2023). Public perception of electric vehicles on Reddit and Twitter: A cross-platform analysis. *Transportation Research Interdisciplinary Perspectives, 21*, 100872. doi:10.1016/j.trip.2023.100872

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206-215. doi:10.1038/s42256-019-0048-x

Ruoso, A. C., & Duarte Ribeiro, J. L. (2022). The influence of countries' socioeconomic characteristics on the adoption of Electric Vehicle. *Energy for Sustainable Development, 71*, 251-262. doi:10.1016/j.esd.2022.10.003

Rye, J., & Sintov, N. D. (2024). Predictors of electric vehicle adoption intent in rideshare drivers relative to commuters. *Transportation Research Part A: Policy and Practice, 179*, 103943. doi:10.1016/j.tra.2023.103943

Sanewal, N., & Khanna, V. (2023). Solar Power Prediction in North India Using Different Regression Models. Paper presented at the *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)*, doi:10.1109/AIC57670.2023.10263912

Sarang, P. (2023). Naive bayes. *Thinking data science: A data science practitioner's guide* (pp. 143-152). Cham: Springer International Publishing. doi:10.1007/978-3-031-02363-7_7

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science, 181*, 526-534. doi:10.1016/j.procs.2021.01.199

Semanjski, I. C. (2023). Chapter 3: The new challenge of smart urban mobility. In I. C. Semanjski (Ed.), *Smart Urban Mobility* (pp. 25-78) Elsevier. doi:10.1016/B978-0-12-820717-8.00011-7.

Semakula, H. M., Song, G., Achuu, S. P., & Zhang, S. (2016). A Bayesian belief network modelling of household factors influencing the risk of malaria: A study of parasitaemia in children under five years of age in sub-Saharan Africa. *Environmental Modelling & Software, 75*, 59-67. doi:10.1016/j.envsoft.2015.10.006

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379-423. doi:10.1002/j.1538-7305.1948.tb01338.x

Shiomi, Y., Toriumi, A., & Nakamura, H. (2022). International analysis on social and personal determinants of traffic violations and accidents employing logistic regression with elastic net regularization. *IATSS Research, 46*(1), 36-45. doi:10.1016/j.iatssr.2021.12.004

Singh, V., Singh, H., Dhiman, B., Kumar, N., & Singh, T. (2023). Analyzing bibliometric and thematic patterns in the transition to sustainable transportation: Uncovering the influences on electric vehicle adoption. *Research in Transportation Business & Management, 50*, 101033. doi:10.1016/j.rtbm.2023.101033

Spencer, A., Ross, S. & Tyson, A. (2023). How Americans view Electric Vehicles. Retrieved from https://www.pewresearch.org/short-reads/2023/07/13/how-americans-view-electric-vehicles/

Srisa-An, C. (2021). Guideline of Collinearity - Avoidable Regression Models on Time-series Analysis. Paper presented at the *2021 2nd International Conference on Big Data Analytics and Practices (IBDAP)*, doi:10.1109/IBDAP52511.2021.9552165

Stajić, D., Pfeifer, A., Herc, L., & Logonder, M. (2023). Early adoption of battery Electric Vehicles and owners' motivation. *Cleaner Engineering and Technology, 15*, 100658. doi:10.1016/j.clet.2023.100658

Stall, S. (2017, Nov 8,). Indiana's battery industry, undeterred by past setbacks, sees bright future. From *Indianapolis Business Journal.* Retrieved from https://www.ibj.com/articles/66200-charged-up-for-growth

Strong, M. (2020, Feb 27,). Utah proposing $50M to create a statewide EV charging network. From *The Detroit Bureau.* Retrieved from https://www.thedetroitbureau.com/2020/02/utah-proposing-50m-to-create-a-statewide-ev-charging-network/

Sullivan, M. (2022, Jul 15). Purdue researchers are developing pavement that can charge electric vehicles as they drive. From *Fox 59 News*. Retrieved from https://fox59.com/indiana-news/purdue-researchers-are-developing-pavement-that-can-charge-electric-vehicles-as-they-drive/

Suvorova, A. (2022). Interpretable Machine Learning in Social Sciences: Use Cases and Limitations. Paper presented at the *Digital Transformation and Global Society*, 319-331. doi:10.1007/978-3-030-93715-7_23

Szczygielski, J. J., Charteris, A., Bwanya, P. R., & Brzeszczyński, J. (2023). Which COVID-19 information really impacts stock markets? *Journal of International Financial Markets, Institutions and Money, 84*, 101592. doi:10.1016/j.intfin.2022.101592

Szczygielski, J. J., Charteris, A., Bwanya, P. R., & Brzeszczyński, J. (2024). Google search trends and stock markets: Sentiment, attention or uncertainty? *International Review of Financial Analysis, 91*, 102549. doi:10.1016/j.irfa.2023.102549doi:10.1016/j.clet.2023.100658

The Salt Lake City Tribune. (2018, Jul 3,). SLC airport adds 24 charging stations for electric cars. Retrieved from https://www.sltrib.com/news/politics/2018/07/03/slc-airport-adds/

The White House. (2023). Biden-Harris administration announces new private and public sector investments for affordable electric vehicles. Retrieved from https://www.whitehouse.gov/briefing-room/statements-releases/2023/04/17/fact-sheet-biden-harris-administration-announces-new-private-and-public-sector-investments-for-affordable-electric-vehicles/

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288. doi:10.1111/j.2517-6161.1996.tb02080.x

Tomaschek, F., Hendrix, P., & Baayen, R. H. (2018). Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics, 71*, 249-267. doi:10.1016/j.wocn.2018.09.004

Torres, A. (2023, Jul 3,). New texas laws aimed at sharp rise in electric vehicle ownership in lone star state. From *The Dallas Morning News*. Retrieved from https://www.dallasnews.com/news/politics/2023/07/03/new-texas-laws-aimed-at-sharp-rise-in-electric-vehicle-ownership-in-lone-star-state/

Tweepy. Available at: https://www.tweepy.org/

Twitter. (2022). Twitter API. Academic Research access. *Twitter Developer Platform*. Retrieved November 25, 2022, from https://developer.twitter.com/en/products/twitter-api/academic-research

U.S. Census Bureau. (2018). *2020 Census Operational Plan: A New Design for the 21st Century*. Retrieved November 25, 2022, from https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/2020-oper-plan4.pdf

U.S. Department of Energy. (2022). Electric Vehicle Charging Station Locations. *U.S. Department of Energy. Alternative Fuels Data Center.* https://afdc.energy.gov/fuels/electricity_locations.html#/find/nearest?fuel=ELEC

U.S. Environmental Protection Agency. (2023). Fast facts: U.S. Transportation sector Greenhouse Gas Emissions 1990-2021. Retrieved from https://www.epa.gov/system/files/documents/2023-06/420f23016.pdf

Utah State University. (2024). NSF ERC ASPIRE overview. Retrieved from https://aspire.usu.edu/about/overview/

UTEP Communications. (2020, Aug 4,). UTEP partnership to advance electric transportation and sustainability on a global scale. Retrieved from https://www.utep.edu/newsfeed/campus/utep-partnership-to-advance-electric-transportation-and-sustainability-on-a-global-scale.html

Wayland, M. (2024, Mar 13,). EV euphoria is dead. automakers are scaling back or delaying their electric vehicle plans. Retrieved from https://www.cnbc.com/2024/03/13/ev-euphoria-is-dead-automakers-trumpet-consumer-choice-in-us.html

Wenando, F. A., Hayami, R., Bakaruddin, & Novermahakim, A. Y. (2020). Tweet Sentiment Analysis for 2019 Indonesia presidential election results using various classification algorithms. Paper presented at the *2020 1st International Conference on Information Technology, Advanced Mechanical and Electrical Engineering (ICITAMEE)*. doi:10.1109/ICITAMEE50454.2020.9398513

Wooldridge, S. (2003). *Bayesian belief networks*. From https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=2e9ba68462bf70a5aef2ffdd6c359e49b2fcbfcd

Wu, Y., Burnside, E. S., Cox, J., Fan, J., Yuan, M., Yin, J., . . . Craven, M. (2017). Breast cancer risk prediction using electronic health records. Paper presented at the *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, doi:10.1109/ICHI.2017.62

Wu, R., Kang, D., Chen, Y., & Chen, C. (2023). Assessing academic impacts of machine learning applications on a social science: Bibliometric evidence from economics. *Journal of Informetrics, 17*(3), 101436. doi:10.1016/j.joi.2023.101436

Xu, M., & Lin, B. (2023). Accessing people's attitudes towards garbage incineration power plants: Evidence from models correcting sample selection bias. *Environmental Impact Assessment Review, 99*, 107034. doi:10.1016/j.eiar.2022.107034

Yang, A., Liu, C., Yang, D., & Lu, C. (2023). Electric Vehicle Adoption in a mature market: A case study of Norway. *Journal of Transport Geography, 106*, 103489. doi:10.1016/j.jtrangeo.2022.103489

Yang, J., Xiu, P., Sun, L., Ying, L., & Muthu, B. (2022). Social media Data Analytics for Business decision-making system to competitive analysis. *Information Processing & Management*, *59*(1), 102751. doi:10.1016/j.ipm.2021.102751

Yang, X., & Wen, W. (2018). Ridge and lasso regression models for cross-version defect prediction. *IEEE Transactions on Reliability, 67*(3), 885-896. doi:10.1109/TR.2018.2847353

Yeom, K., & Choi, H. (2018). Prediction of manufacturing plant's electric power using machine learning. Paper presented at the *2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN)*, doi:10.1109/ICUFN.2018.8436973

Zaleznik, A. (1984). The “Hawthorne effect”. Retrieved from https://www.library.hbs.edu/hc/hawthorne/09.html#fnl11

Zhang, J., Yue, H., Wu, X., & Chen, W. (2019). A brief review of Bayesian belief network. Paper presented at the *2019 Chinese Control and Decision Conference (CCDC)*, doi:10.1109/CCDC.2019.8832649

Zhao, S., Xie, T., Ai, X., Yang, G., & Zhang, X. (2023). Correcting sample selection bias with model averaging for consumer demand forecasting. *Economic Modelling, 123*, 106275. doi:10.1016/j.econmod.2023.106275

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2) doi:10.1111/j.1467-9868.2005.00503.x

Zulfiker, M. S., Kabir, N., Biswas, A. A., Zulfiker, S., & Uddin, M. S. (2022). Analyzing the public sentiment on COVID-19 vaccination in social media: Bangladesh context. *Array, 15*, 100204. doi:10.1016/j.array.2022.100204

Glossary

Bias: It is a systematic tendency in which the methods used to gather data and generate statistics present an inaccurate, skewed, or biased depiction of reality.

Endogenous variable: Variable whose measure is determined by the model.

Exogenous variable: Variable whose measure is determined outside of the model and is imposed on it.

Directed Acyclic Graph (DAG): In the **Bayesian Causal Inference** context, it **is a** Graphic Model which consist of a set of nodes representing the variables in U (Exogenous veriables) & V (Endogenous variables) and a set of edges between the nodes representing the functions in f.

Explainability: Property of an element that allows its mechanisms to be explicitly described, understood, and studied.

Interpretability: It is the characteristic of an element to have concrete physical meaning.

MAE: Mean Absolute Error

Multicollinearity: In the statistics context, it refers to when two or more variables have a high correlation between each other.

RMSE: Root Mean Squared Error

Variance: Defined as the spread or dispersion within a dataset.

Appendix

APPENDIX 1. DATASET VARIABLES NOMENCLATURE FOR THE PRESCRIPTIVE ANALYSIS

Area	Sub-Field	Variables (Proportion on city per year)	Dataset Nomenclature
Education (3)	Educational Attainment (S1501)	Less than HS Graduate	S1501_LHSG
		HS Graduate	S1501_HSG
		Higher than HS Degree	S1501_HHSG
Employment (18)	Employment Status (S2301)	Labor Force Participation Rate	S2301_LFPR
		Employment Rate	S2301_ESER
	Means of transportation to Work by Travel time to work (B08134)	Car, Truck, or van	B08134_CTV
		Public Transportation	B08134_PT
		Walked	B08134_W
		Taxicab, motorcycle, bicycle, or other means	B08134_TMBO
	Travel time to work (B08303)	<5 minutes	B08303_L5
		5-9 minutes	B08303_5T9
		10-14 minutes	B08303_10T14
		15-19 minutes	B08303_15T19
		20-24 minutes	B08303_20T24
		25-29 minutes	B08303_25T29
		30-34 minutes	B08303_30T34
		35-39 minutes	B08303_35T39
		40-44 minutes	B08303_40T44
		45-59 minutes	B08303_45T59
		60-89 minutes	B08303_60T89
		>90 minutes	B08303_M90
Families and Living Arrangements (4)	Marital Status (S1201)	Now Married (except separated)	S1201_NowM
		Widowed	S1201_W
		Divorced	S1201_D
		Never Married	S1201_NevM
Health (1)	Disability Characteristics (S1810)	With a Disability (Different Capability)	S1810_WDC
Housing (28)	Occupancy Characteristics (S2501)	Owner Occupied Unit	S2501_OOU
		Renter Occupied Unit	S2501_ROU
		1-person household	S2501_1PH
		2-person household	S2501_2PH
		3-person household	S2501_3PH

		4-or-more-person household	S2501_4MPH
	Financial Characteristics (S2503)	Monthly Housing Costs_<300	S2503_MHC_L300
		Monthly Housing Costs_300-499	S2503_MHC_300499
		Monthly Housing Costs_<500-799	S2503_MHC_500799
		Monthly Housing Costs_<800-999	S2503_MHC_800999
		Monthly Housing Costs_<1000-1499	S2503_MHC_1K1.49K
		Monthly Housing Costs_<1500-1999	S2503_MHC_1.5K1.9K
		Monthly Housing Costs_<2000-2499	S2503_MHC_2K2.49K
		Monthly Housing Costs_<2500-2999	S2503_MHC_2.5K2.99K
		Monthly Housing Costs_<3000<=	S2503_MHC_3KM
		No Cash Rent	S2503_MHC_NCR
	Physical Housing Characteristics for occupied units (S2504).	YearStructureBuilt_2020 or later	S2504_YSB_2020L
		YearStructureBuilt_2000-2019	S2504_YSB_20002019
		YearStructureBuilt_1980-1999	S2504_YSB_19801999
		YearStructureBuilt_1960-1979	S2504_YSB_19601979
		YearStructureBuilt_1940-1959	S2504_YSB_19401959
		YearStructureBuilt_1939 or earlier	S2504_YSB_1939E
		VehiclesAvailable_NoVehicles	S2504_VA_0
		VehiclesAvailable_1 Vehicle	S2504_VA_1
		VehiclesAvailable_2 Vehicles	S2504_VA_2
		VehiclesAvailable_3 or more Vehicles	S2504_VA_3M
		Telephone Service Available	S2504_VA_TSA
	Financial Characteristics for Housing Units with a Mortgage (S2506)	Owner Occupied Housing Unit with Mortgage	S2506_OHUM
Income and Poverty (10)	Income in the last 12	Income_<10K USD	S1901_I_<10K
		Income_10K-14.999K USD	S1901_I_10KT14.9K

	months (S1901)	Income_15K-24.999 USD	S1901_I_15KT24.9K
		Income_25K-34.999K USD	S1901_I_25KT34.9K
		Income_<35K-49.999K USD	S1901_I_35KT49.9K
		Income_50K-74.999K USD	S1901_I_50KT74.9K
		Income_75K-99.999K USD	S1901_I_75KT99.9K
		Income_100K-149.999K USD	S1901_I_100KT149.9K
		Income_150K-199.999K USD	S1901_I_150KT199.9K
		Income_>200000 USD	S1901_I_>200K
Population and People (13)	Age and Sex (S0101)	Childhood Stage (0-17 years old)	S0101_CS_017
		Early Adulthood (18-34 years old)	S0101_EA_1834
		Middle Age (35-64 years old)	S0101_MA_3564
		Late Adulthood (65 years old and over)	S0101_LA_65M
	Language spoken at home (S1601)	English	S1601_LS_E
		Spanish	S1601_LS_S
		Other Indo-European Language	S1601_LS_IE
		Asian and Pacific Languages	S1601_LS_AP
		Other Languages	S1601_LS_O
	Sex by Age (B01001)	Male	B01001_G_M
		Female	B01001_G_F
		Average Proportion of Population per Zip Code	Pop_Proportion
	Type of Computer and Internet subscription (S2801)	Has Internet Availability	S2801_IA
Race and Ethnicity (8)	Racial group (B02001).	White alone	B02001_WA
		Black or African American alone	B02001_BAA
		American Indian and Alaska Native alone	B02001_AIAN
		Asian alone	B02001_AA
		Native Hawaiian and Other Pacific Islander alone	B02001_NWOPI

	Ethnicity (B03001)	Some other race alone Two or more races Hispanic or Latino	B02001_SOR B02001_>2RG B03001_HL
Charging Infrastruc-ture (2)		Average Charging Station Availability on Year per Zip Code Average number of new Charging Stations openings on the Year per Zip Code	CSA ONCS
Observed Social Media Variables (3)		Gender of user who posted social media publication (Male, Female) Year when the social media user posted social media publication (2016, 2017,…,2021, 2022) City where the social media user posted social media publication (Indianapolis, Salt Lake City, El Paso)	GENDER YEAR CITY
Social Media Post Sentiment		Numerical value of the sentiment compound score given by VADER tool to the Social Media post Sentiment polarity category obtained based on the SENTIMENT variable (Positive (SENTIMENT>0), Neutral (SENTIMENT=0), and Negative (SENTIMENT<0)	SENTIMENT SENTIMENT_CATEGORY

APPENDIX 2. MONTHLY SENTIMENT AVERAGE AND MONTHLY SOCIAL MEDIA USER COUNT PER CITY BY YEAR BY MONTH

		Indianapolis, IN		El Paso, TX		Salt Lake City, UT	
Year	Month	Count	Average Sentiment	Count	Average Sentiment	Count	Average Sentiment
2016	January	13	0.2215	-	-	-	-
2016	February	12	0.1229	-	-	-	-
2016	March	7	0.1536	-	-	-	-
2016	April	7	0.3021	-	-	-	-
2016	May	9	-0.0389	-	-	-	-
2016	June	16	0.2029	-	-	-	-
2016	July	10	0.1656	-	-	-	-
2016	August	2	-0.1250	-	-	-	-
2016	September	8	0.1379	-	-	-	-
2016	October	16	0.0375	1	0.1280	-	-
2016	November	8	0.1839	3	0.0000	-	-
2016	December	11	-0.0644	-	-	-	-
2017	January	6	0.2820	-	-	-	-
2017	February	8	-0.0593	-	-	-	-
2017	March	13	0.0894	-	-	-	-
2017	April	6	0.3031	-	-	-	-
2017	May	16	0.0996	-	-	-	-
2017	June	9	-0.0155	-	-	-	-
2017	July	13	-0.0888	-	-	-	-
2017	August	17	0.1456	-	-	-	-
2017	September	5	0.3260	1	-0.1779	-	-
2017	October	18	0.1781	6	0.2369	-	-
2017	November	17	0.1739	14	-0.0052	-	-
2017	December	15	0.0621	4	0.0000	-	-
2018	January	21	0.1362	3	-0.2036	1	0.2500
2018	February	68	0.0344	4	0.4207	1	0.0000
2018	March	94	0.1283	3	-0.3533	2	-0.1139
2018	April	53	0.1299	-	-	3	-0.0257
2018	May	33	0.1520	2	0.3056	3	0.1681
2018	June	39	0.0646	8	0.2201	5	-0.2348
2018	July	25	0.1920	3	0.2956	4	0.1018
2018	August	41	0.0861	11	0.0249	4	-0.1929
2018	September	36	0.1381	4	0.1233	5	0.2433
2018	October	48	0.1117	2	0.1671	3	0.4374
2018	November	54	0.2216	4	0.2110	6	0.0568
2018	December	51	0.2102	5	0.1174	6	0.1190

2019	January	22	0.1043	7	0.2775	3	0.0833
2019	February	48	0.0397	7	0.1638	2	0.4599
2019	March	26	0.0932	6	0.0503	-	-
2019	April	30	0.0603	8	0.1644	4	0.3284
2019	May	35	0.1119	-	-	2	0.0000
2019	June	25	0.2412	10	0.0514	5	0.0918
2019	July	45	0.1590	3	0.0824	2	-0.3042
2019	August	37	0.2016	5	0.3144	6	0.3651
2019	September	34	0.0175	9	0.0227	2	0.1012
2019	October	13	0.1668	10	0.3305	2	0.0000
2019	November	9	0.0965	15	0.2812	5	0.2185
2019	December	8	-0.0420	18	0.1405	5	0.1721
2020	January	40	0.2517	7	0.1837	1	0.5574
2020	February	39	0.1267	4	-0.0115	2	0.2465
2020	March	36	0.2216	5	0.1376	4	0.1817
2020	April	29	0.2320	-	-	-	-
2020	May	37	0.1832	14	0.0894	3	0.0000
2020	June	27	0.1534	7	0.1691	2	0.3094
2020	July	23	0.0451	8	0.1686	5	-0.3578
2020	August	24	0.0775	4	0.0000	1	0.0000
2020	September	24	0.1951	7	0.0093	-	-
2020	October	11	0.2547	4	0.1681	-	-
2020	November	15	0.1761	5	0.5290	3	0.1808
2020	December	22	0.0834	4	-0.0625	-	-
2021	January	40	0.2517	4	0.2483	-	-
2021	February	39	0.1267	-	-	-	-
2021	March	36	0.2216	-	-	-	-
2021	April	29	0.2320	-	-	-	-
2021	May	37	0.1832	2	-0.0270	-	-
2021	June	27	0.1534	3	-0.0136	-	-
2021	July	23	0.0451	2	0.2379	-	-
2021	August	24	0.0775	-	-	-	-
2021	September	24	0.1951	1	0.6037	-	-
2021	October	11	0.2547	2	0.0129	-	-
2021	November	15	0.1761	-	-	-	-
2021	December	22	0.0834	2	0.3715	-	-
2022	January	10	-0.2033	4	0.0903	-	-
2022	February	42	0.0895	2	0.0640	2	0.0000
2022	March	32	0.1841	4	0.1659	-	-
2022	April	32	0.1063	7	0.1426	10	0.0889
2022	May	38	0.1555	7	0.3197	10	0.1442
2022	June	39	0.1645	5	0.0000	7	0.3288
2022	July	41	0.1810	5	0.4565	3	0.3231

2022	August	42	0.1214	6	0.1795	8	0.0744
2022	September	60	0.1178	5	0.1901	13	0.2858

APPENDIX 3. NAÏVE-BAYES RESULTS FROM ASSOCIATING PREDICTOR VARIABLES WITH THE SENTIMENT CATEGORY

Child	Overall Contribution	p-value	Pearson Correlation
S0101_MA_3564	1.8220%	0.2414%	0.0173
S2504_VA_1	1.8133%	0.2522%	-0.0104
S2503_MHC_300499	1.6369%	0.6064%	-0.0269
YEAR	1.6364%	4.4379%	0.0615
S2504_YSB_1939E	1.6078%	0.6991%	-0.0101
S2504_YSB_19801999	1.5681%	0.8477%	-0.0083
S2501_ROU	1.5419%	0.9618%	-0.0513
S2501_OOU	1.5382%	0.9789%	0.0264
B08303_M90	1.5285%	1.0256%	0.0519
B08303_40T44	1.5177%	1.0804%	0.0288
B01001_G_F	1.5068%	1.1380%	0.0146
B08303_5T9	1.4982%	1.1857%	-0.0174
B08134_W	1.4842%	1.2672%	-0.0310
B08134_TMBO	1.4703%	1.3536%	-0.0258
B02001_WA	1.4134%	1.7689%	-0.0373
B08303_25T29	1.4027%	1.8599%	0.0206
S2504_YSB_19601979	1.4002%	0.5435%	-0.0295
S1501_HHSG	1.3946%	1.9314%	0.0061
B02001_AA	1.3946%	1.9314%	-0.0126
B01001_G_M	1.3917%	1.9574%	-0.0202
S1201_W	1.3727%	2.1377%	-0.0074
S2503_MHC_1K1.49K	1.3677%	2.1877%	0.0237
S1601_LS_O	1.3565%	2.3031%	0.0146
B02001_AIAN	1.3470%	2.4058%	-0.0215
B08303_45T59	1.3320%	2.5774%	0.0159
B08134_PT	1.3314%	2.5842%	-0.0413
S1901_I_15KT24.9K	1.3054%	2.9094%	-0.0273
S2503_MHC_3KM	1.2727%	3.3735%	0.0203
B02001_SOR	1.2591%	3.5856%	-0.0070
B08303_15T19	1.2545%	3.6597%	-0.0256
B08303_20T24	1.2359%	3.9760%	-0.0204
Pop_Proportion	1.2232%	1.3636%	-0.0205
S2501_2PH	1.1919%	4.8272%	0.0082
S1901_I_75KT99.9K	1.1914%	4.8382%	0.0205
S1901_I_>200K	1.1753%	5.1913%	0.0355

S2503_MHC_2K2.49K	1.1610%	1.8702%	0.0405
B08012_L5	1.1527%	5.7253%	-0.0098
S2503_MHC_800999	1.1515%	5.7545%	-0.0048
S0101_CS_017	1.1314%	6.2748%	0.0073
S1901_I_100KT149.9K	1.1127%	6.7961%	0.0537
B08303_10T14	1.1079%	6.9373%	-0.0193
S1901_I_<10K	1.1060%	6.9937%	-0.0487
S2501_1PH	1.1019%	7.1147%	0.0049
S2503_MHC_500799	1.1016%	7.1233%	-0.0339
S2503_MHC_L300	1.0966%	7.2776%	-0.0157
S2504_VA_3M	1.0907%	7.4593%	-0.0026
S2504_YSB_20002019	1.0870%	7.5781%	-0.0232
S1201_D	1.0867%	7.5867%	-0.0142
S1501_HSG	1.0832%	2.7606%	0.0264
S2501_4MPH	1.0807%	2.7952%	-0.0118
S1901_I_10KT14.9K	1.0620%	8.4121%	-0.0272
S1901_I_50KT74.9K	1.0599%	8.4883%	0.0080
S1501_LHSG	1.0586%	3.1177%	0.0002
S0101_LA_65M	1.0541%	8.6919%	0.0303
S1901_I_150KT199.9K	1.0385%	9.2724%	0.0461
S2801_IA	1.0385%	0.8679%	0.0274
S0101_EA_1834	1.0359%	3.4837%	-0.0211
S2501_3PH	1.0341%	3.5159%	-0.0061
S2506_OHUM	1.0291%	3.6013%	0.0067
B08303_35T39	1.0291%	3.6013%	0.0175
S2504_VA_2	1.0281%	9.6788%	0.0167
S1201_NowM	1.0177%	10.0964%	0.0024
S2503_MHC_NCR	0.9886%	1.1529%	-0.0305
S1201_NevM	0.9782%	11.8466%	0.0145
S2504_VA_0	0.9457%	13.4763%	-0.0345
S2504_VA_TSA	0.9437%	1.4858%	0.0275
S2504_YSB_19401959	0.9418%	5.4866%	-0.0249
B08303_30T34	0.9215%	14.8117%	0.0185
B08303_60T89	0.9131%	6.2859%	0.0085
S1601_LS_IE	0.9022%	6.6164%	-0.0149
S2301_ESER	0.8991%	1.9088%	0.0035
S2301_LFPR	0.8752%	7.5070%	-0.0029
S1601_LS_AP	0.8752%	7.5070%	-0.0207

B08134_CTV	0.8646%	18.3937%	-0.0115
CSA	0.8590%	18.7818%	0.0090
S1901_I_35KT49.9K	0.8195%	21.7227%	-0.0567
S1601_LS_S	0.7964%	3.3744%	-0.0041
CITY	0.7964%	3.3744%	-0.0093
B03001_HL	0.7964%	3.3744%	-0.0045
B02001_NWOPI	0.7964%	3.3744%	-0.0206
S1601_LS_E	0.7964%	3.3744%	0.0057
B02001_BAA	0.7964%	3.3744%	0.0154
S2503_MHC_1.5K1.9K	0.6466%	38.8995%	0.0281
B02001_>2RG	0.5752%	27.4288%	0.0247
S1901_I_25KT34.9K	0.5463%	51.9866%	-0.0308
ONCS	0.5402%	52.8420%	-0.0258
S2503_MHC_2.5K2.99K	0.4537%	65.3698%	0.0200
S2504_YSB_2020L	0.3493%	59.9278%	-0.0120
GENDER	0.3115%	13.0035%	0.0314
S1810_WDC	0.1498%	92.3121%	0.0105

APPENDIX 4. PROBABILITY DISTRIBUTION TABLES BASED ON BOXPLOT FROM FIGURE 4.22

Results for Probability when Sentiment = Negative

City	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
El Paso	14	0.17952	0.00532	0.01989	0.14770	0.16543	0.17965	0.19290	0.21920
Indianapolis	14	0.21544	0.00623	0.02332	0.17700	0.20050	0.21445	0.22942	0.26280
Salt Lake City	14	0.21206	0.00458	0.01715	0.18660	0.19798	0.21175	0.22703	0.24120

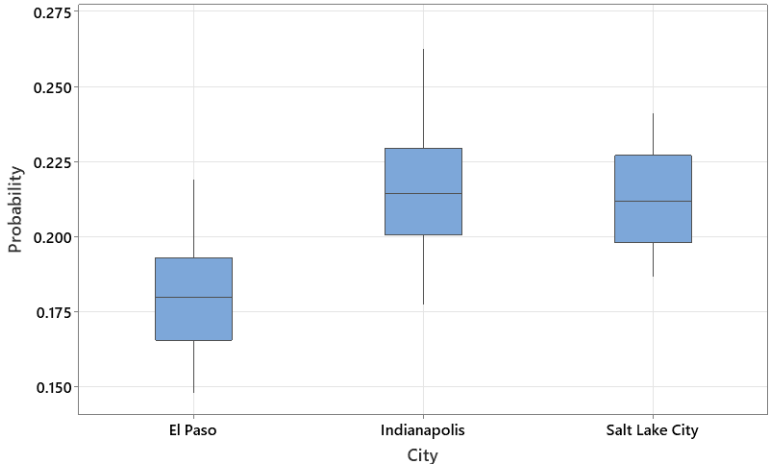


Figure A4.1. Boxplot of probability when the Sentiment is Negative per City

Results for Probability when Sentiment = Neutral

City	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
El Paso	14	0.36219	0.00835	0.01989	0.31910	0.33475	0.35905	0.39017	0.41330
Indianapolis	14	0.29791	0.00623	0.00722	0.26130	0.27365	0.29460	0.32175	0.34130
Salt Lake City	14	0.30349	0.00458	0.00980	0.24240	0.27963	0.30230	0.32580	0.37700

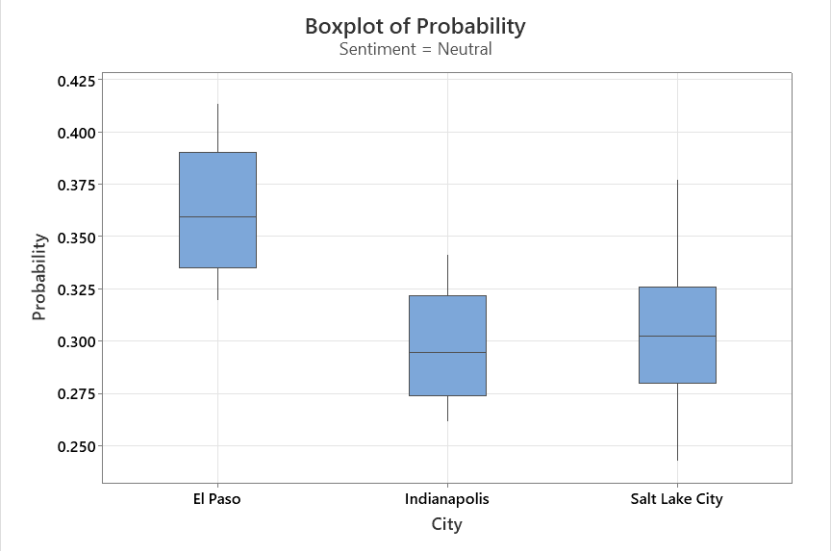


Figure A4.2. Boxplot of probability when the Sentiment is Neutral per City

Results for Probability when Sentiment = Positive

City	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
El Paso	14	0.4583	0.0123	0.0462	0.3675	0.4275	0.35905	0.39017	0.41330
Indianapolis	14	0.4866	0.0123	0.0461	0.3959	0.4567	0.29460	0.32175	0.34130
Salt Lake City	14	0.4845	0.0136	0.0507	0.3818	0.4562	0.30230	0.32580	0.37700

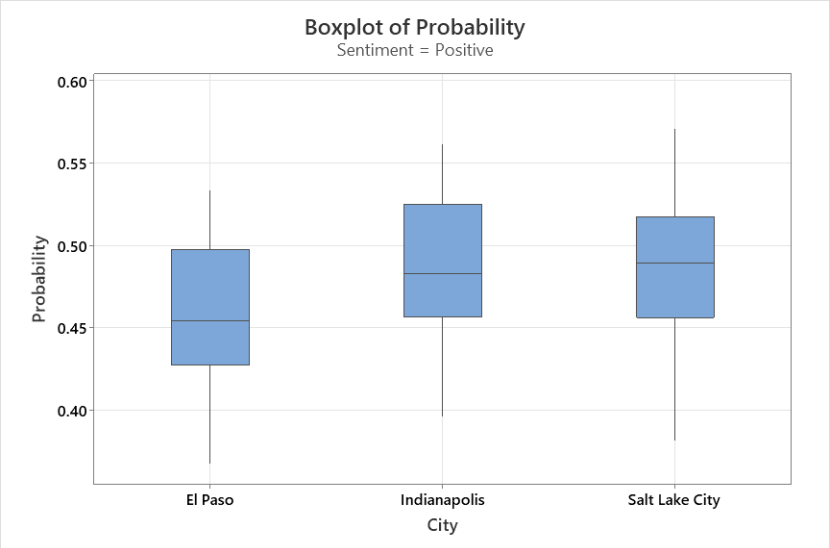


Figure A4.3. Boxplot of probability when the Sentiment is Positive per City

APPENDIX 5. BAYES FACTORS RESULTS OF THE SENTIMENT PROBABILITY WHEN YEAR, CITY AND GENDER ARE BEING CONTROLLED

Sentiment	Year	City	Gender	Probability	Bayes Factor
Negative	2016	El Paso	female	0.2192	1.095428
Negative	2016	El Paso	male	0.1477	0.676193
Negative	2016	Indianapolis	female	0.2628	1.390987
Negative	2016	Indianapolis	male	0.177	0.839182
Negative	2016	Salt Lake City	female	0.2412	1.240318
Negative	2016	Salt Lake City	male	0.1866	0.895139
Negative	2017	El Paso	female	0.2081	1.02538
Negative	2017	El Paso	male	0.1826	0.871664
Negative	2017	Indianapolis	female	0.2487	1.291651
Negative	2017	Indianapolis	male	0.2233	1.121807
Negative	2017	Salt Lake City	female	0.2286	1.156324
Negative	2017	Salt Lake City	male	0.2366	1.209173
Negative	2018	El Paso	female	0.1925	0.930189
Negative	2018	El Paso	male	0.1686	0.791185
Negative	2018	Indianapolis	female	0.2289	1.158292
Negative	2018	Indianapolis	male	0.2047	1.004315
Negative	2018	Salt Lake City	female	0.2108	1.042237
Negative	2018	Salt Lake City	male	0.2165	1.078206
Negative	2019	El Paso	female	0.1817	0.86652
Negative	2019	El Paso	male	0.1731	0.816821
Negative	2019	Indianapolis	female	0.2153	1.07059
Negative	2019	Indianapolis	male	0.2106	1.040984
Negative	2019	Salt Lake City	female	0.1986	0.96697
Negative	2019	Salt Lake City	male	0.2229	1.119222
Negative	2020	El Paso	female	0.1941	0.939782
Negative	2020	El Paso	male	0.1756	0.83103
Negative	2020	Indianapolis	female	0.231	1.17211
Negative	2020	Indianapolis	male	0.214	1.062366
Negative	2020	Salt Lake City	female	0.2127	1.054169
Negative	2020	Salt Lake City	male	0.2265	1.142591
Negative	2021	El Paso	female	0.1779	0.844373
Negative	2021	El Paso	male	0.1559	0.720582
Negative	2021	Indianapolis	female	0.2105	1.040226
Negative	2021	Indianapolis	male	0.1879	0.902818
Negative	2021	Salt Lake City	female	0.1943	0.940984
Negative	2021	Salt Lake City	male	0.1983	0.965148
Negative	2022	El Paso	female	0.1814	0.864666
Negative	2022	El Paso	male	0.1549	0.715198

Negative	2022	Indianapolis	female	0.2149	1.068057
Negative	2022	Indianapolis	male	0.1866	0.895139
Negative	2022	Salt Lake City	female	0.1983	0.965148
Negative	2022	Salt Lake City	male	0.197	0.957149
Neutral	2016	El Paso	female	0.4133	1.589425
Neutral	2016	El Paso	male	0.3191	1.057388
Neutral	2016	Indianapolis	female	0.3413	1.169067
Neutral	2016	Indianapolis	male	0.2613	0.798109
Neutral	2016	Salt Lake City	female	0.377	1.36535
Neutral	2016	Salt Lake City	male	0.2424	0.721911
Neutral	2017	El Paso	female	0.39	1.442532
Neutral	2017	El Paso	male	0.4087	1.559508
Neutral	2017	Indianapolis	female	0.3213	1.068129
Neutral	2017	Indianapolis	male	0.3387	1.1556
Neutral	2017	Salt Lake City	female	0.3554	1.243993
Neutral	2017	Salt Lake City	male	0.3159	1.041735
Neutral	2018	El Paso	female	0.3573	1.254341
Neutral	2018	El Paso	male	0.3727	1.340311
Neutral	2018	Indianapolis	female	0.2931	0.93551
Neutral	2018	Indianapolis	male	0.3076	1.002351
Neutral	2018	Salt Lake City	female	0.325	1.086351
Neutral	2018	Salt Lake City	male	0.2863	0.9051
Neutral	2019	El Paso	female	0.3349	1.136277
Neutral	2019	El Paso	male	0.3842	1.407695
Neutral	2019	Indianapolis	female	0.2738	0.850683
Neutral	2019	Indianapolis	male	0.3176	1.050104
Neutral	2019	Salt Lake City	female	0.3042	0.986428
Neutral	2019	Salt Lake City	male	0.2958	0.947748
Neutral	2020	El Paso	female	0.3608	1.273563
Neutral	2020	El Paso	male	0.3907	1.446544
Neutral	2020	Indianapolis	female	0.2961	0.949114
Neutral	2020	Indianapolis	male	0.3231	1.076969
Neutral	2020	Salt Lake City	female	0.3282	1.102273
Neutral	2020	Salt Lake City	male	0.3011	0.972045
Neutral	2021	El Paso	female	0.327	1.096285
Neutral	2021	El Paso	male	0.3401	1.162662
Neutral	2021	Indianapolis	female	0.267	0.821748
Neutral	2021	Indianapolis	male	0.2794	0.874828
Neutral	2021	Salt Lake City	female	0.2968	0.952304
Neutral	2021	Salt Lake City	male	0.2596	0.791096
Neutral	2022	El Paso	female	0.3343	1.133049
Neutral	2022	El Paso	male	0.3376	1.149934
Neutral	2022	Indianapolis	female	0.2732	0.848118

Neutral	2022	Indianapolis	male	0.2773	0.86573
Neutral	2022	Salt Lake City	female	0.3035	0.983169
Neutral	2022	Salt Lake City	male	0.2576	0.782781
Positive	2016	El Paso	female	0.3675	0.607411
Positive	2016	El Paso	male	0.5332	1.194112
Positive	2016	Indianapolis	female	0.3959	0.685113
Positive	2016	Indianapolis	male	0.5617	1.339735
Positive	2016	Salt Lake City	female	0.3818	0.645643
Positive	2016	Salt Lake City	male	0.571	1.391441
Positive	2017	El Paso	female	0.4019	0.702474
Positive	2017	El Paso	male	0.4087	0.722574
Positive	2017	Indianapolis	female	0.43	0.788641
Positive	2017	Indianapolis	male	0.438	0.814749
Positive	2017	Salt Lake City	female	0.416	0.744674
Positive	2017	Salt Lake City	male	0.4476	0.846922
Positive	2018	El Paso	female	0.4502	0.856025
Positive	2018	El Paso	male	0.4588	0.886077
Positive	2018	Indianapolis	female	0.478	0.957289
Positive	2018	Indianapolis	male	0.4877	0.995209
Positive	2018	Salt Lake City	female	0.4642	0.905708
Positive	2018	Salt Lake City	male	0.4972	1.033765
Positive	2019	El Paso	female	0.4833	0.978021
Positive	2019	El Paso	male	0.4427	0.830436
Positive	2019	Indianapolis	female	0.5109	1.092004
Positive	2019	Indianapolis	male	0.4718	0.933782
Positive	2019	Salt Lake City	female	0.4972	1.033765
Positive	2019	Salt Lake City	male	0.4813	0.970031
Positive	2020	El Paso	female	0.4451	0.83855
Positive	2020	El Paso	male	0.4338	0.800809
Positive	2020	Indianapolis	female	0.4729	0.937912
Positive	2020	Indianapolis	male	0.4629	0.900986
Positive	2020	Salt Lake City	female	0.4591	0.887312
Positive	2020	Salt Lake City	male	0.4724	0.936033
Positive	2021	El Paso	female	0.4951	1.025117
Positive	2021	El Paso	male	0.5041	1.06248
Positive	2021	Indianapolis	female	0.5226	1.144147
Positive	2021	Indianapolis	male	0.5327	1.191716
Positive	2021	Salt Lake City	female	0.5089	1.083299
Positive	2021	Salt Lake City	male	0.5421	1.237641
Positive	2022	El Paso	female	0.4843	0.981755
Positive	2022	El Paso	male	0.5075	1.077248
Positive	2022	Indianapolis	female	0.5119	1.096383
Positive	2022	Indianapolis	male	0.5361	1.208112

Positive	2022	Salt Lake City	female	0.4982	1.037908
Positive	2022	Salt Lake City	male	0.5455	1.254444

Vita

Jesus Alejandro Gutierrez Araiza was born in Delicias and raised in Ciudad Juarez, Mexico in 2000. After he finished his High School studies at the Escuela Preparatoria Central de Ciudad Juarez (EPCCJ), he started his B.S. in Industrial and Systems Engineering major in the University of Texas at El Paso in Fall 2018. Being a member of the University Honors Program (UHP) allowed him to be a recipient of the Fall 2020 Houston Endowment Professional and Leadership Development Fund, graduating two years later in Summer 2022.

Alejandro decided to pursue graduate studies in the Industrial Engineering program, specializing in Smart Manufacturing under the mentorship of Dr. Sergio Luna collaborating with the NSF-ERC ASPIRE Co-Director, Dr. Ivonne Santiago, as a Graduate Research Assistant. Alejandro showcased his research prowess by presenting a research poster and a proceeding paper, on the IISE 2023 and IEEE-SYSCON 2024 annual conferences respectively.

During his two years at Graduate School, he served as Secretary of the Institute of Industrial and Systems Engineers-UTEP Chapter. He was also a Assistant Instructor for the *Data Visualization for Decision-Making* online course. Thanks to his compromise not only to his research but also to service to his academic department, he received the IMSE Outstanding Service Award Summer 2024.

In the Summer of 2023, he expanded his practical experience as a Metrology and Calibration Engineering Intern in the Central Quality Administration of Cummins Inc. at the Rocky Mount Engine Plant (RMEP) in North Carolina. Looking ahead, Alejandro anticipates returning to Cummins Inc. as a Manufacturing Development Program Associate in August 2024, leveraging his academic and professional expertise to contribute to the Industrial Engineering field.
Contact Information: *ise.jagutierrez@gmail.com*