University of Texas at El Paso
ScholarWorks@UTEP

Open Access Theses & Dissertations

2020-01-01

Forecasting Consumer Consumption Behavior Of Water Bottles Using Generlized Linear Model For Supply Chain Resilience

Abdulaziz Alidrees
University of Texas at El Paso

Follow this and additional works at: https://scholarworks.utep.edu/open_etd

 Part of the Engineering Commons

Recommended Citation

Alidrees, Abdulaziz, "Forecasting Consumer Consumption Behavior Of Water Bottles Using Generlized Linear Model For Supply Chain Resilience" (2020). *Open Access Theses & Dissertations*. 3135.
https://scholarworks.utep.edu/open_etd/3135

This is brought to you for free and open access by ScholarWorks@UTEP. It has been accepted for inclusion in Open Access Theses & Dissertations by an authorized administrator of ScholarWorks@UTEP. For more information, please contact lweber@utep.edu.

FORECASTING CONSUMER CONSUMPTION BEHAVIOR OF WATER BOTTLES USING GENERLIZED LINEAR MODEL FOR SUPPLY CHAIN RESILIENCE

ABDULAZIZ ALIDREES

Master's Program in Industrial Engineering

Approved

Jose Espiritu Nolasco, Ph.D. Chair

Eric D Smith, Ph.D.

Methaq S. Abed, Ph.D.

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

Copyright ©

by

Abdulaziz Alidrees

2020

FORECASTING CONSUMER CONSUMPTION BEHAVIOR OF WATER BOTTLE SUSING GENERLIZED LINEAR MODEL FOR SUPPLY CHAIN RESILIENCE

By

ABDULAZIZ ALIDREES

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 and System Engineering

THE UNIVERSITY OF TEXAS AT EL PASO

DECEMBER 2020

ACKNOWLEDGMENTS

Foremost, I would like to thank God and my parents and siblings for their blessing and guidance. It revolves around the journey, not the destination. My earnest gratefulness goes to all of those who have escorted me on this journey of learning, shared their insight and care, and without whom the work outlined here would not be the same. Specifically, I wish to acknowledge.

Special thanks go out to my country for supporting me financially all the way during this journey. I am extremely thankful and privileged to UTEP for providing me with the opportunity to pursue my masters. I would not be able to take the first step without having an arm extended to me.

I thank my project supervisor, Dr Jose Espiritu Nolasco, for his continuous guidance and encouragement through the entire course of this study. His direction and impetus were truthfully motivating not just to this study, but also to my professional career. In absence of his support this study would have been impossible. All my colleagues from the university for spiritual support. My friends, some who are fellow master's students, here and there. Particularly Cate, Falah, Lara, Franco, Anna, Jacquelin and Mohammed, for comprehending, for numerous fun and light, and more thoughtful dialogues, and for your friendship all the way through.

And most notably, my dearest family, to whom I dedicate this thesis for listening uncomplainingly, and for offering me the opportunity and inspiring me to chase my dreams no matter what. Mom and Dad, who were an encouragement, shared with me their taste for writing and intellectual struggle, and whose pride in my accomplishments meant a lot to me.

ABSTRACT

When forecasting the amount of water consumed by individuals at Kuwait's Mubarak Al-Kabeer city Block 8, it is important for any water producing company to tell the amount of water they should produce for its target consumers. After interviewing a couple of residents in this block, it was discovered that they were restricted by being unable to shop for their essentials, including water packs. Data on the amount of water consumption in milliliters for different individuals was collected and recorded. Generic scanners and smartphone scanners were installed in 20 different households that rely on water packs instead of central water supply. Every individual in each household was asked to scan the water bottle after consumption. This information gets transferred to an excel spreadsheet, indicating that the individual has consumed a 500ml water bottle. The data was collected over a period of 30 days, and other factors for these individuals were also recorded. Also, qualitative data was collected from manufacturing employees about the circumstances that they are facing. These factors are the age of the individuals, weight and height, whether they have any of the following diseases: hypertension, diabetes, kidney problems, thyroid problems, allergy, asthma, heart disease, post-traumatic stress disorder, and irritable bowel syndrome. The total number of individuals who were recorded was 79. In this study, we used the generalized linear regression model to forecast the amount of water consumption, since this model has the ability to forecast the response variables from the normal distribution and uses categorical explanatory variables on the response variable. We began by cleaning and preparing the data by treating missing values and outliers, performed feature selection, and finally proceeded to fit the Generalized Linear Model. Parameter estimation of the regression coefficients of the model was done using the maximum likelihood estimation formula. The best regression model was selected using stepwise regression based on the AIC values. The

model that reduced the Akaike Information Criterion most was selected. Further investigation was done on this model, such as multicollinearity, which was absent, and tested the interaction between independent variables where the interaction between age and diabetic variables was considered. The model was finally used to forecast the amount of water consumption, and the accuracy of the final model was tested using the adjusted R squared and was found to be 29%. This is because the model was trained on a dataset that had a large number of observations in the range of 3-5 and was able to accurately predict only the values that lay in this range, compared to the values outside this range. The study finally recommended using clustering models for further research, such as the K-means clustering algorithm for forecasting.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Water is life. To maintain this life, there needs to be a way in which water consumption is managed as it is such a precious commodity. Distribution of water consists of a complex decision-making process where the water distribution managers have to ensure that all the customers get enough water and their supply should not run out. This water should be enough for all the customers. In order to achieve this particular goal, the management has to be prepared to meet all the demands from the customers and this can only be achieved when there is proper planning. In planning, the management will have to come up with a prediction of what amount of water may-be demanded in a particular time frame. This prediction can only be done if there is a model involved. The model should be reliable and should not require too many inputs to work. Such a model is the generalized linear model. The model seeks to have the equation of a line which, when drawn, can be used to forecast future demand. The generalized linear model is easy to use and also easily understandable due to certain features that can be applied. This study will use the generalized linear model to forecast future water demand. Preparedness will help to preserve the water supply of those who depend on it.

1.1 Supply Chain Management

In commerce, supply chain management (SCM) is the flow of goods and services that focus on storage and proper movements of the raw materials, finished goods, and focus on all of the work-in-process inventory (Hess, 2010). SCM is the interconnected or interlinked networks that have a particular node of the businesses, which combine the products and services' provision that have been required by the end customers, or the node of businesses. Furthermore, SCM has always been defined as the process following the steps like "designing, planning, execution, controlling,

and monitoring" of the activities associated with the supply chain, which has a core objective of creating the net value. In addition, SCM aims to build a competitive infrastructure, synchronize the supply with good demand, measure all activity's performance, and have certain leverage on worldwide logistics (Mentzer, 2001). SCM practices take place in different industries and domains. These are areas like system engineering, logistics, operations management and industrial engineering. Procurement, marketing, and information technology always strives for an integrated approach and ensures a proper flow of goods and services (Gölgeci & Kuivalainen, 2020). The supply chain works in a pattern that each node in this chain is dependent on others. If one node's logistics or functioning would disturb this, then the whole process would suffer likewise. In the process, the marketing channels play a vital role as they are the main drivers of SCM. It is essential in SCM to manage all the risks and be sustainable in all parts (Lam, 2018). Many things should be considered when the SCM is being handled. These factors include human capital and talent, or the management associated with it, the visibility/transparency, integration of ethical issues, and all the other parts that play a vital role in managing the supply chain in a particular manner (Wieland et al., 2016). SCM has several techniques aiming for the right coordination in several parts of the supply chain, from supplying the raw materials to delivering the resumptions of the products. This basically tries to minimize several total costs to the existing conflicts among the various partners. There might be several risks associated with the supply chain, which need to be solved or addressed so that all the supply chains could be managed better.

1.1.1 Functions of Supply Chain Management

Supply chain management (SCM) is basically a cross-functional approach that will incorporate managing raw-materials and moving into an organization. This also tends to manage the end customers who are the end node of the SCM and several aspects of the internal processes. The more an organization strives to concentrate on its core competencies, the more agile it would be (Eltantawy, 2011). They also strive to reduce the ownership of several distributions and the sources of raw materials. All these outsourced functions might help regulate the functionalities that help in efficiently keeping the supply chain activities. The main purpose of SCM is to improve the cooperation and trust among the various supply chain partners. (Eltantawy, 2011). This would all take place by taking care of inventory visibility, as well as the inventory levels. From that viewpoint, there is a need to regulate communication with the suppliers and manufacturers

1.1.2 Risks to Supply Chain Management

Supply chain management (SCM) is a commercial process that needs to be maintained with the right values. Furthermore, any type of process that involves an end customer requires a stable supply chain to ensure the smoothness of service and product flow. Moreover, any interference could extend the level of damage to the enterprise level. Any change in the supply unavoidable, especially in supply chain management (Brun & Caridi, 2008). This calls for risk management which is essential for SCM success. As time passes, the risk related to a given supply chain appears. Many of the risk management strategies that take place over the past years are inadequate in today's risk , because the risk is in a continuous cycle of change having that said, methods that have been effective in the past do not necessarily mean that they would be effective against present or potential obstacles.

SCM goes through many of the risks, and these risks are associated with air, land, and ocean freight (Brun & Caridi, 2008). These all are inevitable and cannot be taken easily. Supply chain always demands a strong grasp of the processes. Poor grasp always results in increased expenses to compensate for the losses associated with it. The risk varies with the industries, and some are unavoidable and might directly impact the profit gains of the company along with its brand image (Brun & Caridi, 2008). The top ten global supply chain risks are discussed below, describing how this can have an effect on the operations.

Political and government changes: Global trade might be impacted drastically on, in the presence of frequent changes in political and governmental policies. Businesses are globalizing, and when international transactions come into play, it is essential to look after the political and governmental considerations. Each region has its own regulations. Frequent changes in political and governmental obligations might bring unstable factors that are not good for the business's initiatives. Businesses always need to consider other countries' policies with where trade is taking place, as a violation might bring several problems which might have an economic impact.

Economic stability: Economic stability is essential for business growth. Today, we live in an era where the supply chain is operating on a global level. It is common to have one of the supply chain nodes present in a different country, unlike other business operations. In one supply chain, several organizations and entities are involved. So, any instability in the economic area may impact all of the nodes involved in the supply chain. During the COVID pandemic, companies like eating hubs, grocery manufacturers, and other businesses have faced a huge collapse due to the stoppage of supply chain operations (Lu, 2020). International businesses are on the brink of bankruptcy. This is a realistic example of how a barrier or instability has a negative economic impact on the entire supply chain.

Extreme weather events: This is a major risk, as the supply chain is primarily dependent on ocean transportation, one of the key drivers of all supply chains, and links the nodes regionally. Poor weather conditions could lead to problems and could be an obstacle to the regulation of business practices. The ocean does much of the transportation that is chosen to control the supply chain globally, and the risks associated with the ocean are dramatic. Times are unpredictable, as we can see in the natural world. Environmental changes take place in various parts of the world-earthquakes, cyclones, tropical storms, and much more. Companies may have economic losses. This is the worst hazard since it is unpredictable, and greater preparation can fail.

Catastrophes: There are man-made threats associated with the supply chain. Risks such as hurricanes and starvation are considered in this category. Various riots or protests may also be used. Recently, a variety of demonstrations or protests have taken place in the United States that have had an impact on companies, while uncertain events have taken place that have compromised the supply chain (Garver, 2020). For instance, if there is some protesting in one of the areas where the production unit of the brand is located. Due to these reasons, the whole supply chain might suffer due to the fact that resources would not be able to move into the supply chain patterns.

Connectivity: The supply chain is highly dependent on connectivity through the nodes. There is a strong demand for 24/7 communication when it comes to working with the global and local supply chain in order to provide more regulation in the network without running into the associated risks. Any disruption may lead to a variety of defects in the supply chain. It is possible to attain better communication and regulate business activities with the help of the integration of

open-source software (Barrios, 2018). As a result, a lack of communication could lead to an error, as no status of operation could be changed, and the business process could be interrupted.

Environmental risk: This is one of the risks that businesses need to be concerned about. These include sudden changes exerted from the environment, such as global warming, pollution, and the residue of wars that lead to disasters. From this perspective, choosing a safe place for the supply chain nodes is very important.

Cyberattacks: When companies use online connectivity to regulate their business globally and locally, this is an important element to take into consideration. The internet goes through its own cybercrimes. As technology is developing, threats are also growing. Companies have started merchandising online, which involves critical data of associated concerned parties. Any malicious activities in the network might be very costly to the associated parties. So, this is one of the major risks as it deals with a significant amount of information that should not be leaked. These days companies are using new technologies like the Cloud, which makes it riskier. So, it is essential to pay attention to this factor as well. It is essential to know that if the technology is rising, then so do the risks. Security needs to be properly managed, especially when it comes to online transactions.

Data integrity and quality: The supply chain involves many customers participating in the supply chain, from the supplier node to the customer node. Data quality should be controlled, secured, and not undermined because it is the key driver on which the processes are based. Without consistent data quality, operations and facilities do not exist in the first place.

Transport loss: Transportation is one of the key activities of the supply chain, as it is the link to the entire supply chain. Several supply chains also rely on rail, ocean, air, and other modes of transportation. Often the material being transported can be very fragile and requires

special care and attention. Any infringement in this regard could jeopardize our supply chain reliability. The danger associated with transport is adverse. For instance, products such as wine require a proper temperature to be preserved, oil requires several restrictions, glass products need proper packaging, and many more.

Supplier consistency: Suppliers play a crucial role in the supply chain as they help regulate individual goods' movement to the next nodes. Inconsistency on behalf of the suppliers could lead to problems with the reputation of the brand. Notorious reputation could decrease demand for the product in the marketplace due to inconsistency that could lead to a potential decline in the economy. This reputation will create issues for manufacturers as they will face loss too. So, it would be a threat involved in the SCM.

These could be dealt with by taking some of the steps that could help bring better changes. Supply Chain Resilience is one solution that could be really beneficial to businesses and can bring progress, a more stable role, and continued risk reduction for the whole supply chain method. These negative experiences call for strategies and commitment to deliver competitive outcomes.

1.1.3 Supply Chain Resilience

The supply chain is like a company that has sensitive and complex processes. Companies receive their goods from large manufacturers, who outsource their components or materials to others, and who may outsource them to others in a certain way. When one part of the supply chain network is exposed to risk, all other parts are vulnerable and disruptive (Deloitte, 2018). It is necessary to concentrate on developing the right form of supply chain resilience that will, in some way, be part of the risk management strategy.

This will allow one company to transform the risk view into an opportunity that generates certain benefits. The key goal of supply chain resilience is to build value to nodes of the supply chain and reduce the risk from the whole supply chain.

1.1.4 Need for Resilience in Supply Chain

Resilience is the heart of supply chain management thinking, which is currently being followed, and it is about understanding the concept of where to invest in resilience (Melnyk et al., 2015). If this understanding is achieved, this can lead to a faster adaptation to interruptions or some sudden risk, or even help with faster recovery. It is observed that resilience is becoming a serious concern. Such disruptions must be dealt with properly and accurately and should be convincing (Pires Ribeiro & Barbosa-Povoa, 2018). In this, the use of resilience metrics is a must to help control the whole process. Resilience is a necessity for any supply chain. Resilience is all about being alert to, responding to, and adapting to changes brought about by a particular disruption (Ambulkar et al., 2014). These characteristics could be attained by having either the right type of organizational supply chain agility, or organizational supply chain robustness. These are two faces of the same currency. A robust supply chain should be closely linked to its activities and specific strategies to respond to risks that affect its capability. Supply chain resilience is basically improved by inter-firm and inter-personal collaborations that go across the supply chain network. Therefore, it expects the right kind of support to adapt to the processes(GT Nexus, 2020). It is not all about just responding to a one-time crisis or just having a flexible supply chain. It is all about continuous adjustments and the anticipation of the disruptions that might permanently impair the core business. Resilience requires strategies. Strategies require continuous process innovation and product structures and look after corporate behavior (Willke &Willke, 2012).

1.1.5 Risk Mitigation With Supply Chain Resilience

While all risks are viewed as a threat and must be eliminated or at least reduce their effects. Resilience is thought of as an opportunity that we must utilize. The supply chain also revolves around the concept of adding values, which has strategic themes. These values include increased market share value, competitive advantage, shareholder values - and a typical brand image that values their customers. All of the risks could be mitigated through strategies, but this requires proper management alongside it. For several practical purposes, a detailed description of the business processes and the business discipline is required. These processes include breaking down real concepts into several components to better grasp the causes of changes or even delays.

1.2 Effect on Supply Chain due to COVID-19

Starting in December 2019, new coronavirus (COVID-19) outbreaks arose from Wuhan, China, worldwide. About 1,100 people died because of the virus. At the beginning of February 2020, it was observed that about 44,000 people had been infected with COVID-19 worldwide at that time. The problem has been on the rise since then (WHO, 2020).

Human well-being is immensely important, but the rules that have been passing lately from the governments impacted economic functionality as a whole. This global pandemic has had a considerable influence on the supply chain. Much of the effect has been seen to date in parts of Asia, but consequences are growing tremendously across the globe (Civil Daily, 2020). Many things have occurred that have prevented normal activities, and nations have been locked up, and curfews have taken place in several nations. All industries, schools, stores, manufacturers, and everything else have been ordered to be closed due to the outbreak of COVID-19. The governments made different decisions and took actions to reduce the disease transmission, for example by applying the rules and charged fines on people who were out without any good

reason. Many manufacturers have remained closed in countries, which has directly impacted the business rate and economy rates. This outbreak has banned shipment processes, which has a fundamental rule to supply chain operations. Also, the suspension of flights has affected business operations and the supply chain due to the restrictions (The New York Times, 2020). The restrictions were not just limited to airline services, but there were also restrictions on ship docking that have been increasing all the time. Considering the accumulation and the development of these indications these are critical threats to the supply chain at the global and local levels, and operations that have been limited would directly affect the economical rate in a significant way.

1.3 Global Supply Chain Risks due to COVID-19

Many supply chains seek globalization, and their operations are becoming a regular part of today's business. In particular, many shifts in the supply chain have occurred in the past decade, and globalization has essentially become the backbone of business operations worldwide. However, the supply chain may be exposed to certain risks, especially when it comes to diversification (The Economist, 2020). Since the processes in the supply chain are all mutually related, any risk occurring in one region may cause problems in other regions as well. Several companies have faced major threats in recent coronavirus circumstances. Unemployment rates are hard to ignore since more than 7.5 million people are unemployed in the United States alone (Saphir, 2020). Businesses are closing, and this has a massive effect on business areas. Supply chain problems and companies' deterioration are taking place because suppliers and organizations are not adequately planned for the disruption. The organization needs to choose some better practices that might help them to be resilient and robust when these kinds of situations occur. Businesses such as the food industry, grocery stores, and related businesses

have been significantly impacted. There have been shortages in the food supply chain, toilet paper, cereals, and packaged goods due to COVID-19 and the outbreak, which severely affected supply and demand operations.

1.3.1 Reason for Shortage in Supply Chain in Grocery Stores due to COVID-19

COVID-19 disrupted global food supplies and also triggered labor shortages in the food industry. When the lock-down took place, it created panic-buying situations all over the place and caused a mess in grocery stores. The shopping behavior of some customers cleared the shelves of the supermarkets and stores. People have purchased rice, pasta, and other essentials in a large amount, and this has impacted global suppliers. Suppliers involved with items such as meat, dairy, fruit, and vegetables have been struggling to move supplies between restaurants and grocery stores, which basically causes consumers to face shortages. Many goods do not reach the market due to the interruption of production and transportation, creating severe damage to the supply chain. COVID-19 has revealed chaos in the food supply chain in a matter of a few weeks. Several organizations are preparing and distributing their food in a certain way to catch up. However , COVID-19 has dramatically affected grocery stores which resulted in empty shelves of essentials . There are also issues with costs that are also rising. The bulk-purchase behavior of shoppers has a detrimental effect. The empty shelves' images have been circulated with the footage of people fighting over toilet paper because people were so afraid that they would not survive. This has caused a spike in the sales price of rice, which have grown by 25%, dried beans by 37%, and pasta by 10% (Rubinstein, 2020). From a sales point of view, stores are making revenue out of this crisis, which is a more positive way of looking at the crisis. Therefore, it was surprising that sales and demand would grow at a tremendous pace.

Nevertheless, unexpected demands shockingly changed the whole chain. So, the whole shortage has emerged hugely, due to the unreasonable management of the supply chain

Aims and Objectives

The objectives and aims of the research are listed below:

1. Trying to build supply chain resilience by being able to forecast customer water needs
2. Help manufacturers to operate despite the various shortages in resources and labor.
3. Determine the right amount of water production in the absence of a precise knowledge of demand and a lack of cooperation in the supply chain nodes.
4. Build resilient-based manufacture that can operates despite various shortages.

An Optimal way to Solve the Problem

Supply shortages were a critical problem in the supply chain of grocery items during the COVID-19 situation due to the lack of effective management , planning and inconsistencies in the supply chain. This was an external type of risk that happened in the supply chain, which was uncertain beforehand. It is essential to understand the multi-dimensional risks and their nature. Therefore, it is crucial to create a resilient and reliable supply chain to improve the given conditions and improve the capacity to respond rapidly to changes (Benard, 2004). The whole supply chain nodes should be able to cope with the situation and be flexible in changing their operations and implement the practices as necessary (Wisner et al., 2014). Risk management procedures should be taken into consideration such as :-

Risk assessment: This stage takes care of evaluating the loopholes of the organization and try solving them in order to ensure the preparedness to a given disruption

Risk mitigation and response planning: This step is where the severity and the probability of a given disruption is evaluated. The potential disruptions are divided into different

categories. The first category is low, medium, and high. The probability is rated on a scale of 5. Based on the probability and severity, mitigation plans are developed.

Event management and coordination: There is a required operational capability that helps in effectively managing the incident such as communication during disruption that might take place across the nodes and these capabilities need to be present at all times.

Response execution: Which is the final step of the whole process. It refers to implementing the plans that have been developed in the early stage according to its severity and probability of taking place. This stage includes monitoring the implementation and making sure that everything is going as planned and the risk does not occur again

1.5 Statement of the problem

Water producing companies seek to predict the expected amount of water to be consumed by their consumers with a high level of accuracy so as to reduce the risk of overproducing or under-producing the amount of water for sale, in the presence of a lack of different resources due to the COVID-19 pandemic. This is essential for the company to efficiently determine capital and resource allocation to its various sectors, especially in the absence of the several resources. An accurate prediction of the water consumption by the customers of the water company will therefore be a very important thing for the company to prevent jeopardizing people's lives during these severe days, where water can be barely found in grocery stores. Water-producing companies usually produce on a steady rate all year long, regardless of the season. Furthermore, water companies would not be able to cover a demand which is almost 10 times higher than the demand throughout the years. The demand that was experienced was due to irrational behavior, which was triggered by human instinct, which is survival for living. During my visit to the grocery stores and manufacturers, and interviewing a couple of employees, they mentioned that

manufacturers do not know exactly the demand of customers, so the manufacturer ended up producing the incorrect quantity. Adequate information exchanging of demand and capacity between the supply chain nodes would greatly help the flow of the goods and eventually reduce the lead time. The study will therefore use the generalized linear models to model a base data that could be used to help the company estimate the respective amount of water consumption and reduce the level of uncertainty in the city where the study was conducted, in order to help on the short-term basis.]

1.6 Objectives of the Study

The general objective of this study is to forecast the amount of water consumed by consumers and help the companies in determining the exact amount of water to be produced. This is to supply a given demand without running into the risk of over-producing or under-producing with the presence of limited resources using generalized linear models.

Specific Objectives

(i) Collect water consumption data

(ii) Clean and prepare the data

(iii) Determine the appropriate model to use for forecasting and fit it to the study

(iv) Testing the accuracy of the final model

(v) Forecast the values of water consumption

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This literature review aims to understand supply chain management (SCM), the issues related to it and the techniques available to optimize it. Organizational and supply chain resilience is a very important aspect of any organization. Companies must have an integrated framework effectively to encounter the issues that have been raised in SCM. Every company must have emergency operating groups to act rigorously when there is a problem in SCM. Efficiency in SCM could be achieved through effective human resource management. If the right people are set to manage things, then all operations will run smoothly. The information system should be accurate and up to date in order to effectively manage SCM. Enterprise Resource Planning tools can be used to handle SCM without errors and optimum output. Technologies such as a fuzzy logic controller can be used to organize the flow of materials from source to destination. Transportation issues such as airline attacks and tropical storms in oceans have to be managed with predetermined strategies to handle the situations efficiently.

2.2 Literature Review

It is very important to understand the theory of organizational resilience Vogus & Sutcliffe, (2007) explain the concept of organizational resilience by taking into consideration the associated theory associated. The authors have focused on exploring the importance of resilience in the organization. They are of the view that organization theory does not hold great importance at present. Further, there have been varying definitions of the concept demonstrated by the authors. Moreover, they have also considered the affective, cognitive, relational, and structural mechanisms in association with organization resilience. At the conclusion, the authors emphasize the importance of the theory, which relies on the ability of the organization to adjust during

adverse conditions. So, for this purpose, there must be some framework of organizational resilience. This great research has been carried out earlier as well, but they do lack certain things.

Kantur & İseri Say, (2012) propose the conceptual integrated framework of organizational resilience for the optimization of SCM. According to the authors, as there has been a surge in the chaos in today's business environments, organizations should be more resilient. The concept of organizational resilience is mainly based on disaster management and emergency management. However, the paper has focused on the notion in the context of organization management. The research suggests some of the basic factors that contribute to the emergence of resilience in an organization. Corresponding to all this, the authors have proposed the integrative framework for organizational resilience and developed a new concept related to organizational evolvability. The concept focuses on increasing sensitivity and enhances the wisdom and understanding of the post-event organization. The authors have categorized organizational resilience as contextual integrity, strategic capacity, perceptual stance, and organizational resilience, according to the proposed model. Also, McManus, Seville, Vargo, & Brunsdon, (2008) discuss the facilitated process that helps to improve organizational resilience. In the views of the authors, resilient organizations play a significant role in resilient communities. However, it is difficult to translate the concept of resilience into working conditions for the organizations. Further, the reports state that the concept of resilience is related to crisis or emergency management issues. The organizations have not well-developed the link between the day-to-day operations and instead have the recovery of the crisis through the organization's resilience. The researchers have proposed a resilient management system, which had three principal attributes: situation awareness, management of keystone vulnerabilities, and

adaptive capacity. Based on these attributes, the paper aims to introduce the facilitated process that helps the organization to improve the performance in association with those three attributes.

Gittell, Cameron, Lim, & Rivas, (2006) explain the organizational resilience in the context of airline industry responses to September 11, 2001. Talking about the terrorist attack of that day that impacted the airline industry immensely, as compared to the other industries, there has been a successful response by most of the industry to the attack. However, there were a few industries that were not able to cope with the situation and became the victims. The paper focuses on those airline companies that successfully recovered from the attacks. The researchers have further stated that it is the development and preservation of the relational reserves that led to a reliable business model. This model has been taken into consideration in this study. Finally, the authors concluded that adequate financial reserves and preservation of relational reserves contribute to organizational resilience. Similarly, Mamouni Limnios et al., (2014) propose the resilience architecture framework. There has been research on organizational resilience in the past as well. However, according to the authors, the research fails to conclude whether resilience is a desirable or undesirable system depending on the system state. The paper aims to develop an organizational typology, that is, the Resilience Architecture Framework (RAF). This concept is a source of integration of various research streams based on organizational rigidity, organizational ambidexterity, and versatile capabilities. The authors came to the conclusion that organizational resilience with this framework will provide the future scope of research.

Also, Lengnick-Hall, Beck, & Lengnick-Hall, (2011) discuss the development of organizational resilience through strategic human resource management. The authors stated that resilience organizations have the ability to survive even during uncertain, adverse, and unstable conditions. In the views of the authors, the organization's capacity is developed with the help of

human resource management. This helps to create those competences among the employees which lead to the ability of the organization to respond in a resilient manner during a difficult situation. The authors suggest three elements that are important to develop the capacity of the organization for resilience. The three elements are specific cognitive abilities, contextual conditions, and behavioral characteristics. Further, the contributions made by the employees at the individual level have been identified. Moreover, the authors have also explained the way in which a strategic human resource management system helps to influence individual behavior and attitude towards the organization. The authors have concluded that these elements are identified at the organizational level in context to the double interact and attraction selection attrition that creates the capacity for resilience. Next, Riolli & Savicki, (2003) propose the organizational resilience of the information system. This paper has examined the data on individual and organizational resilience. The authors have developed a theoretical model to determine the lack of research and theory regarding the resilience in the information system. Both the individual and organizational level of response has been considered to understand the organizational resilience in the information system field. The authors have concentrated on the organizational structures and processes in addition to the organizational factors that represent the sources of vulnerabilities and protection on the organizational level. Considering the individual level, the factors that address the matter of resilience are situational demands, constraints, and deficient resources. They are integrated with individual differences such as base values, skills, personality, and dispositions. This paper focuses on the elaboration of the model to support the individual level research findings. The authors have discussed the information system framework and future research regarding the concept of organizational resilience. Similarly, Aleksić et al., (2013) assess the organizational resilience potential in Small and Medium-sized Enterprises (SMEs) of

the process industries. According to the authors, the concept of organizational resilience has been established in the context of the modern business environments that help the organization to overcome the crisis and emerging situation. The concept of organizational resilience holds importance during the normal period of operation. However, it is even more needed during the time of crisis. The failure during a crisis may cause significant problems in the other processes as well. In this paper, the fuzzy mathematical model has been developed for the identification of organizational resilience in the process industry. The authors have taken the illustrative example, where the data gathered states the factors that help to improve the strategies of the business and organizational resilience. Moreover, the survey taken by the authors has been included in the research by taking a significant organization from one region.

On the other hand, Tadić, Aleksić, Stefanović, & Arsovski, (2014) present the evaluation and ranking of organizational resilience factors based on Two-Step Fuzzy AHP and Fuzzy TOPSIS. The authors have illustrated the novel fuzzy multi-criteria decision-making approach that helps in the evaluation and ranking of organizational factors. The user preference orders have been considered for this approach. The authors suggest the view that due to the vagueness of the decision data, numerical data is not adequate for situations related to real-life business. Moreover, fuzzy sets present linguistic expressions that express human judgments. The treated problem has been solved with the help of the two fuzzy multi-criteria models. The authors have applied the Fuzzy Analytic Hierarchical Process (FAHP). This helps the authors to identify the importance of the business processes and the organizational resilience factors based on the business processes. Further, the ranking of the organizational resilience factors has been determined with the help of the extension of the fuzzy technique for order preference by similarity to ideal solution. Concerning the complexity and the type of management problem, the

paper comprises the introduction to a modified fuzzy decision matrix. The researchers have applied the proposed algorithm for the purpose of assessment of organizational resilience factors concerning the SMEs of the process industry. Similarly, Crichton, Ramsay, & Kelly, (2009) suggest information about enhancing organizational resilience through emergency planning. In the views of the authors, the learning can be grabbed from all the emergency exercises or actual incidents. The paper aims to examine the recurring themes to be applied across various sectors, from the lessons learned. The lessons learned from those events are expressed in the specific form, that is, in context to the actual event and the sector in which it has occurred. The report has considered seven incidents that occurred in the United Kingdom and at an international level. The wide range of sectors with varying parameters has also been identified. The authors are of the view that through the lessons learned from the incidents, the organization can become wiser. The recurring themes are used to explore the resilience of the emergency plans. Moreover, the authors have also proposed recommendations to implement the best practices in improving the learning of lessons within the organizations.

Hillmann, Duchek, Meyr, & Guenther, (2018) recommend the ways which help future managers to develop organizational resilience. In the views of the authors, managers play an important role in developing organizational resilience. The managers are supposed to employ the long-term visioning in volatile and uncertain times. In other words, managers are responsible for promoting organizational resilience capabilities. However, on the basis of the research, strategic management education lacks the proper methodology of providing accurate learning experiences that helps to develop the capabilities related to organizational resilience. The paper contains the use of qualitative research design by taking the experimental character as an example. This is to make the difference between the groups of students learning interventions and the students who

took part in the case study. Based on the results, it has been stated that the former group is more superior and efficient in the matter of the strategy process. They are also better in terms of performance outcomes, learning outcomes, plausibility, creativity, transferability. The analysis of the study shows that there is a positive impact on resilience capabilities such as sense-making and anticipation. Similarly, research by Bouaziz & Smaoui Hachicha, (2018) reflects information about Strategic Human Resource Management practices and organizational resilience. The paper aims at determining the relationship between strategic human resource management and organizational resilience in the context of the Tunisian democratic transition. The authors have assumed that there is an influence of SHRM practice on the organizational resilience dimensions. The deductive approach has been used in this paper to accomplish the aim of identifying the matter of organizational resilience. On the basis of the results, it has been analyzed that SHRM practices have a great impact on the resilience dimensions. This practice helps to boost the robustness of the organization and this is frequent in the second period. Moreover, it also impacts the agility and integrity of the organization.

Sahebjamnia et al., (2015) proposed the research about the integrated business continuity and disaster recovery planning that should take place towards organizational resilience. The businesses can easily be disrupted, and it is nearly impossible to have a prediction about their time, extent, and nature. Thus, these types of decision support frameworks for the protection are needed against disruptive events as a proactive approach. This article provided the framework for integrated disaster recovery planning and business continuity for the critical functionalities. The study proposed a model that was successful for the decision problems at all tactical, operational, and strategic levels. The concept of organizational resilience was first explored at the strategic level followed by a multi-objective mixed-integer linear programming. This was acquired for an

allocation of internal and external resources to recover and resume the business plans. The model has a goal to have control over the loss of resilience by maintaining the minimum recovery time objectives and maximizing the recovery point. When it came to the operational level, the evaluation of hypothetical disruptive events occurred for examining the plans' applicability. The authors also developed an augmented ε-constraint method for finding a compromise solution. The authors have validated these methods through a real case study for better outcomes to be achieved. Ignatiadis & Nandhakumar, (2007) proposed a study based on the concept that how the enterprise systems have some impacts on organizational resilience. It is a fact that the enterprise systems are firmly used for facilitating the exchange of information between the departments and the seamless integration within an organization. For all of this to be attained, it is necessary to control mechanisms and deal with the systems. These help in safeguarding the organization's data or avoid the unintended, yet unauthorized, use of the system as well. However, this method is attainable to a certain extent and is ideal for total control. The article has the purpose of having an organization where the enterprise system has been deployed. It has been suggested that the enterprise system develops the power differentials which further serve to enhance the control in a company. All of the processing assists in enhanced rigidity or decrement in the organizational resilience and flexibility.

Whereas, in different circumstances, these enterprise systems root drift, resulting from certain unexpected circumstances associated with these power differentials along with the people's discernment in solving the problem within the enterprise systems. It has been concluded in this study that the reduction in control might serve in some circumstances that can enable organizational resilience. Next, Ortiz-de-Mandojana & Bansal, (2015) recommends an insight over the long-term benefits of organizational resilience through the help of certain sustainable

business processes. The author identified the facts about the benefits of business sustainability that often are applied to data analysis and short-term casual logic. Authors have argued about the social and the environmental practices (SEPs) that have been associated with business sustainability. These have basically not contributed to the short-term outcomes but are applicable to organizational resilience as well. In this article, the authors defined concepts like correct maladaptive tendencies, the firm's ability to sense, and coping with the unexpected situation positively. The authors have identified the advantages and disadvantages associated with organizational resilience. Similarly, the study found that the concept is a path-dependent construct and latent. The authors assessed it through several long-term outcomes with the inclusion of the sales growth, improvised financial volatility, and the survival rates. Authors of this study have tested the same hypotheses with a dataset fetched from 121 matched pairs based in the US - in total 242 individual firms. The whole process took place over a 15-year period. The authors concluded that they were unable to find out any relationships among the short-term financial performance and the SEPs after all of the intended tests.

Next, Powley, (2009) presented research based on reclaiming resilience and safety and identified that resilience activation can be a critical period of crisis. An organization has a latent capacity for rebounding the activities to enable the bounce back and positive adaptation. This takes place in the case when the normal flow of the organization, as well as the relational practices and routines, can be disrupted. This literature firmly examined the organizational crisis that is unexpected in nature. This presents a precise model of how the whole resilience has been activated in certain similar situations. It has been observed that resilience activation can be described with the help of three social mechanisms. In this study, it has been described that liminal suspension has a description of how the crisis alters the formal relational structures, then

how a crisis temporarily undoes and opens a temporal space for the company. This all takes place to renew the right relationships between partners. It has been noticed that compassionate witnessing explains how the members of an organization's opportunities for the management response needed to be maintained for an individual's needs. The authors have also explained about the relational redundancy, in that it explains the social capital of the organizational members and the connections that have been built across the functional and organizational boundaries. These boundaries will be activating some relational networks that enable resilience. It has been noticed that these narrative accounts can have the support of the induced model.

Annarelli & Nonino, (2016) explained about the strategic and operational management of organizational resilience. The authors put an insight over the current state as well as future directions for the research. It has been noticed that the article is based on critical analysis and the literature search for the right investigation of the research associated with organizational resilience. It is observed that the research stream is based on the operational management of resilience and organizational resilience too. As per the authors' findings, these are the terms that are distant from its infancy, but these are required to be considered in a developing phase. The authors have found a piece of evidence from some academic literature that surely has helped in sharing the consensus on the definition of resilience, characteristics, and the foundations in certain recent years. It has been observed that the major focus of the research is based on supply chain resilience. It is mentioned by the authors that reaching consensus on the implementation of the subject is still too far for the literature. It has been noticed that the literature has explored topics like how to create resilience processes and maintain them as well. This also focused on the ways to reach operational resilience. The authors have imposed certain processes and methods and find out various results based on that. The results explain that they have found certain future

right directions on operational, organizational, and strategic analysis. This has been born out from the various research accessed for the study.

Similarly, Linnenluecke et al., (2011) addressed research about extreme weather events and the certain critical importance of anticipatory adaptation along with the organizational resilience and looked after certain associated impacts as well. Authors have stated that growing scientific evidence explains that certain severe weather extremes like flooding, hurricanes, droughts, and heat waves will be impacting organizations, entire economies, and also on the industries heavily. It has been observed that the findings are associated with the practical and theoretical frameworks for strengthening the organizations' capacity for having the responses to those impacts, though it would be necessary to understand the requirement for building up the anticipatory adaptation. It also includes facts about the organizational theory literature that have certain offers related only to the limited insights. The article basically proposed the framework about organizational resilience and adaptation for extreme weather events. It was necessary to address the several effects of discontinuities which take place in ecology. Finally, this study has predicted some of the measures and the suggestions that can be used in future research. Next, Somers, (2009) presented a study based on measuring resilience potential and marked it as an effective or adaptive strategy for organizational crisis planning. The author firstly addressed the fact that whether crisis planning, and the crisis' effective adaptive behaviors, have some of the causal relationships between them or not. It has been stated that traditional planning has been viewed for the right crisis planning which will be basically an outcome of a process. The process is required to be utilized during the crisis too in the step-by-step fashion. This study basically challenges the orthodox view and plans on suggesting a new paradigm. These are completely focused on the creation of organizational processes and the structures that help in building the

potential organizational resilience. The main objective was focused on developing the scales to measure latent resilience in the organization. It has been observed in this research that it has started building the critical foundation that will be helping search for the new paradigm for disaster planning. The new paradigm is all based on the organizational resilience potential that should be precisely focused on the future of research. Also, Orchiston et al., (2016) addressed a study based on organizational resilience in the tourism sector. The researchers have used a tourism organization's data which belonged to Canterbury, New Zealand. It is estimated or observed that this study has identified resilience's dimensions for the tourism organization for the post-disaster context. This has provided a successful quantitative assessment associated with organizational resilience. This has been measured within the different sectors of the industry (tourism industry). This also provides a supporting hand for having the findings associated with the key attributes of resilience that are referred to as the culture and planning or the innovation and collaboration. The research also presented certain methods that could assist in the resilience assessment and could be adopted in certain other studies as well.

Mallak, (2002) provided a study towards the theory of organizational resilience. The author mentioned that the workers are going through rapid change from many sources. Resilience is basically an ability of an organization or an individual that can help in implementing the positive adaptive behaviors and the expeditious design. These are picked up because they have matched to certain immediate situations and work on enduring minimal stress. This issue has been addressed with various disciplines and with the matter of the different types of perspectives. One question was essential to be addressed in this study - how the organizations and individuals will respond to the outcomes that come from the effects that take place because of change. The article somehow managed working towards the unified theory of resilience. This

theory is used to effectively manage and embrace the changes taking place in the organizational space effectively. It has been mentioned by the author that the concept of resilient organization is evolving by the time for the purpose of coping and understanding. This will all be taking place with the associated work stress and the pace of change occurring in the modern-day generation.

Similarly, Chewning et al., (2012) proposed a research-based concept of organizational resilience and rebuilding the communication structures with the help of information and communication structures. The authors have employed a perspective on the organizational resilience to evaluate how Information and Communication Technologies (ICTs) were being used by the organizations to support the recovery from Hurricane Katrina in the USA, which occurred in 2006. It has been observed that longitudinal analysis was carried out that was in the contact of ICT use and it is done by carrying out in-depth interviews. The results of interviews showed the organizations have enacted the resilient behaviors with a huge variety of conditions. This was possible with the help of information sharing, adaptive ICT use, resource acquisition, and (re)connection. The findings for all the interviews and the analysis emphasize the ICT transition that is used in the different stages of recovery. The author also mentioned in the research that the stages of recovery also involve an anticipated stage. Organizational resilience was advancing with the association of the external availability along with the additional sources used for the reliance. The authors have discussed various other contributions of ICTs in the context of disaster and concerning resilience.

Sawalha, (2015) discusses managing adversity and understanding the dimension of organizational resilience. The aim of the study was to find out about how organizations in the insurance operations interpret organizational resilience. The goal was to identify the several elements, potential objectives, and practices of the concept - i.e. organizational resilience. The

authors also aimed to investigate several impacts that culture might have on resilience. The research acquired an approach to work on the desired outcomes. The study was cordially taken out in Jordan's insurance industry where a total of 28 companies were registered at the Amman Stock exchange. The researchers collected the information with the help of the conduction of the survey that was followed by three semi-structured interviews. The results of the study revealed that many of the respondents were aware of the concept of organizational resilience but understood the meaning differently. The concept is constituted with various factors in which some of them have the potential to improvise, but they were missing the access to organizational resilience. It has been noticed that culture is one factor that influences organizational resilience levels. This has explored some of the practical implications which enable a company to withstand future risks which help in ensuring long-term survival. This paper contributed to Jordan organizing policymakers to start with the active existing resources. This paper suggested considering several cultural trends that can help in initiating certain new frameworks. The study used both qualitative and quantitative approaches and made a solid context in the emerging economy for the betterment.

Wicker et al., (2013) addressed the topic of organizational resilience for various community sports clubs that are impacted by natural disasters. It is seen that when these sports clubs suffer then it impacts organizational resilience which is critical to recovery. The authors of this study have conceptualized the concept of organizational resilience as a function of redundancy, rapidity, resourcefulness, and robustness. This has been applied to certain community sports clubs. The study has investigated the data collected through the survey from 200 Australian sports clubs which were affected by natural disasters like cyclones and flooding. The findings of the survey show that those clubs have used financial resources along with the

human resources for their recovery conduction. The use of government, the number of members, and the organizational resilience had a positive effect on the overall recovery of clubs. There are some factors that can assist in recovering from such issues. These factors include suitable insurance coverage, government grants, and inter-organizational relationships. This paper recommends expanding and refining the measurement of the concept for further investigation.

Spiegler et al., (2015) addressed a study based on the nonlinear control theory's values in the proper investigation that is underlying the resilience and dynamic of the grocery supply chain. In this research, an empirical context considers several methods that are required to use the non-linear control theory in a proper supply chain resilience is dynamic which is firmly developed and tested. The method used in this research utilizes the proper block diagram development and describes the function representation of the stimulation and non-linearities, and the transfer of the function formulation. The researchers have used two types of response lenses in the study: 'filter' or the frequency response lenses, and the 'shock' or the step response which is the system dynamics model created to firmly analyze the performance of the resilience of the replenishment system's distribution center at a larger grocery retailer. The authors of this study also discussed the potential risks for the right sort of performance, based on resilience, that includes the mismatch possibility between the supply and demand. This is all about serving the cause of on-shelf stock-outs and storing in an inefficient manner. This research completely explains the proper resilience that is determined in the grocery shipment. This has investigated the right behavior of shipment responses and stock responses of a dynamic nature. The authors have chosen the right methods that firmly allow better structures based on non-linear system control which would not be evident using proper simulation alone. The whole method includes the better type of understanding which has an insight based on the non-linear system control

structure. This will regulate better identification of the inventory system that potentially leads to the ‘drift’, the right sort of impacts on the non-linearities on the performance of the supply chain and minimizes the simulation changes. This study would be helpful in a proper management system for resilience in grocery stores.

Similarly, Hecht et al., (2019) addressed a qualitative study based on the urban food supply chain resilience for the crisis threatening food security. The study involves growth, distribution, and supplying food business and organizational functionalities. It is addressed in this particular research that the supply chain of food and grocery might face disruption from several hazards that might be the natural or human-generated, that might range from political issues to weather extremities. The main aim of the researchers in this study was to have an identification of the factors that might be associated with the resilience of the organizational level food system. The authors also looked at how these factors should be responded to and how it might relate to all the confidentialities in the disruptive events. The study has used a method of interviews - semi-structured in-depth interviews. The authors have conducted these interviews with the associated representatives of the key food system organizations and businesses by means of proper sampling. In the research, 26 food systems’ organization representatives have been satisfied by the two informant categories. There might be categories like the governmental offices and the non-profit associations as well that are all involved in the supplying and distribution business. The interviews were analyzed in a proper manner and several results were carried out based on them. The results depicted that there are 10 factors that might contribute to organizational resilience. These factors are: formal emergency planning; staff attendance; service providers; staff training; post-event learning; redundancy of food supply, food suppliers, infrastructure, location, insurance. Researchers focused on stressing the fact that reliability access could ensure

safe food for all of the people and the organizations or the supply chain should strengthen their operations in the case of disruption in an optimal manner. The authors concluded that this method would bring improvised resilience for better functioning. In the other hand , Antonio Candelieria, Davide Soldia and Francesco Archettia (2015) studied short-term forecasting of hourly water consumption by using automatic metering readers data. In their study, they used a regression model to forecast the amount of water that would be demanded by customers. They found that availability of more data, for example one year of consumption data, could help to identify, automatically, if the behavior is cyclic – a typical pattern – or a completely casual variation in consumption pattern just for that day (anomaly). The proposed approach for short-term water demand forecasting adopted a two-stage learning schema based on time-series data clustering (first stage) and Support Vector Machine for regression (second stage). This was a completely data-driven, fully adaptive and self-learning approach that had been designed and developed to be applicable both at aggregated level (i.e., urban water demand data from SCADA) and at individual customers' level. The approach was found to be reliable on both individual data and a wider set of data even though a wider set was needed to make the results more concrete and statistically concrete. Furthermore, Adam Piasecki, Jakub Jurasz, and Bartosz Kaźmierczak (2018) used Artificial Neural Networks and Multiple Linear Regression to predict future daily water consumption. The study was based on three antecedent records of water consumption and humidity forecast for a given day, which were considered to be independent variables. The study found that using the artificial neural networks and the multiple linear regression were both significant as they gave similar results. The study found that the (ANN) was superior to the (MLR) even though the mean absolute percentage error had a limited difference. The study also presented a novel approach to the division of sets into testing, training

and validation subsets. The conducted analysis confirmed the good accuracy of ANN and slightly lesser accuracy of MLR models. However, both those approaches were superior in comparison to a benchmark approach. The applied model in this paper can be also used for different parts of the water and wastewater system in each city. It is especially important in case of the volume of wastewater which is delivered to the wastewater treatment plants or the changes of the water pressure in pipelines. In both cases accurate forecasts can provide important information for people involved in various decision-making processes. David Walker, Enrico Creaco, Lydia Vamvakeridou-Lyroudia, Raziyeh Farmani, Zoran Kapelan, and Dragan Savic (2015) presented a preliminary study into forecasting domestic water usage with data collected during an ongoing project. The study employed an artificial neural network to predict the next time step's water usage for nine users from the project's Greek case study. Models were trained using an evolutionary algorithm, with parameterizations determined experimentally. The models resulting from the work used a range of input schemas, contrasting the use of historical input readings with inputs based on summary statistics constructed from those readings. While the peak and mean performance of the different schemas was similar, the worst case of the schemas using real historical values was lower than those using the statistical inputs. This can be attributed to the complexity of optimizing network weights for greater numbers of inputs. In all cases, the models failed to accurately predict the magnitude of consumption in peak cases.

CHAPTER 3: METHODOLOGY

3.1 Introduction

This chapter discusses the methodology that will be implemented in the study. It explains the assumptions of the model and the reason for its suitability. In section 3.2, the generalized linear model, its assumptions and its components are discussed. Section 3.3 discusses how parameter estimation using the maximum likelihood estimation method will be achieved, section 3.4 will discuss how the adequacy of the model is tested and section 3.5 will discuss how the forecasting of the response variable is done. The last section discusses how the accuracy of the forecasted response variable is achieved.

3.2 Generalized Linear Model

The GLM is a generalized form of linear models that extends the scope of linear models to allow for non-normality in the response variable. It provides a mean for modeling the relationship between one or more explanatory variables and a response variable whose distribution can be expressed in the form:

$$f(y,\theta,\phi) = \left[\frac{y(\theta) - b(\theta)}{a(\phi))}\right] + c(y,\phi) \qquad (3.1)$$

The above equation is derived from Julian J. Faraway, (2016),

where $a(\phi), b(\theta)$ *and* $c(y,\phi)$ are specific functions and Y is the response variable. The parameter Θ is the natural parameter and is used in relating the response Y to the covariates and it is a function of $\mu = E(y)$ while ϕ is the dispersion or the scale parameter.

The assumptions that underlie the GLM include:

i) The data $Y_1, Y_2, \ldots \ldots \ldots Y_n$ which represents the dependent variable is independently distributed.

ii) The dependent variable Y_i does not need to be normally distributed but should follow a distribution from the exponential family of distributions.

iii) GLM does not assume a linear relationship between the dependent variable and the independent variables, but it does assume a linear relationship between the transformed response variable in terms of the link function and the explanatory variables, for example, $log(\mu) = \alpha + \beta x$, for the Poisson regression.

iv) It is not a must for errors to be normally distributed though they should be independent.

v) Heteroskedasticity, that is, the variance need does not have to be the same since it is hard to achieve this, especially when overdispersion is present.

vi) It relies on large sample approximations since it uses MLE in parameter estimation.

The GLM model is made up of three components:

3.2.1 The Random Component

The random component refers to the probability distribution of the response/dependent variable Y_i. Each component Y_i is independent and is from one of the exponential family of distributions. The response variable should follow any of the distributions that belong to the exponential family. These include binomial, negative-binomial, Poisson, geometric, exponential, gamma, and normal, among others. The study considers the normal distribution which is a continuous distribution.

3.2.2 Normal Distribution

If the response variable follows a normal distribution with parameters μ and δ^2, then the PDF of the distribution is given by:

$$f_Y(y;\mu,\delta^2) = \left(\frac{1}{(\sqrt{2\pi\delta^2})}\right) exp\left(-\frac{(y-\mu^2)^2}{2\delta^2}\right) \quad (3.2)$$

The above equation is derived from Julian J. Faraway, (2016),

This can also be expressed as:

$$= exp\left[\frac{\left(y\mu - \frac{\mu^2}{2}\right)}{\delta^2} - \frac{1}{2}\left(\frac{y^2}{\delta^2} + log2\pi\delta^2\right)\right] \quad (3.3)$$

The above equation is derived from Julian J. Faraway, (2016),

Where $\theta = \mu$, $\emptyset = \delta^2$, $a(\phi) = \phi$, $b(\theta) = \frac{\theta^2}{2}$, $c(y,\theta) = -\frac{y^2}{\varphi} + log2\pi\delta^2$

this clearly shows that if the random variable *Y* follows a normal distribution, it can be fitted into a generalized linear model.

If the variable *Y* does not follow the normal distribution, it can be transformed using the Box-Cox method of transformation or other transformation methods to follow the normal distribution. Using the Box-Cox method of transformation transforms the data to follow a normal distribution which belongs to a two-parameter exponential family. The transformation of Y is of the form:

$$y = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases} \quad (3.4)$$

The above equation is derived from Julian J. Faraway, (2016),

where, λ is chosen to obtain the optimal transformation (one that results in the best approximation of a normal distribution curve) and the test only works for positive data. After this

transformation, the study will check for normality by visualizing the transformed data's histogram superimposed with a normal curve. The skewness and kurtosis coefficients will also be checked since the visualized test is subjective. If the skewness coefficient is close to 0 and the kurtosis coefficient close to 3, then the data will be assumed to follow normal distribution and can be visualized as shown in the graph below:

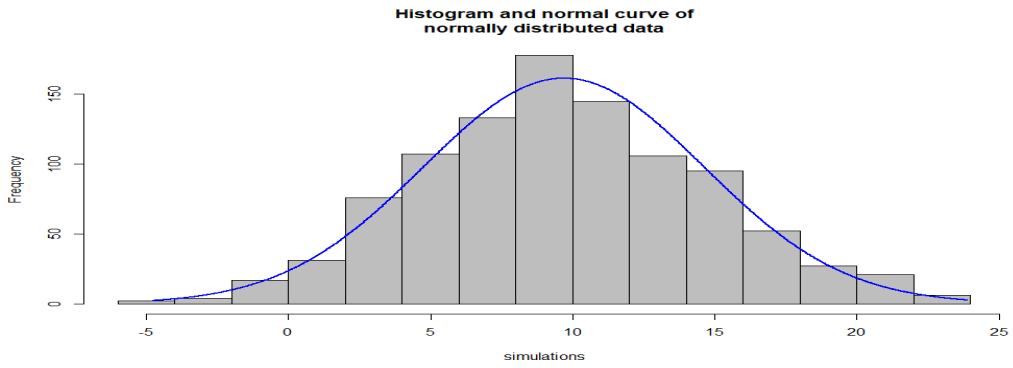

Figure 1- Histogram and normal curve of normally distributed data

3.2.3 The Linear Predictor

The study considers data which has explanatory variables that can be categorized into numeric variables and categorical variables. It focuses on the independent variables in the model and their linear combination in creating a linear predictor. GLM is a generalization of the general linear models that extends the scope of explanatory variables from being numeric to including categorical variables. The linear predictor for the model will be of the form:

$$\eta_i = \beta_0 + \beta_{1x_1} + \dots\dots + \beta_{p-1x_{p-1}}$$

$$= \beta_0 + \sum_{i=1}^{p-1} \beta_{ix_i} \qquad (3.5)$$

The above equation is derived from Julian J. Faraway, (2016),

where, η_i is the dependent variable, $X_1, X_2 \dots \dots X_p$ are the independent variables and β_i for $i = 1,2, \dots \dots p$ is the regression coefficient corresponding to variable i.

Similarly, the linear predictor for the factors will be of the form;

$$\eta_i = \alpha_j + \alpha_k + \dots + \alpha_z$$

where the $\alpha_b^{'s}$ for $b = j, k, \dots \dots, z$ are the parameters for the factors.

The study will also consider interaction terms and not only the main effects. The term yielded is referred to as an interaction term. This term will vary depending on its constituents as $\beta_{ix_ix_j}$ would be as a result of interaction between two variables, x_i and x_j, α_{ij}, as a result of interaction between two factors *i* and *j* and $\beta_{ix_ix_j}$ will be as a result of interaction between a variable *i* and a factor *j*.

The generalized linear predictor would be of the form:

$$\eta_i = \beta_0 + \sum_{i=1}^{p=1} \beta_i x_i + \sum_{b=j}^{z} \alpha_b + \sum_{i \neq j} \beta_{ix_ix_j} \qquad (3.6)$$

The above equation is derived from Julian J. Faraway, (2016),

3.2.4 The Link Function

This is a function that connects the mean response to the linear predictor, $g(\mu) = \eta_i$. Since what we are trying to do in a GLM is determine a relationship between the mean of the response variable and the covariates, then, by setting the link function $g(\mu) = \eta_i$ and assuming that the link function is invertible, we can make mean μ the subject of the formula as follows:

$$\mu = g^{-1}(\eta) \qquad (3.7)$$

There are a number of link functions associated with different distributions. However, they can be used interchangeably as they fit best to the given data. The commonly used link functions that could be chosen are shown in the table below:

Table 1 - Commonly used link functions

Distribution	$g(\mu)$	Link function
Normal	μ	identity
Poisson	$\log \mu$	log
Binomial	$\log\left(\frac{\mu}{1-\mu}\right)$	logit
Gamma	$\frac{1}{\mu}$	inverse

The study will use an appropriate link function to connect the mean response to the linear predictor.

3.3 Feature Selection

We need to reduce the dimensionality of the data by considering which features are important in determining the response variable. We will use the Boruta algorithm in R which uses the random forest algorithm of classification to select the important features with respect to the outcome variable. The algorithm will first duplicate the dataset and then shuffle the values for each column. The values are called shadow features. It will then train a random forest classifier on the dataset. The algorithm will then check the real features to determine if they have a higher importance. It does this by checking if the feature has a higher Z-score than the maximum Z-score of the shadow features compared to the best of the shadow features. If they do, it will record this in a vector. These are called hits. It repeats another iteration until the defined number of iterations are over and gives you a table of these hits. At each repetition, the algorithm will compare the Z-score of the reordered copies of all the features and the original features to determine whether the latter performs better than the former. If it does, it marks the

feature as an important feature. In essence, the algorithm tries to validate feature importance by comparing with random shuffled copies which increase the robustness. This can be achieved by using a binomial distribution to compare the times a function has done better with the shadow features.

If, let's say, 20 iterations do not report a function as a hit, we will reject it and delete it from the original matrix. After all the features have been verified, after the required number of iterations, the algorithm stops and prints the results.

3.4 Parameter Estimation of the Fitted Model

In this section, the maximum likelihood estimation method of parameter estimation will be discussed. The maximum likelihood estimation method is used to estimate the parameters of the fitted model because it possesses desirable properties which are the invariance property, consistency and minimum variance. The invariance property states if g is a monotonic function and $g(\theta)$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $\hat{\theta}$. In consistency, the estimate $\hat{\theta}$ is consistent if $\hat{\theta} \rightarrow \theta$ in probability as $n \rightarrow \infty$, where θ is the true parameter of the distribution of the sample. Finally, in the class of all estimators, for large samples, $\hat{\theta}$ has the minimum variance and is therefore the most precise estimate possible. This property is known as the asymptotic normality because as $n \rightarrow \infty$, then θ, is approximately normal, and is unbiased with variance given by the Cramer-Rao lower bound, that is:

$$CRLB = -nE\left[\frac{\sigma^2}{\sigma\theta^2} logf(x:\theta)\right]^{-1} \tag{3.8}$$

The above equation is derived from Julian J. Faraway, (2016),

The last property that makes the MLE to be desirable is that it is asymptotically efficient in that, for large n, it is unbiased with a variance equal to the lowest possible value of unbiased estimators. The MLE for the normal distribution under consideration will be obtained as follows:

3.4.1 Normal Distribution

Consider a random variable $Y \sim N(\mu, \delta^2)$ whose PDF is:

$$f(y) = \frac{1}{\sqrt{2\pi\delta^2}} \times \exp\left(-\frac{(y-\mu)^2}{2\delta^2}\right) \quad (3.9)$$

Where $y_i's$ for $i = 1,2,3, \ldots \ldots, n$ are n sample response variables.
The likelihood function of the normal distribution will be given by:

$$L(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\delta^2}} \times \exp\left[-\frac{(y-\mu)^2}{2\delta^2}\right]$$

$$= (2\pi\delta^2)^{\frac{-n}{2}} exp\left(-\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\delta^2}\right) \quad (3.10)$$

Where $y_i's$ for $i = 1,2, \ldots \ldots, n$ are n sample response variables.
to get the maximum estimate easily, the study uses the log-likelihood function which takes the form:

$$l(y_i) = \log \mathrm{L}(y_i)$$

$$= -\frac{n}{2}\log(2\pi\delta^2) - \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\delta^2} \quad (3.11)$$

The identity link function is used to connect the mean response and the linear predictor in case of a normal distribution. The identity link function is given by:

$$\mu = \sum_{i=1}^{n} \beta_i x_i + \sum_{i \neq j} \beta_{ij} x_i \quad (3.12)$$

The above equations are derived from Julian J. Faraway, (2016),

3.5 Model Selection

The adequacy and accuracy of the model will be tested using Akaike Information Criterion (AIC) and the R-squared and the adjusted R-squared as discussed below. Further model selection factors that will be included are a test for multicollinearity and interaction between the independent variables.

3.5.1 Akaike Information Criterion

The Akaike information criterion (AIC) will be used to select the most suitable generalized linear model for water consumption. The model with the least AIC will be considered as the best model. This is because the criterion gives an estimate of the relative information lost when a given model is used to represent the process that generated the data. The selection of the best model based on the AIC will be done using the stepwise regression method which uses both backward and forward selection methods. This stepwise regression uses the AIC values and selects the model with the least AIC value.

To calculate the value of AIC, the formula below is used:

$$AIC = -2\log(\tau) - 2\mathrm{K} \tag{3.13}$$

The above equation is derived from Julian J. Faraway, (2016),

where τ is the maximized value of the likelihood function of the distribution, and K is the number of estimated parameters in the model.

3.5.2 R Squared

The coefficient of determination also referred to as R^2, is a statistical measure of the goodness of fit of the model. It measures how well the fitted regression line approximates the observed/actual data. It tells how much variation (which can be expressed as a percentage) of the response variable is explained by the explanatory variables in the regression model. Higher values indicate smaller differences between the observed values and the predicted values. It takes the values between 0 and 1 as it is a percentage. For the ordinary least squared models, it is calculated by using the following formula:

$$R^2 = 1 - \frac{sum\ squared\ regression(SSR)}{total\ sum\ of\ squares\ (SST)}$$

$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (3.14)$$

The above equation is derived from Julian J. Faraway, (2016).

Where y_i represent the observed values, $\hat{y}_i$represents the predicted values of y and $\bar{y}$ is the mean of the observed values.

R-squared has a limitation of not being able to tell whether the predictions are biased, which is why we will use a residual plot. We will use the adjusted R squared since as more variables are added to a model the mode the value of R squared rises. That's why R squared is best for simple linear regression while the adjusted R squared is best for regression with more than one independent variables. The difference between the formulas of the two is that the adjusted R squared includes the degrees of freedom as shown below:

$$R^2 = 1 - \frac{SSR/df_r}{SST/df_t} \quad (3.15)$$

The above equation is derived from Julian J. Faraway, (2016),

Where df_r is the degrees of freedom, n – 1 of the estimate of the population variance of the response variable, and df_t is the degrees of freedom, n – p – 1 is the estimate of the underlying population error variance.

3.5.3 Multicollinearity and Interactions

Multicollinearity is when there is correlation between the independent variables in the model. In the presence of multicollinearity in a model, the standard errors of the regression coefficients estimate become increasingly large and the parameters of the model also become indeterminate. The variance inflation factor is used to detect multicollinearity in the regression model. It is calculated by using the formula below:

$$VIF = \frac{1}{1 - R_i^2} \quad (3.16)$$

The above equation is derived from Julian J. Faraway, (2016),

Where R_i^2 is the value of R-squared calculated by regressing the j^{th} predictor on the other remaining predictors.

A VIF value above 5 should be further investigated while values above 10 show presence of serious multicollinearity. Interaction in regression model is where the effect of an explanatory variable on the response variable changes depending on the value of one or more other explanatory variable(s). It is represented as the product between two or more explanatory variables. Below are two regression equations where the first one has no interaction and the second one has:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 \quad (3.17)$$

The above equation is derived from Julian J. Faraway, (2016),

We always avoid higher-order or levels of interactions since they are complex and hard to interpret.

3.6 Forecasting Methods and it Types

Forecasting methods are used by the analysts in order to make decisions when there is a 'what-if' question on certain processes or activities. It is used to identify the appropriate responses for the demand level changes, price reduction by the competitors, and fluctuations in economics. To find the greatest benefits from the forecasts, analysts have to choose the correct forecasting method out of all the available methods and the best suitable one for the particular need .The predictability of an event or a quantity depends on the several factors including:

- How well the factors are understood and contributed towards the decision.
- How much data are available to make a decision.

- Whether the forecast will affect the thing that we forecast.

The forecasting techniques must be accurate in some cases. For example, electricity consumption forecasting, water consumption and weather forecasting requires accurate forecasting methods. In this case, a good model is required to link the variables and decisions. On the contrary, when calculating currency exchange rates, we can rely on simple forecasting methods such as moving averages.

3.6.1 Types of Forecasting

There are three major categories of forecasting; they are short-term, medium-term, and long-term forecasts. Short-term forecasting is required for personnel scheduling, production of goods, and transportation of vehicles. Medium-term forecasting is required to assess potential capital needs for the procurement of new raw material. Long-term forecasting is required in strategic planning. These decisions are based on market opportunities, environmental factors, and internal resources. Forecasting systems demand expertise in the identification of forecasting problems, implementing a range of forecasting methods, a selection of methods based on the problem, and evaluating the methods with respect to time and refining forecasting methods over time (Hyndman R.k , 2011). There are many types of forecasting methods qualitative , quantitative forecasting and judgmental forecasting

3.6.2 Advantages of Forecasting

Forecasting approaches provide knowledge that can help managers focus on potential plans and on the organization's objectives. Forecasting is also used to forecast raw material costs, schedule acceptable numbers of staffing, help to define inventory levels and a variety of other operations. Forecasting is carried out in two ways: local factors and external factors where the

external factors are controllable and sometimes non-controllable. The non-controllable entities are the national economy, governments, customers, and competitors (Slideshare, 2018).

3.6.3 Qualitative and Quantitative Forecasting Methods

Personal views are the basis for qualitative forecasting, and historical numerical data are the basis for predicting the future. Examples of qualitative forecasting approaches are the usage of the Delphi system educated opinions and the historical life-cycle comparison. Similarly, simple exponential smoothing, multiplicative seasonal indexes, simple and weighted moving averages are examples of quantitative forecasting methods (Bizfluent, 2018).

Quantitative forecasts use historical data, production and financial reports, and statistics. The use of forecasting depends on the statistical modeling, trend analyses and other sources. There are some other sources that are used for forecasting - they are government agencies, trade associations, and academic institutions. Qualitative forecasting is developed from the experience and opinions of business experts. These forecasts are made based on the interpretation of data combined with professional expertise over time (Bizfluent, 2018).

Qualitative judgments are made by human judgment and opinions. They are subjective and non-mathematical. This can incorporate the latest information in the environment and 'inside information this sometimes can lead to bias in forecasting accuracy. Qualitative judgments have unstructured data, and they are based on interviews, focus groups, and observations.

Quantitative judgments are based on numerical data and mathematical calculations. They are consistent and objective in nature since the decisions are taken based on numerical calculations. They have the possibility of including more information. Quantitative judgments are based on structured data and include statistical analysis. They have objective conclusions, surveys, and experiments to support the decisions.

There are three quantitative tools available. They are trend analysis, seasonal judgment, and graphical method. Trend analysis is based on the method for forecasting sales data when there is an upward and downward pattern that exists in the data or process. In seasonal judgments, they are based on the variations from season to season. In the graphical method, the information is plotted in the graphical form. The information in the spreadsheet is converted into a graphical form. Extrapolation can be used to predict future demands whereas trends and patterns are easier to spot. Similarly, quantitative forecasting is based on market research, which is taken based on surveys. Focus groups are a type of qualitative forecasting method, which consists of all panels of customers who will be providing opinions about the product or service. Another method is the panel consensus, where a group of people provide the forecasting and the facilitator brings a consensus decision (Slideshare, 2017).

3.6.4 Naïve Forecasting Methods

This is the method of future forecasting based on past records. A naïve forecasting method will be based on the prior period's data. This method also depends on the average of the actual values for certain periods. It makes no adjustments in past records in order to estimate future records. It is mostly used to create a forecast to check the results of more sophisticated forecasting methods (Bizfluent, 2018).

In this method, all forecasts are set to the value of the last observation. That is,

$$y_{T+h/T} = y_T \text{……...(3.18)}$$

Equation derived from Otexts. (2020).

This method works well in many economic and financial time series. Naïve forecasting is the starting point for much statistical forecast development. For some products, it is difficult to

improve the accuracy by using only the naïve forecast. This model is available in many forms and they are simpler.

3.6.5 Mean Absolute Percent Error (MAPE)

It measures the error in percentages. It is calculated as the average of the unsigned percentage error.

$$\frac{1}{n}\sum \frac{|actual - forecast|}{|actual|} * 100 \quad \text{.....(3.19)}$$

Equation derived from Forecastpro. (2020).

3.6.6 Mean Absolute Deviation (MAD)

MAD measures the value of the size of the error in units. It is calculated as the average of the unsigned errors.

$$\frac{1}{n}\sum |actual - forecast| \quad \text{.....(3.20)}$$

Equation derived from Forecastpro. (2020).

3.6.7 Mean Squared Error (MSE)

MSE measures the average value of the squares of the errors in units.

$$MSE = \frac{1}{n} \Sigma (y_i - \widetilde{y_i})^2 \quad \text{.....(3.21)}$$

Equation derived from Freecodecamp. (2018, October 8).

3.7 Judgmental Forecasting Methods

This method is applied when historical data is inaccessible when a new product is introduced, when the market faces new competitors or exceptional marketing situations. As per research, judgmental forecasting works better when the forecast has important domain knowledge, more timely and up-to-date information. The judgmental forecasting has been

improved significantly in recent years because of well-structured and systematic approaches. It is subjective and comes with limitations.

The judgmental forecasting is used based on three general settings:

- There is a non-availability of data and no possibility to apply statistical methods and judgmental forecasting is the feasible approach.
- There is the availability of data and statistical forecasts are generated and they are adjusted using judgment.
- There is the availability of data and statistical and judgmental forecasts are generated independently and combined (Otexts, 2020).

3.7.1 Limitations

Judgmental forecasts can be inconsistent since the decisions are taken based on the opinions of the experts. They mostly depend on human cognition and they are vulnerable to its limitations. Human judgment can vary based on psychological factors. Judgments can be combined with personal and political agendas. Another important factor in judgmental forecasting is attachment where the subsequent forecasting tends to converge or to be close to an initial reference point. The forecaster is influenced by the prior information and gives more weight in the forecasting process. Attaching will lead to a bias and undervaluing new information and create a systematic bias (Otexts, 2020).

3.7.2 Key Principles

The objective of the forecasting task must be clear and well defined. The definitions should be clear and comprehensive, and it must avoid ambiguity and vague expressions. It is better to have prior data collection before starting the task. The systematic approach must be implemented so that accuracy of forecasting and consistency can be improved by having a

checklist of categories of information that are relevant to the forecasting results. The decisions have to be documented and formalized in order to be used as a reference in the future.

In judgmental forecasting, the irregularities in the forecasting would be identified by systematic monitoring. Feedbacks of the forecasting can be recorded, and it can be used as a reference. Since time is in constant change, the forecaster should have feedback and records to back up their decisions by which the forecasting accuracy will be improved. The isolation between the forecasters and users must be remained all periods. It is important for the forecasters to communicate to potential users thoroughly. The process can be explained to the users and it can be justified by the forecasting methods in order to make assurance to the users (Texts, 2020).

3.7.3 Recommendations

- The guidelines for forecasting new policy must be developed to encourage more systematic and a structured forecasting approach.
- Forecasting methodology must be documented, including all assumptions made in the forecasting.
- New policy forecasts must be framed by at least two people from different sections of the organization.
- Once in a year revision must be carried out on the forecasting, especially on the new policies, (Texts, 2020).

3.8 Delphi Method

In this method, decision making by a group of people is encouraged instead of individuals. A facilitator will coordinate the process in this method. There are many stages in Delphi method. The first one is forming a panel of experts. Then, the distribution of forecasting tasks is assigned to the experts. In the third stage, the initial findings are collected and provided

for feedback. At the fourth stage, the experts will review the forecasting based on the feedback and this process will continue until the final consensus decision is taken (Texts, 2020).

In the Delphi method, the challenge is finding experts from diverse fields. It requires finding five to 20 experts for a panel. It is important to keep the experts anonymous so that they won't be influenced by any political or social pressures in their forecasts. In this method, the experts are given the chance to speak and be accountable for their forecasts. Here, the group meeting is avoided so that there is no domination of a few people and no seniority or other influential factors. This method increases the chances of communicating with experts with a variety of skills and expertise from many places. This process makes the method cost-effective by eliminating the expenses of travel, and others.

3.8.1 Limitations

This method is very time-consuming. The decision could be taken in a few minutes or in hours in group meetings, and experts may lose interest and cohesiveness after a long-time. There is an alternative to the Delphi method, known as the "Estimate talk estimate" method, in which experts can interact between the iterations even though the forecast submissions remain anonymous (Texts, 2020).

The Delphi method, scenario building, statistical surveys, and composite forecasts are judgmental forecasting methods. They are considered to be judgmental methods based on intuition and subjective estimates. The methods produce a prediction based on the opinions of the managers and experts (Disfluent, 2018).

3.9 Time Series & Regression Models

In this regression models, the time series of interest 'y' assumes that it has a linear relationship with other time series. The forecast variable 'y' is called the dependent variable. The predator variables 'x', independent or explanatory variables.

3.9.1 Simple Linear Regression

In this method, the linear relationship between the forecast variable y and a single predator variable x is given by:

$$Y_{et}=\beta_0+ \beta_1 x_t+\xi_t \text{……...(3.27)}$$

Equation derived from Texts. (2020).

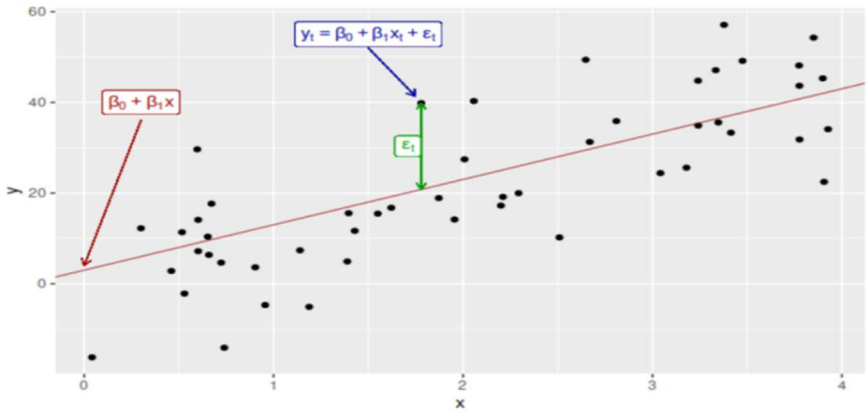

Figure 2 - Linear regression model (Otexts, 2020)

In this analysis, the coefficients β_0 and β_1 represent the intercept and slope of the line. When x=0, the intercept β_0 represents the predicted value of 'y'. The slope β_1 represents the average predicted change in 'y' resulting from a one-unit increase in 'x'. The observations don't fall on the straight line, but they are scattered around it. Each observation y_{et} consists of the systematic and explained part of the model, $\beta_0 + \beta_1 x_t$, and the random error, a_t. The error doesn't imply a mistake, but a deviation from the underlying line model (Texts, 2020).

3.9.2 Multiple Linear Regression

When there are two or more predictor variables involved, then this model is called a multiple regression model. The general form of a multiple regression model is:

$$Y_{et}=\beta_0+\beta_1x_{1,t}+\beta_2x_{2,t}+\beta_3x_{3,t}+\beta_4x_{4,t}+\ldots\ldots+a_t \quad \text{.....(3.28)}$$

Equation derived from Texts. (2020).

Where 'y' is the variable to forecast and x_1, x_2, ….X_u is the k predictor variables. The coefficients β_1, ... β_k measure the effect of each predictor after taking the accounts of all the other predictors in the model.

The linear regression model is a reasonable approximation to reality. The relationship between the dependent variable and the predictor variables satisfies the linear equation. As far as errors are concerned, the following are the assumptions:

- They have mean zero, otherwise, there will be a systematic biasing in forecasting.
- They are auto correlated. Otherwise, the forecasts will be inefficient.
- They are unrelated to the independent variables.

In order to produce prediction intervals, the errors have to be distributed with constant variance. The most essential thing in the linear regression is that each predictor 'x' is not a random variable. If an experiment is controlled in a laboratory, the values of each 'x' are controlled, and the resulting values of 'y' are observed (Texts, 2020).

3.9.3 Mean Absolute Percent Error (MAPE)

It measures the error in percentages. It is calculated as the average of the unsigned percentage error.

$$=\frac{1}{n}\sum\frac{|actual-forecast|}{|actual|}*100 \quad \text{.....(3.29)}$$

Equation derived from Forecaster. (2020).

3.9.4 Mean Absolute Deviation (MAD)

MAD measures the value the size of the error in units. It is calculated as the average of the unsigned errors.

$$= \frac{1}{n} \sum |actual - forecast| \quad \text{.....(3.30)}$$

Equation derived from Forecaster. (2020).

3.9.5 Mean Squared Error (MSE)

MSE measures the average value of the squares of the errors in units.

$$MSE = \frac{1}{n} \Sigma (y_{es} - \bar{y}_{es})^2 \quad \text{.....(3.31)}$$

Equation derived Freecodecamp. (2018, October 8).

3.9.6 Exponential Smoothing Methods

Forecasting methods used in this approach are weighted averages of past observations along with the weights decreasing exponentially with respect to time. There are two categories of exponential smoothing methods. They are based on the application in forecasting time series with various characteristics. The second part includes exponential smoothing methods. It is a more advanced method than the moving average. It uses very little record of data for forecasting. In this method, the new forecast is given by the following equation:

New forecast = last period forecast + α (last period's actual demand – last period's forecast) where α is the smoothing constant and has values between 0 and 1.

Equation derived from FORECASTING FUNDAMENTALS. *(*2020)

3.9.7 Simple Exponential Smoothing

This method looks at forecasting data with no clear trend or seasonal pattern. In exponential forecasting methods, they are calculated based on weighted averages, where the weights decrease exponentially, and the smallest weights are related to the past observations.

$$y_{T+1|T}=\alpha y_T+\alpha(1-\alpha)y_{T-1}+\alpha(1-\alpha)^2y_{T-2}+\cdots,$$

In this equation, $0 \leq \alpha \leq 1$ is the smoothing parameter. The parameter α controls the rate at which the weights decrease. The following table shows the weights attached to the different weights for a small value of 'α' that will be approximately calculated as follows:

	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$
y_T	0.2000	0.4000	0.6000	0.8000
y_{T-1}	0.1600	0.2400	0.2400	0.1600
y_{T-2}	0.1280	0.1440	0.0960	0.0320
y_{T-3}	0.1024	0.0864	0.0384	0.0064
y_{T-4}	0.0819	0.0518	0.0154	0.0013
y_{T-5}	0.0655	0.0311	0.0061	0.0003

Figure 3 - Different weights for the value of α

In this experiment, the values of weights are decreasing exponentially for 'α' between 0 to 1. That is why it is called an exponential smoothing technique. In this experiment, if α is small and if it is close to 0, then more weight is given to observations from the past. If 'α' is large, and if it is close to 1, more weight is given to recent observations.

The alternative for simple exponential smoothing is component form. This form represents the forecast equation and the smoothing equation for all the components that are added in this method. This component form includes forecast equations and smoothing equations. The forecast equation indicates that forecast time at time t+1 is nothing but it is the estimated level at the time, t. The smoothing equation for the level provides the estimated level of the series at the point at each period, t.

Optimization is very essential in simple exponential smoothing. It needs the smoothing parameters and the choice of the initial values appropriately. The smoothing parameters can be chosen in a subjective manner. The better way to obtain values for the unknown parameters is to

estimate them from the observed data. The minimization of SSE is done in order to estimate the unknown parameters and the initial values for exponential smoothing (Otexts, 2020).

3.9.8 Holt's Linear Trend Method:

This method decides on the forecasting of data along with a trend and this method involves a forecast equation and two smoothing equations.

Forecast equation: $y_{t+h|t}=\ell_t+hb_t$……...(3.34)

Level equation: $\ell_t=\alpha y_t+ (1-\alpha)(\ell_{t-1}+b_{t-1})$ ……...(3.35)

Trend equation: $b_t = \beta* (\ell_t-\ell_{t-1}) + (1-\beta*) b_{t-1}$……...(3.36)

Equations derived from Otexts. (2020).

Where, ℓ_t denotes an estimate of the level series at time t, b_t denotes the estimate of the trend of the series of the time t, α is the smoothing parameter for the level, $0 \leq \alpha \leq 1$, and β is the smoothing parameter for the trend, $0 \leq \beta \leq 1$. With simple exponential smoothing, the level equation shows that ℓt is the weighted average of observation yt and the one-step-ahead training forecast for time, t that is given by lt-1 +bt-1. The forecast function is no longer flat, but it is trending. The h-step ahead forecast is equal to the last estimated level, added with 'h' times the last estimated trend value. Hence, the forecasts are considered as a linear function of 'h.'

3.9.9 Damped Trend Methods

This method indicates that they are used to over-forecast, especially for longer forecast horizons. The methods that include the damped trend have proven that they are more useful. They are popular individual methods when forecasts are required automatically for many series (Otexts, 2020).

3.9.10 Estimating Error, Trend and Seasonal Models

The maximization of likelihood is an alternative method to estimate the parameters by minimizing the sum of squared errors. The likelihood is the probability of the data is developed from the specified model. In the additive error model, the results are the same in maximizing the likelihood and minimizing the sum of the squared errors. Different results would be obtained for multiplicative error models. In this model, the smoothing parameters α, β, Υ and ϕ, and the initial states $l_o, b_o, s_o, s_{-1}, \ldots, s_{-m+1}$ by maximizing the likelihood. The parameters have been manipulated to lie between 0 and 1 so that equations can be interpreted as weighted averages.

State-space models are another way to view the parameters, through a consideration of the mathematical properties. The parameters are constrained to restrict the observations in the distant past, which will be having a continuing effect on current forecasts. This leads to admissibility constraints on the parameters are less restrictive than the traditional constraint regions (Otexts, 2020).

3.10 Model Selection

The advantage of the ETS statistical framework is that information criterion will be used for model selection. The AIC , AICc and BIC can be used to determine the most appropriate Error ,Trend and Seasonal models for a given time series.

AIC, Akaike's Information Criterion is defined as:

$$AIC = -2 \log (L) + 2k \quad \ldots\ldots\ldots\ldots\ldots\ldots (3.37)$$

Where 'L' is the likelihood of the model and 'k' is the total number of parameters and represents the initial states that have been estimated.

The AIC correlated for sample small sample bias (AICC) is defined as:

$$AIC = AIC + k [\log (T) - 2] \quad \ldots\ldots\ldots\ldots\ldots\ldots (3.38)$$

The Bayesian Information Criterion (BIC) is:

$$BIC = AIC + k [\log (T) - 2] \text{ ……...(3.39)}$$

Equations derived from Otexts. (2020).

Models with multiplicative errors are useful when the data is strictly positive, but they are not being numerically constant when there are zero and negative values. Multiplicative error models will not be considered if the time series is not strictly positive where only six additive models can be applied (Otexts, 2020).

3.10.1 ARIMA Model

This is another method in time series forecasting. Exponential smoothing and the Auto Regression Integrated Moving Average model are mostly used in time series forecasting. ARIMA models are based on the correlation of data. The autoregressive model is one of the types of ARIMA models. In this model, the variable of interest is forecasted by implementing the linear combination of the variable's past values. Autoregression is nothing, but it is a regression of the variable against it.

The autoregressive model of the order 'p' can be written as:

$$Y_t = c + \phi_{1\ yt-1} + \phi_1\ y_{t-1} + \ldots\ldots + \phi_1\ y_{t-1} + \xi_t \text{……...(3.40)}$$

Equation derived from Otexts. (2020).

Where ξ_t is the white noise. It is like a multiple regression but lagged values of y_t as predictors. This is referred to as an AR (p) model, known as an autoregressive model of order. Autoregressive models are good in handling the large range of different time-series patterns. The variance of the error term ξ_t will only change the scale of the series, not the patterns (Otexts, 2020).

3. 10.2 Non-Seasonal ARIMA Model

It is obtained by combining the autoregression and moving average model. This equation can be written as:

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon t$$

where y′t is the differenced series. It is called as ARIMA (p, d, f) model, where:

p = order of the autoregressive part

d = degree of the first differencing involved

q = order of the moving part

The stationary and invertibility are applied for autoregressive and moving average models, which are also applied to the ARIMA model. The ARIMA function in R uses the Hyndman-Khandakar algorithm variation, which combines unit root tests, minimization of the AIC, and MLE to get the ARIMA model. The argument auto. ARIMA () provides many variations in the algorithm. There is a modeling procedure, which is used to fit the ARIMA model.

1. The data is to be plotted, and any unusual observations have to be identified.
2. Transform the data to stabilize the variance.
3. If the data are non-stationary, the data's first differences are taken first until the data are stationary.
4. The chosen model is tried, and AIC is used to search for a better model.
5. When the residuals look like, forecasts are calculated (Otexts, 2020).

3. 10.3 Mean Absolute Percent Error (MAPE)

It measures the error in percentages. It is calculated as the average of the unsigned percentage error.

$$= \frac{1}{n} \sum \frac{|actual - forecast|}{|actual|} * 100 \quad \text{.....(3.42)}$$

Equation derived from Forecastpro. (2020).

3. 10.4 Mean Absolute Deviation (MAD)

MAD measures the value the size of the error in units. It is calculated as the average of the unsigned errors.

$$= \frac{1}{n}\sum |actual - forecast| \text{……...(3.43)}$$

Equation derived from Forecastpro. (2020).

3. 10.5 Mean Squared Error (MSE)

MSE measures the average value of the squares of the errors in units.

$$MSE = \frac{1}{n}\Sigma (y_i - \widetilde{y}_i)^2 \text{……...(3..44}$$

Equation derived from Freecodecamp. (2018, October 8)

CHAPTER 4: RESULTS AND DISCUSSION

4.1 Introduction

This chapter describes data analysis and findings. Section 4.2 covers the data description and the data cleaning and preprocessing while section 4.3 covers the selection of the most appropriate model using the stepwise regression based on the Akaike Information Criterion values while section 4.4 discusses the results of the generalized linear model fitted using the normal distribution. Section 4.5 deals with prediction of values that the response variable takes under the fitted model and then in section 4.6, the accuracy of the fitted model is assessed using R squared and the adjusted R squared values.

4.2 Data Description and Pre-processing

4.2.1 Data Description

Data on the amount of water consumption in milliliters for different individuals was collected and recorded. The data was collected for a period of 30 days and other features of these individuals were also recorded. These features are the age of the individuals, weight, height and whether they had any of the following diseases: hypertension, diabetes, kidney problems, thyroid problems, allergy, asthma, heart disease, post-traumatic stress disorder and irritable bowel syndrome. The total number of individuals who were recorded was 79 and the total number of observations made were 2,150

4.2.2 Data cCeaning and Pre-Processing

Before doing the analysis, it was important to first clean and prepare the data. After reading the dataset into R, we ensured that the different types of variables were recognized correctly as numeric or factors. The variables with yes and no should be recognized as factors in R, but R read them as characters. We converted them into factors. We proceeded to check for missing values in the dataset and summed them using the function sum(is.na()) which gave a zero, meaning that there were no missing values in the data.

The next step was to do data visualization using ggplot2 package in R and plot boxplots and scatter plots. These visualizations helped in getting more insights about the data. From the boxplots we saw that most of the variables contained outlier observations. The scatter plots of age, height and weight against consumption show where most of the observations are distributed using the dots. The Figure 4 below shows combined box plots and scatter plots for consumption against age, weight, height and diabetes.

We then proceeded to check for any outlier observations in the dataset, study them and determine whether to drop them. We did this by creating a function to check for outliers and plot histograms and boxplots of the data with outliers and without outliers. We dropped these outliers and proceed to use the data without outliers, since they increased the variability in our data and reduced the statistical accuracy of our model. The Figure 5 below shows the plot of the histograms and boxplots of the data with outliers and without outliers.

The next important step was to determine the probability distribution of the data so as to know which model we should use for forecasting. The response variable to be forecasted was consumption. We extracted this variable and analyzed it. We ran the function basicStats (Consumption) in R and got the basis statistics of the response variable. We noticed that the

mean (4.497) is greater than the variance (1.567), therefore the distribution cannot be Poisson or Binomial. We proceed to check if the data is normally distributed by plotting a histogram and fitting a normal line. The line had a bell shape which confirmed that the data was normally distributed. The qqplot had the points lying along the qqline which further confirmed that the data was normally distributed. Figure 6 shows both plots. However, from the basicStats results, the skewness is close to zero which is perfect, but the kurtosis is far from 3, which deviates from the assumption of normal distribution. We tried the box cox method of transformation into normal distribution, which was the best method after iteratively going through all the methods using the bestnormalize() function from bestnormalize package in R, but it does not work. Figure 7 shows the plots of the transformed response variable using the box cox method of transformation and how they deviate from a normal distribution. In fact, they increase the skewness and lower the kurtosis to negative values totally deviating it from a normal distribution.

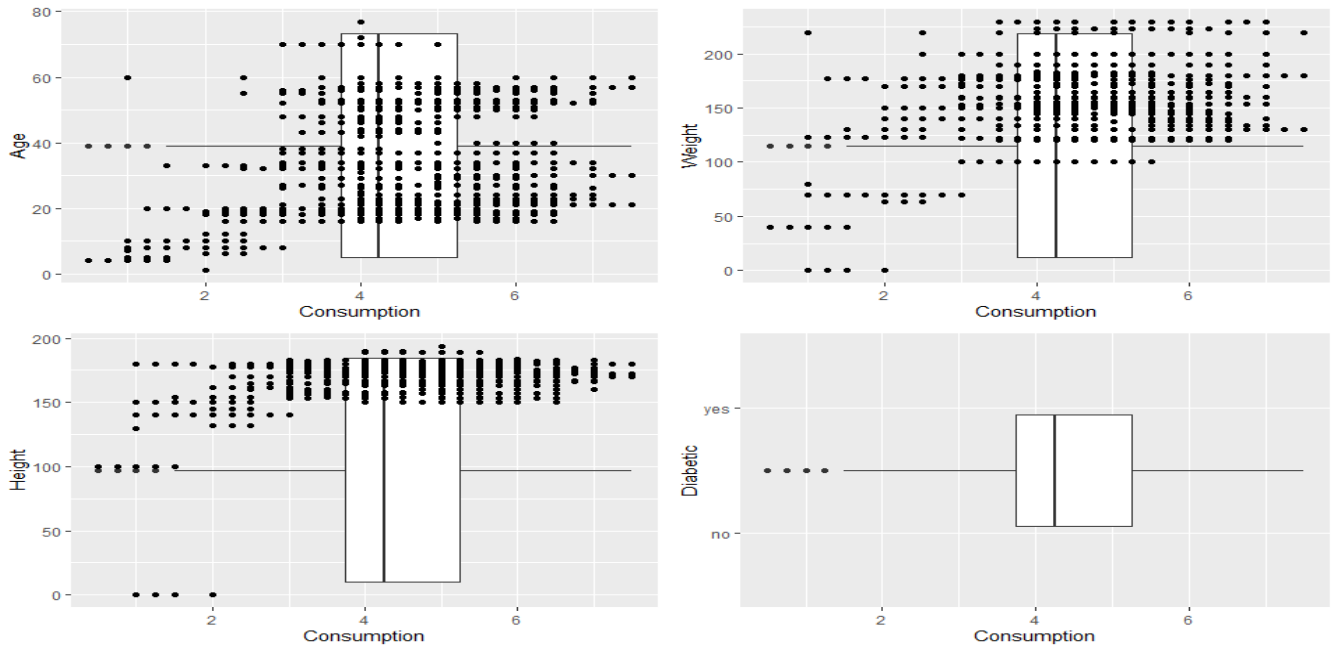

Figure 4 - combined box plots and scatter plots for consumption against age, weight, height and diabetes.

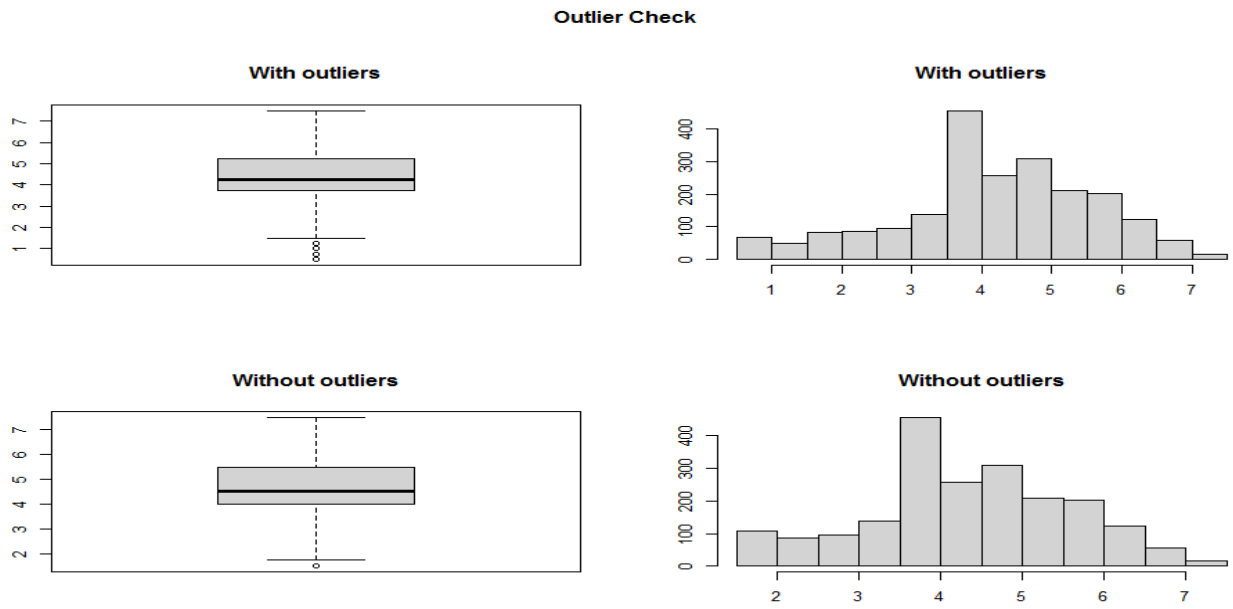

Figure 5 - Histograms and boxplots of the data with outliers and without outliers.

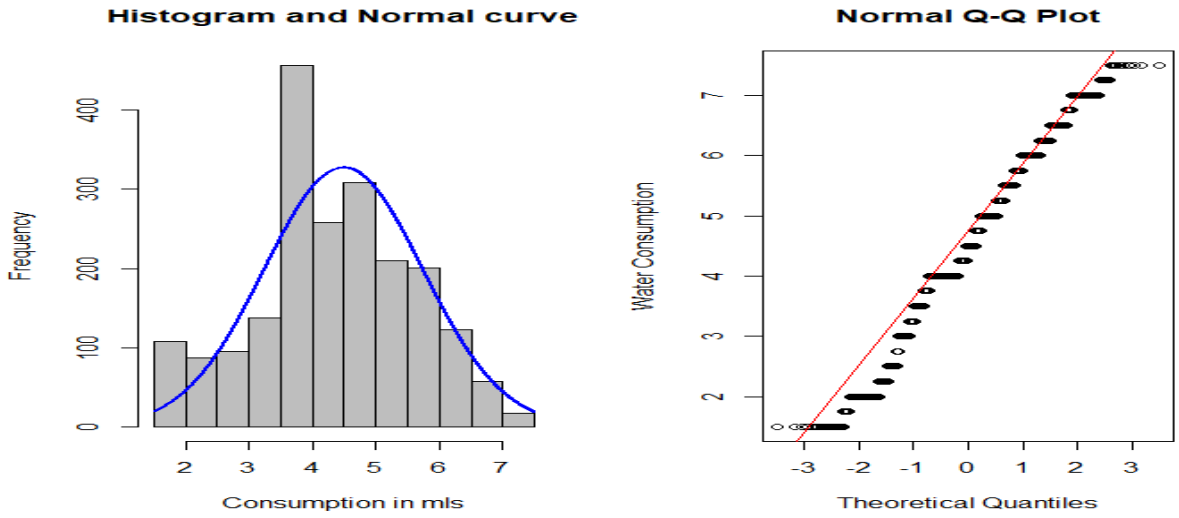

Figure 6 - Histogram and Normal Curve/Normal Q-Q Plot

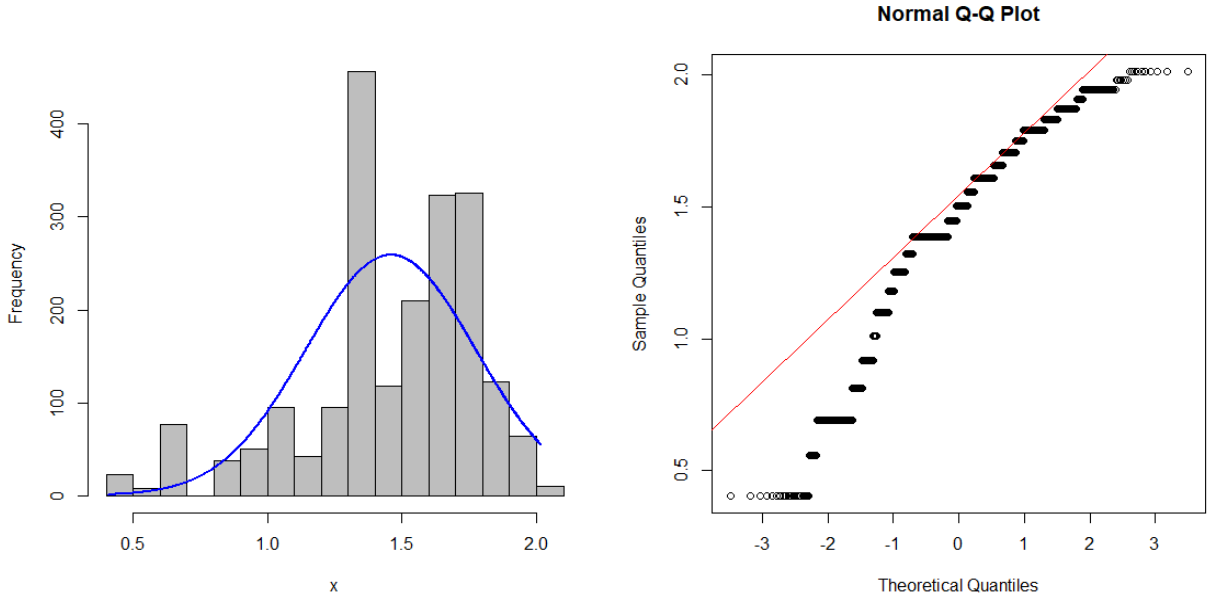

Figure 7 - Normal histogram and qq-plot of the response variable

4.2.3 Feature Selection

We then proceeded to do variable selection (feature selection) using a Boruta package. We want to determine which variables are important in determining the amount of water consumption. The Boruta package performed the Boruta algorithm based on a random forest classification algorithm. The results we got were that all variables are important in determining the amount of water consumed. The result from R below confirms that, as seen in the decision column:

Table 2 - Attstats from feature selection

	MeanImp	MedianImp	MinImp	MaxImp	NormHits	Decision
Age	45.45903	45.37472	43.70023	47.35892	1	Confirmed
Weight	38.42920	38.49991	35.83016	40.11152	1	Confirmed
Height	39.73609	39.58693	38.31872	42.67456	1	Confirmed
Hypertension	19.20157	19.20805	17.96975	20.35018	1	Confirmed
Diabetes	16.15565	15.80109	14.13186	17.84538	1	Confirmed
Kidney	17.59151	17.35635	16.75370	18.71334	1	Confirmed
Thyroid	21.63427	21.65884	19.99021	23.30543	1	Confirmed
Allergy	16.46132	16.86732	14.74359	17.09706	1	Confirmed

Asthma	17.09477	17.09665	15.77908	19.10066	1	Confirmed
Heart disease	18.53837	18.89933	16.95968	19.89445	1	Confirmed
Post-traumatic stress disorder	18.63838	18.49358	18.49358	19.93836	1	Confirmed
Irritable bowel syndrome	21.30209	21.30885	20.03864	22.09509	1	Confirmed

4.3 Model Fitting and Selection

After the data was cleaned and prepared, we proceeded to create the model which was to be used for prediction. We divided the dataset into two partitions, one for training the model and the second one for testing the model. We did this to ensure that the model would make predictions on a new dataset it was not trained on. We then fitted a generalized linear model on the training dataset. This was the first and the initial model.

We then used the backward and forward stepwise regression to select the best model with the highest accuracy in prediction the response variable. We did this by carrying out the following on the initial model: first was stepwise regression, second was test for multicollinearity, then tested for interaction and finally treating outliers in the model.

In the stepwise regression method, we used both forward selection and backward selection based on the AIC values. The forward selection involved starting with a model with zero explanatory variables and adding one variable at a time and selecting the model with the first variable that reduces the AIC the most. We then added the second variable and selected the model with the combination of two variables that reduced the AIC the most. This was repeated until a model with the combination of variables that reduced the AIC the most was achieved. The backward selection involved starting with an initial model with all the variables and dropping the variables one at a time until the model with the least AIC was achieved. The method then combined the forward and backward selection to select the best model.

We noticed that the stepwise regression still did not improve the accuracy of the model based on the results from the R-squared value. We then proceeded to test for multicollinearity among the variables in the new model. We noticed that all the VIF values of the variables were below the threshold of 5, as shown in Table 3 below, which meant that there was no multicollinearity in the model.

Table 3 - VIFs of variables in the model

Age	2.298757
Weight	3.351027
Height	2.550361
Hypertension	1.892192
Diabetes	1.321881
Thyroid	1.435767
Asthma	2.310273
Post-traumatic stress disorder	3.422400
Kidney	1.133192
Allergy	4.688636
Heart disease	1.178331
Irritable bowel syndrome	1.012261

We then proceeded to check for interactions in our model. This was done by stepwise update of the model with interaction terms between all the variables. That is, the model was updated with the interaction between age and weight (Age: Weight) and the AIC value was recorded, the

model was again updated with the interaction between age and height (Age: Height) and the AIC was again recorded, and so on until all the combinations were exhausted. Interactions were tested on the first order and only the interaction terms that reduced the values of the AIC were selected. The interaction term selected was between age and diabetic variables which reduced the AIC from 3,739 to 3,693. Its regression coefficient was -0.1032 which means that this interaction reduced the amount of water consumption.

On this new model, we finally proceed to check for outliers in the model. We created a function that identified the outliers based on the standardized residuals, the leverage values, cook's distance and diffts. Standardized residuals are residuals that have been transformed so they approximately follow a standard normal distribution. If they are above 3 in terms of absolute value, then investigate. Leverage is a measure of how far away the explanatory variable values are from those of the other observations. The leverage values are like residuals but in the horizontal direction. The rule of thumb is that a leverage value greater than $2\frac{p}{n}$ is a potential problem. p is the number of predictors+1 (i.e. the number of estimated parameters) and n is the sample size. The cook's distance finds influential outliers in the set of the explanatory variables. Values above 1 were considered to be high and are investigated. Finally, the diffts - which is an alternative to cook's distance - was also used to identify the outliers, and if it is above $2\sqrt{\frac{p}{n}}$ for large data sets we investigate. The function printed the values which were identified as outliers by at least two among the four criterions considered. These values were then removed from the training dataset and a new model was created from the dataset less the outliers. We then checked the accuracy of this model using the adjusted R-squared. The model seemed to improve the accuracy from 28.8% to 29.0% and therefore we decided to drop the outliers. We have now attained our final model.

4.4 Results of the Fitted Generalized Linear Model

Below gives a summary of the fitted model, that is, the variables, parameter estimates, standard error, the p-value of the fitted model. “Estimate” and “Std.Error” are the regression coefficients and the standard deviation of the explanatory variables respectively.

Table 4 - Summary of the model

Coefficient					
	Estimate	Std.Error	t-value	Pr(>|t|)	
Intercept	1.186451	0.196622	6.034	2.12e-09	***
Age	0.023570	0.002883	8.175	7.39e-16	***
Weight	0.005901	0.001363	4.328	1.62e-05	***
Height	0.010715	0.001834	5.843	6.57e-09	***
Hypertensionyes	-0.546107	0.098945	-5.519	4.16e-08	***
Diabeticyes	4.851000	0.992255	4.889	1.15e-06	***
Kidneyyes	0.688538	0.148190	4.646	3.75e-06	***
Thyroidyes	-0.381699	0.196164	-1.946	0.05191	*
Allergyyes	1.355527	0.420955	3.220	0.00132	**
Asthmayes	-1.312589	0.324578	-4.044	5.59e-05	***
Heart_Diseaseyes	-1.425142	0.267442	-5.329	1.18e-07	***
Post_traumatic_Stress_Disorderyes	-2.181952	0.518150	-4.211	2.73e-05	***
Irritable_Bowl_Syndromeyes	-0.626549	0.278214	-2.252	0.02450	*
Age:Diabeticyes	0.087473	0.015498	5.644	2.06e-08	***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

The value of the coefficient for each independent variable gives the size of the effect that the variable has on the dependent variable. The sign of the coefficient (positive or negative), gives the direction of the effect. If positive, it means that the independent variable increases the value

of the response variable whereas a negative coefficient decreases the value of the dependent variable. All the continuous variables, that is Age, Height and Weight had positive regression coefficients which meant that a unit increase of these variables increased the amount of water consumption.

The interpretation of the regression coefficients of the categorical variables is slightly different. First, all the categorical variables in our dataset contained two levels and so they can be referred to as binary variables. When we include a binary variable in regression, the baseline or reference level is chosen which the other level is compared to. For all the variables in our model, yes is taken as the reference level. The regression coefficient are also interpreted as the effect of the I dependent variable when in the reference level. For example, in Diabetic variable where yes is the reference, it can be interpreted as a person with diabetes will consume water 4.8 times more than those without diabetes. For those with Asthma, they will consume water 1.3 times lesser than those without Asthma. The same can be explained for the rest of the variables.

The p-value is the probability of obtaining the observed results of a test and it is obtained by getting the probability that the z-value is greater than the level of significance under consideration. If the p-value for an independent variable is less than the significance level, there is enough evidence to reject the null hypothesis that the regression coefficient of this variable is insignificant in the model. Basing on the 10% significance level, all the independent variables were found to be significant except Thyroid and irritable bowl syndrome since they all had a p-value greater than 0.1 as shown in table 4. We however do not drop these variables since dropping them lowers the accuracy of the model as seen by the reduced adjusted R-squared. Most of the standard errors of the significant variables are less than 0.3 which shows that the absolute measure of the distance that the data points fall from the regression line is small.

4.5 Forecasting and Accuracy of the Model

With the final model obtained, we did forecasting of the variable consumption on the test data, the second partition of the data which the model was totally not trained on and is a new dataset to the model. The model was able to predict most of the values, which were between 3 and 5 accurately and even some of those outside the range. We predicted new values on consumption from the test and created a data-frame with one column containing the predicted values and the other column containing the actual or the observed values which were in the test dataset. The table below shows randomly selected rows from this data-frame, to show the ability and accuracy of our model in making the predictions or forecasts.

Table 5 - Randomly selected rows of predicted against forecasted values

#Observation	Predicted	Actual
1165	4.481043	5.00
1177	4.391447	4.75
915	3.908193	3.75
1096	4.974828	3.50
343	4.695628	6.75
627	5.120187	2.50
184	3.769953	6.00
158	4.463750	5.00
846	4.252046	6.00
227	4.191706	6.75

A plot of the observed or actual values against the predicted values was also created to help in understanding the accuracy of our model. If, for example, the actual value is 5 and the predicted value is 5, and on another occasion the actual value is 10 and the predicted value is 10, when plotted they will form a diagonal line on the plot. If the model that is used for prediction has a high R Square, then the predicted values will be close to the actual values and the plot of actual against the predicted will form a general diagonal line. The lesser the R Square of our model the more the plot does not appear to form a diagonal shape. In the case of our model, the plot was not able to assume a clear diagonal line due to the low R Square whose reason was discussed earlier.

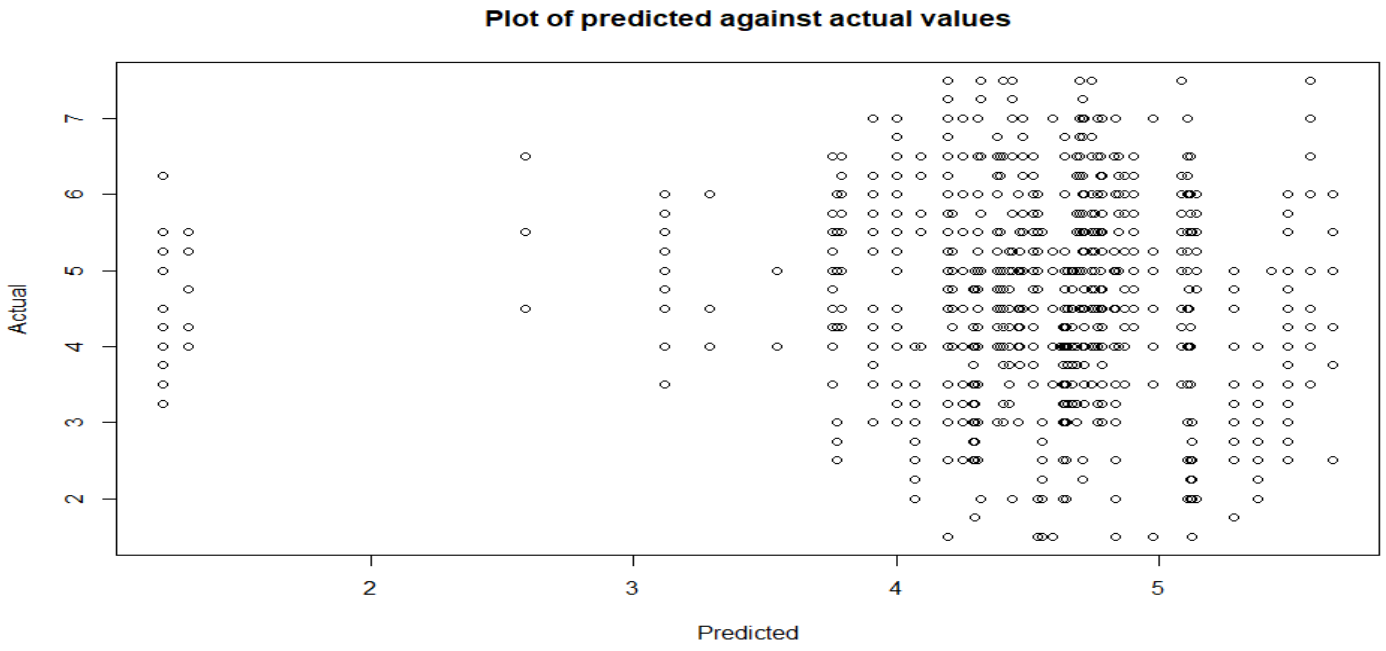

Figure 8 - Plot of actual values against the predicted values

CHAPTER 5: CONCLUSION AND RECOMMENDATION

5.1 Introduction

This chapter covers the summary of all the findings from the research, conclusions drawn from the findings and suggested recommendations for further research in modeling the amounts of water consumption. Section 5.2 gives the conclusions drawn from the project and section 5.3 has the recommendations given for further research.

5.2 Conclusion

The best distribution which the response variable fitted was normal distribution, and any further attempt to transform it to a normal distribution so as to make the kurtosis closer to 3 made it deviate further from being normal distribution. In some way, the data appeared to be discrete and the number of values of the amount of water consumption was 25. That is, when we converted the consumption variable to factors, the number of levels were 25. In the data mining step to determine the best model for our analysis, timeseries models could not be considered since the data is not a timeseries data. It does not have characteristics of timeseries data such as trend and seasonality and though it was collected over a period of 30 days, the number of individuals recorded were 79 and so we would have to create 79 timeseries models if we were to transform the data into a timeseries and use timeseries models.

We therefore considered using regression models and the generalized linear regression model was considered the best, since the response variable was normally distributed and some of the explanatory variables were factors. As it was observed that the response variable appeared discrete with 25 levels, we tried fitting regression models for count data (Poisson regression and negative binomial regression) but they could not fit since the variance of the data is higher than the mean. We then tried to fit a multinomial logistic regression model, but its initial accuracy

was much lower (24%) compared to that of the generalized linear model (29%). The accuracy of our final model was at 29% based on the adjusted R squared. This is so because the model accurately predicts the values between 3 and 5 since the training dataset contained observations with a lot of values lying between 3 and 5. Our model was therefore biased to forecast that new observations would lie in this range.

5.3 Recommendations for Further Research

Firstly, the model was not good enough at forecasting the values not within the range of 4-5. For further research and study, we recommend using more data where the total number of observations are evenly distributed along the entire range. This will also make the data not appear as discrete and the accuracy of the model will improve. We also recommend using other machine learning algorithms and models for forecasting, such as clustering models like the K-means clustering algorithm along with simulation models to ensure the forecasting is contribution supply chain resilience. This is because the dataset had a characteristic of forming clusters around the 1-2 range and 3-5 range.

CHAPTER 6: REFERENCES

Adam P., Jakub J. & Bartosz K. (2018). Forecasting Daily Water Consumption: A Case Study in Torun, Poland

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989) *Statistical Modelling in GLIM*. Oxford Science Publications, Oxford

Aleksić, A., Stefanović, M., Arsovski, S., & Tadić, D. (2013). An assessment of organizational resilience potential in SMEs of the process industry, a fuzzy approach. *Journal of Loss Prevention in the Process Industries*, *26*(6), 1238-1245. doi:10.1016/j.jlp.2013.06.004

Ambulkar, S., Blackhurst, J., & Grawe, S. (2014). Firm's resilience to supply chain disruptions: Scale development and empirical examination. *Journal of Operations Management*, *33-34*(1), 111-122. doi:10.1016/j.jom.2014.11.002

Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E. and Thandi, N. (2005) *A Practitioner's Guide to Generalized Linear Models*. Casualty Actuarial Society, Virginia

Annarelli, A., & Nonino, F. (2016). Strategic and operational management of organizational resilience: Current state of research and future directions. *Omega*, *62*, 1-18. doi:10.1016/j.omega.2015.08.004

Antonio C., David S. & Francesco A. (2015). Short-term forecasting of hourly water consumption by using automatic metering readers' data

Barrios, K. (2018, October 10). Top 10 global supply chain risks. Retrieved from https://www.xeneta.com/blog/global-supply-chain-risks

Benard, B. (2004). *Resiliency: What we have learned*. WestEd.

Bouaziz, F., & Smaoui Hachicha, Z. (2018). Strategic human resource management practices and organizational resilience. *Journal of Management Development*, *37*(7), 537-551. doi:10.1108/jmd-11-2017-0358

Brun, A., & Caridi, M. (2008). Assessing improvement opportunities and risks of supply chain transformation projects. *Supply Chain*. doi:10.5772/5355

Chewning, L. V., Lai, C., & Doerfel, M. L. (2012). Organizational resilience and using information and communication technologies to rebuild communication structures. *Management Communication Quarterly*, *27*(2), 237-263. doi:10.1177/0893318912465815

Civil daily. (2020). Coronavirus – Health and governance issues. Retrieved from https://www.civilsdaily.com/story/coronavirus-health-and-governance-issues/

Crichton, M. T., Ramsay, C. G., & Kelly, T. (2009). Enhancing organizational resilience through emergency planning: Learnings from cross-sectoral lessons. *Journal of Contingencies and Crisis Management*, *17*(1), 24-37. doi:10.1111/j.1468-5973.2009.00556.x

David W., Enrico C., Lydia V., Raziyeh F., Zoran K. & Dragan S. (2015). Forecasting Domestic Water Consumption from Smart Meter Readings using Statistical Methods and Artificial Neural Networks

Deloitte. (2018, February 14). Supply chain resilience: Five areas for managing risk | Deloitte US. Retrieved from https://www2.deloitte.com/us/en/pages/risk/articles/improving-supply-chain-resilience.html

Ellis, S. (2020). Coronavirus: Impact on, and Implications for, the Global Supply Chain. Retrieved from https://www.ibm.com/downloads/cas/WMAJZAGB

Eltantawy, R. (2011). Supply management governance role in supply chain risk management and sustainability. *Supply Chain Management - New Perspectives*. doi:10.5772/19992

Gittell, J. H., Cameron, K., Lim, S., & Rivas, V. (2006). Relationships, layoffs, and organizational resilience. *The Journal of Applied Behavioral Science*, *42*(3), 300-329. doi:10.1177/0021886306286466

Gölgeci, I., & Kuivalainen, O. (2020). Does social capital matter for supply chain resilience? The role of absorptive capacity and marketing-supply chain management alignment. *Industrial Marketing Management*, *84*, 63-74. doi:10.1016/j.indmarman.2019.05.006

GT Nexus. (2020). The Four Level of Supply Chain Maturity. Retrieved from https://www.gtnexus.com/blog/cloud-supply-chain/the-four-levels-of-supply-chain-maturity/

Hecht, A. A., Biehl, E., Barnett, D. J., & Neff, R. A. (2019). Urban food supply chain resilience for crises threatening food security: A qualitative study. *Journal of the Academy of Nutrition and Dietetics*, *119*(2), 211-224. doi:10.1016/j.jand.2018.09.001

Hess, J. (2010). Globales supply chain management braucht supply chain engineering. *Supply Chain Engineering*, 17-22. doi:10.1007/978-3-8349-6357-4_3

Hillmann, J., Duchek, S., Meyr, J., & Guenther, E. (2018). Educating future managers for developing resilient organizations: The role of scenario planning. *Journal of Management Education*, *42*(4), 461-495. doi:10.1177/1052562918766350

History.com Editors. (2009, October 29). Great Depression history. Retrieved from https://www.history.com/topics/great-depression/great-depression-history

Ignatiadis, I., & Nandhakumar, J. (2007). The impact of enterprise systems on organizational resilience. *The Transfer and Diffusion of Information Technology for Organizational Resilience*, *22*(1), 259-274. doi:10.1007/0-387-34410-1_18

Julian J. Faraway, (2016), Extending the Linear Model with R: Generalized Linear, *Mixed Effects and Nonparametric Regression Models (2nd ed). Taylor & Francis Group LLC*.

KANTUR, D., & İŞERİ SAY, A. (2012). Organizational resilience: A conceptual integrative framework. *Journal of Management & Organization*, 2155-2181. doi:10.5172/jmo.2012.2155

Kasemsap, K. (2020). Advocating sustainable supply chain management and sustainability in global supply chain. *Supply Chain and Logistics Management*, 1462-1490. doi:10.4018/978-1-7998-0945-6.ch071

Lam, H. K. (2018). Doing good across organizational boundaries. *International Journal of Operations & Production Management*, *38*(12), 2389-2412. doi:10.1108/ijopm-02-2018-0056

Lengnick-Hall, C. A., Beck, T. E., & Lengnick-Hall, M. L. (2011). Developing a capacity for organizational resilience through strategic human resource management. *Human Resource Management Review*, *21*(3), 243-255. doi:10.1016/j.hrmr.2010.07.001

Linnenluecke, M. K., Griffiths, A., & Winn, M. (2011). Extreme weather events and the critical importance of anticipatory adaptation and organizational resilience in

responding to impacts. *Business Strategy and the Environment*, *21*(1), 17-32. doi:10.1002/bse.708

Mallak, L. (2002). Toward a theory of organizational resilience. *PICMET '99: Portland International Conference on Management of Engineering and Technology. Proceedings Vol-1: Book of Summaries (IEEE Cat. No.99CH36310)*. doi:10.1109/picmet.1999.808142

Mamouni Limnios, E. A., Mazzarol, T., Ghadouani, A., & Schilizzi, S. G. (2014). The resilience architecture framework: Four organizational archetypes. *European Management Journal*, *32*(1), 104-116. doi:10.1016/j.emj.2012.11.007

McManus, S., Seville, E., Vargo, J., & Brunsdon, D. (2008). Facilitated process for improving organizational resilience. *Natural Hazards Review*, *9*(2), 81-90. doi:10.1061/(asce)1527-6988(2008)9:2(81)

Melnyk, S. A., Closs, D. J., Griffis, S. E., Zobel, C. W., & Macdonald, J. R. (2015, November 20). Understanding supply chain resilience - Supply chain 24/7. Retrieved from https://www.supplychain247.com/article/understanding_supply_chain_resilience/

Mentzer, J. T. (2001). *Supply chain management*. SAGE.

Nelder, John Ashworth and Robert WM Wedderburn (1972). "Generalized linear models". In: *Journal of the Royal Statistical Society: Series A (General)* 135.3, pp. 370–384

Norris, P., & Inglehart, R. F. (2018). Understanding Brexit: Cultural resentment versus economic grievances. *SSRN Electronic Journal*. doi:10.2139/ssrn.3231896

Orchiston, C., Prayag, G., & Brown, C. (2016). Organizational resilience in the tourism sector. *Annals of Tourism Research*, *56*, 145-148. doi:10.1016/j.annals.2015.11.002

Ortiz-de-Mandojana, N., & Bansal, P. (2015). The long-term benefits of organizational resilience through sustainable business practices. *Strategic Management Journal*, *37*(8), 1615-1631. doi:10.1002/smj.2410

Pires Ribeiro, J., & Barbosa-Povoa, A. (2018). Supply chain resilience: Definitions and quantitative modelling approaches – A literature review. *Computers & Industrial Engineering*, *115*, 109-122. doi:10.1016/j.cie.2017.11.006

Powley, E. H. (2009). Reclaiming resilience and safety: Resilience activation in the critical period of crisis. *Human Relations*, *62*(9), 1289-1326. doi:10.1177/0018726709334881

Riolli, L., & Savicki, V. (2003). Information system organizational resilience. *Omega*, *31*(3), 227-233. doi:10.1016/s0305-0483(03)00023-9

Rubinstein, P. (2020). Why grocery shelves won't be empty for long. Retrieved from https://www.bbc.com/worklife/article/20200401-covid-19-why-we-wont-run-out-of-food-during-coronavirus

Sahebjamnia, N., Torabi, S., & Mansouri, S. (2015). Integrated business continuity and disaster recovery planning: Towards organizational resilience. *European Journal of Operational Research*, *242*(1), 261-273. doi:10.1016/j.ejor.2014.09.055

Sakate, DM and DN Kashid (2016). “A new robust model selection method in GLM with application to ecological data”. In: *Environmental Systems Research* 5.1, p. 9. DOI: p0-ef839/ue5.2016.03. 009. URL: http://xehw.org/abs/2016.3019

Saphir, A. (2020, May 8). Explainer: Why 14.7% unemployment rate doesn't capture the true state of the coronavirus economy. Retrieved from https://www.reuters.com/article/us-usa-economy-unemployment-data-explain/explainer-why-fridays-u-s-jobless-figures-wont-capture-the-true-state-of-the-coronavirus-economy-idUSKBN22K0HW

Sawalha, I. H. (2015). Managing adversity: Understanding some dimensions of organizational resilience. *Management Research Review*, *38*(4), 346-366. doi:10.1108/mrr-01-2014-0010

Somers, S. (2009). Measuring resilience potential: An adaptive strategy for organizational crisis planning. *Journal of Contingencies and Crisis Management*, *17*(1), 12-23. doi:10.1111/j.1468-5973.2009.00558.x

Spiegler, V., Potter, A., Naim, M., & Towill, D. (2015). The value of nonlinear control theory in investigating the underlying dynamics and resilience of a grocery supply chain. *International Journal of Production Research*, *54*(1), 265-286. doi:10.1080/00207543.2015.1076945

Supply chain dive. (2017, August 7). The 5 types of supply chain risk. Retrieved from https://www.supplychaindive.com/news/5-supply-chain-risk-spotlight/448580/

Tadić, D., Aleksić, A., Stefanović, M., & Arsovski, S. (2014). Evaluation and ranking of organizational resilience factors by using a two-step fuzzy AHP and fuzzy TOPSIS. *Mathematical Problems in Engineering*, *2014*, 1-13. doi:10.1155/2014/418085

The Economic Times. (2020, April 3). How the coronavirus crisis is affecting food supply. Retrieved from https://economictimes.indiatimes.com/news/international/world-news/how-the-coronavirus-crisis-is-affecting-food-supply/are-we-facing-food-shortages/slideshow/74960847.cms

The Economist. (2020, April 11). The changes COVID-19 is forcing on to business. Retrieved from https://www.economist.com/briefing/2020/04/11/the-changes-covid-19-is-forcing-on-to-business

The New York Times. (2020, February 21). Coronavirus outbreak deepens its toll on global business. Retrieved from https://www.nytimes.com/2020/02/21/business/coronavirus-global-business.html

Vogus, T. J., & Sutcliffe, K. M. (2007). Organizational resilience: Towards a theory and research agenda. *2007 IEEE International Conference on Systems, Man and Cybernetics*. doi:10.1109/icsmc.2007.4414160.

WHO. (2020). Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). Retrieved from https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf

Wicker, P., Filo, K., & Cuskelly, G. (2013). undefined. *Journal of Sport Management*, *27*(6), 510-525. doi:10.1123/jsm.27.6.510

Wieland, A., Handfield, R. B., & Durach, C. F. (2016). undefined. *Journal of Business Logistics*, *37*(3), 205-212. doi:10.1111/jbl.12131

Willke, H., & Willke, G. (2012). *Political governance of capitalism: A reassessment beyond the global crisis*. Edward Elgar Publishing.

Wisner, J. D., Tan, K., & Leong, G. K. (2014). *Principles of supply chain management: A balanced approach*. Cengage Learning.

Yang, G. (2014, March 30). Building a resilient organization-why and how | OBlog. Retrieved from https://blog.nus.edu.sg/audreyc/2014/03/30/building-a-resilient-organization-why-and-how/

Garver, R. (2020, June 2). Economic damage from civil unrest may persist for decades. Retrieved from https://www.voanews.com/usa/nation-turmoil-george-floyd-protests/economic-damage-civil-unrest-may-persist-decades

Lu, S. (2020, March 26). How might COVID-19 affect apparel sourcing and trade. Retrieved from https://shenglufashion.com/2020/03/25/how-might-covid-19-affect-apparel-sourcing-and-trade/

Forecasting references

Bizfluent. (2018). The disadvantages of long-term cash flow forecasting. https://bizfluent.com/info-8174633-disadvantages-longterm-cash-flow-forecasting.html

Bizfluent. (2020). The differences between qualitative and quantitative forecasting techniques. https://bizfluent.com/info-12042327-differences-between-qualitative-quantitative-forecasting-techniques.html

Otexts. (2020). 5.2 least squares estimation | Forecasting: Principles and practice. OTexts. https://otexts.com/fpp2/least-squares.html

Slideshare. (2017, April 13). Quantitative and qualitative forecasting techniques om. Share and Discover Knowledge on SlideShare. https://www.slideshare.net/HallmarkBschool/quantitative-and-qualitative-forecasting-techniques-om

Slideshare. (2018, May 30). Forecasting methods. Share and Discover Knowledge on SlideShare. https://www.slideshare.net/vaibhavagarwal75436/forecasting-methods-99591465

GLM REFERENCES

Adam P., Jakub J. & Bartosz K. (2018). Forecasting Daily Water Consumption: A Case Study in Torun, Poland

Antonio C., David S. & Francesco A. (2015). Short-term forecasting of hourly water consumption by using automatic metering readers' data

David W., Enrico C., Lydia V., Raziyeh F., Zoran K. & Dragan S. (2015). Forecasting Domestic Water Consumption from Smart Meter Readings using Statistical Methods and Artificial Neural Networks

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989) *Statistical Modelling in GLIM*. Oxford Science Publications, Oxford

Nelder, John Ashworth and Robert WM Wedderburn (1972). "Generalized linear models". In: *Journal of the Royal Statistical Society: Series A (General)* 135.3, pp. 370–384

Sakate, DM and DN Kashid (2016). "A new robust model selection method in GLM with application to ecological data". In: *Environmental Systems Research* 5.1, p. 9. DOI: p0-ef839/ue5.2016.03. 009. URL: http://xehw.org/abs/2016.3019

Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E. and Thandi, N. (2005) *A Practitioner's Guide to Generalized Linear Models*. Casualty Actuarial Society, Virginia

GLM APPENDIX

R CODE:

```
rm(list=ls())

cat("\f")

 #loading required R packages

library(readxl)

library(rcompanion)

library(fBasics)

library(ggplot2)

library(Boruta)

library(gridExtra)

library(car)

library(rsq)



dat = read_xlsx("F:/GeniusBen/Fiverr/Richard/finaldata.xlsx")

View(dat)

nrow(dat)   #no. of observations made



#Changing the type of variables into their respective types

str(dat)  #Viewing the structure of the data

sapply(dat, class) #viewing the class of the variables

col_names <- names(dat)

dat[,col_names] <- lapply(dat[,col_names], factor)

sapply(dat, class)
```

```
#change back to numeric from factors

dat$Consumption  <-  as.numeric(gsub("[\\%,]", "", dat$Consumption))

dat$Age  <-  as.numeric(gsub("[\\%,]", "", dat$Age))

dat$Weight  <-  as.numeric(gsub("[\\%,]", "", dat$Weight))

dat$Height  <-  as.numeric(gsub("[\\%,]", "", dat$Height))



#confirm the new structure and class of the dataset if it is correct

str(dat)

sapply(dat, class)

#check for missing values

sum(is.na(dat))

#data visualization with gg plots (boxplots and scatter plots)

p1 <- ggplot(data = dat,aes(x = Age, y = Consumption, group = 1)) + geom_boxplot() +

  coord_flip() ;p1

p1 <- p1 + geom_point() ;p1

p2 <- ggplot(data = dat,aes(x = Weight, y = Consumption)) + geom_boxplot() +

  coord_flip() ;p2

p2 <- p2 + geom_point() ;p2

p3 <- ggplot(data = dat,aes(x = Height, y = Consumption)) + geom_boxplot() +

 coord_flip() ;p3

p3 <- p3 + geom_point() ;p3

p4 <- ggplot(data = dat,aes(x=Diabetes, y=Consumption,group=1))+geom_boxplot()+
```

```
  coord_flip() ;p4

p5 <- ggplot(data = dat,aes(x=Kidney, y=Consumption,group=1))+geom_boxplot()+

  coord_flip() ;p5

p6 <- ggplot(data = dat,aes(x=Thyroid, y=Consumption,group=1))+geom_boxplot()+

  coord_flip() ;p6

p7 <- ggplot(data = dat,aes(x=Allergy, y=Consumption,group=1))+geom_boxplot()+

  coord_flip() ;p7

p8 <- ggplot(data = dat,aes(x=Asthma, y=Consumption,group=1))+geom_boxplot()+

  coord_flip() ;p8

p9 <- ggplot(data = dat,aes(x=Heart_Disease, y=Consumption,group=1))+geom_boxplot()+

  coord_flip() ;p9

p10 <- ggplot(data = dat,aes(x=Post_traumatic_Stress_Disorder, y=Consumption,group=1))+geom_boxplot()+

  coord_flip() ;p10

p11 <- ggplot(data = dat,aes(x=Irritable_Bowel_Syndrome, y=Consumption,group=1))+geom_boxplot()+

  coord_flip() ;p11

grid.arrange(p1,p2,p3,p4)

grid.arrange(p5,p6,p7,p8)

grid.arrange(p9,p10,p11)



#check for outliers

outlierKD <- function(dt, var) {

  var_name <- eval(substitute(var),eval(dt))

  tot <- sum(!is.na(var_name))
```

```
na1 <- sum(is.na(var_name))

m1 <- mean(var_name, na.rm = T)

par(mfrow=c(2, 2), oma=c(0,0,3,0))

boxplot(var_name, main="With outliers")

hist(var_name, main="With outliers", xlab=NA, ylab=NA)

outlier <- boxplot.stats(var_name)$out

mo <- mean(outlier)

var_name <- ifelse(var_name %in% outlier, NA, var_name)

boxplot(var_name, main="Without outliers")

hist(var_name, main="Without outliers", xlab=NA, ylab=NA)

title("Outlier Check", outer=TRUE)

na2 <- sum(is.na(var_name))

message("Outliers identified: ", na2 - na1, " from ", tot, " observations")

message("Proportion (%) of outliers: ", (na2 - na1) / tot*100)

message("Mean of the outliers: ", mo)

m2 <- mean(var_name, na.rm = T)

message("Mean without removing outliers: ", m1)

message("Mean if we remove outliers: ", m2)

response <- readline(prompt="Do you want to remove outliers and to replace with NA? [yes/no]: ")

if(response == "y" | response == "yes"){

 dt[as.character(substitute(var))] <- invisible(var_name)

 assign(as.character(as.list(match.call())$dt), dt, envir = .GlobalEnv)

 message("Outliers successfully removed", "\n")
```

```
    return(invisible(dt))
  } else{
    message("Nothing changed", "\n")
    return(invisible(var_name))
  } }
dat2 = dat
outlierKD(dat2, dat2$Consumption)
sum(is.na(dat2))
dat2=na.omit(dat2)
str(dat2)
dat3 = dat2[,-c(14,15)]
View(dat3)


#Variable Selection using Boruta package
set.seed(111)
boruta <- Boruta(Consumption~., data = dat3, doTrace=2, maxRuns=500)
print(boruta)
par(mfrow=c(1,1))
plot(boruta, las=2, cex.axis=0.7)
bor<-TentativeRoughFix(boruta)
print(bor)
attStats(boruta)
```

```
## Descriptive statistics of the response variable

Consumption = dat3$Consumption

summary(Consumption)



#Check for normality

basicStats(Consumption)

par(mfrow = c(1,2))

plotNormalHistogram(Consumption,main = "Histogram and Normal

  curve",

                xlab="Consumption in mls")

qqnorm(Consumption, ylab = "Water Consumption")

qqline(Consumption, col="red")



#We try transforming the response variable using box cox transformation, as suggested by bestNormalize function which iteratively

suggests the best method for transformation but it deviates the data farther from a normal distribution.

library(bestNormalize)

bestNormalize(Consumption)

length(Consumption)

trans<-numeric(2150)

x<-dat2$Consumption

for (i in 1:2150){

 c<--0.00001

 trans[i]<- (x[i]^c-1)/c
```

```
}

plotNormalHistogram(trans)  #transformed data is far away from a normal distribution as seen in these three plots.

basicStats(trans)

qqnorm(trans)



#normally distributed should have a plot similar to

plotNormalHistogram(rnorm(1000,10,5), main="Histogram and normal curve of

normally distributed data",  xlab = "simulations")



#Data Partition

set.seed(222)

ind = sample(2, nrow(dat3), replace = TRUE, prob = c(0.6,0.4))

training = dat3[ind == 1,]

test = dat3[ind == 2,]



# Model creation

# Initial model with all variables included

model0 <- glm(formula =  Consumption ~ Age + Weight + Height + Hypertension +

              Diabetes + Kidney + Thyroid + Allergy + Asthma + Heart_disease +

              Post-traumatic_stress_disorder + Irritable_bowel_syndrome,

            data = training, family = gaussian())

summary(model0)

model1 = step(model0, direction = "both")
```

```
summary(model1)

rsq(model1,adj = TRUE)



#Test for multicollinearity

vif(model1)



### Test of interactions

model2 <- update(model1, ~.+ Age:Weight)

summary(model2)

model3 <- update(model1, ~.+ Age:Height)

summary(model3)

model4 <- update(model1, ~.+ Age:Hypertension)

summary(model4)

model5 <- update(model1, ~.+ Age:Diabetes)

summary(model5)

model6 <- update(model1, ~.+ Age:Kidney)

summary(model6)

model7 <- update(model1, ~.+ Age:Thyroid)

summary(model7)

#new best model

model5 <- update(model1, ~.+ Age:Diabetes)

summary(model5)
```

```
rsq(model5, adj = TRUE)



# This function displays the outlier statistics. Note that there is no limit

# to the number of these so it could blow up if there are a lot of them. You

# can modify the function to deal with this if you like.

show_outliers <- function(the.linear.model, topN) {

 # length of data

 n = length(fitted(the.linear.model))

 # number of parameters estimated

 p = length(coef(the.linear.model))

 # standardised residuals over 3

 res.out <- which(abs(rstandard(the.linear.model)) > 3) #sometimes >2

 # topN values

 res.top <- head(rev(sort(abs(rstandard(the.linear.model)))), topN)

 # high leverage values

 lev.out <- which(lm.influence(the.linear.model)$hat > 2 * p/n)

 # topN values

 lev.top <- head(rev(sort(lm.influence(the.linear.model)$hat)), topN)

 # high diffits

 dffits.out <- which(dffits(the.linear.model) > 2 * sqrt(p/n))

 # topN values

 dffits.top <- head(rev(sort(dffits(the.linear.model))), topN)

 # Cook's over 1
```

```
  cooks.out <- which(cooks.distance(the.linear.model) > 1)

  # topN cooks

  cooks.top <- head(rev(sort(cooks.distance(the.linear.model))), topN)

  # Create a list with the statistics – can't do a data frame as different

  # lengths

  list.of.stats <- list(Std.res = res.out, Std.res.top = res.top, Leverage = lev.out,

                        Leverage.top = lev.top, DFFITS = dffits.out, DFFITS.top = dffits.top,

                        Cooks = cooks.out, Cooks.top = cooks.top)

  # return the statistics

  list.of.stats

}

Outliers <- show_outliers(model5, 10)  ;Outliers



#In practice you often get points that violate more than one criterion. we look for these points.

#We use the function intersect () which gives us the intersection of its two arguments.

group_of_outliers <- Reduce(intersect,list(Outliers$DFFITS,Outliers$Leverage))



#We extract the outliers from the dataframe and compare the summary statistics

#to that of the original data

Outlier_Values = training[group_of_outliers,]



summary(training)

summary(Outlier_Values)
```

```
#data less outliers

Data_less_outliers = training[-group_of_outliers,]



#new model less outliers

model8 <- glm(formula = Consumption ~ Age + Weight + Height +  Hypertension + Diabetes + Kidney + Thyroid + Allergy + Asthma +

Heart_disease + Post_traumatic_stress_disorder +  Irritable_bowel_syndrome +  Age:Diabetes, family = gaussian(), data = Data_less_outliers)

summary(model8)

rsq(model8, adj = "TRUE")




#diagnostic plots for the model

par(mfrow=c(2,2))

plot(model8)



#compare predicted vs actual on the test dataset

pred = predict(model8, data = test)

forecastVSactual =as.data.frame(cbind(pred, test$Consumption))

names(forecastVSactual)[1] <- "Predicted"

names(forecastVSactual)[2] <- "Actual"

head(forecastVSactual)



library(dplyr)
```

```
sample_n(forecastVSactual, 10)

#################################################################

#Let's try a multinomial logistic model

dat4 = dat3

dat4$Consumption = as.factor(dat4$Consumption)

summary(dat4$Consumption)

str(dat4$Consumption)



library(nnet)

model2 = multinom(Consumption~., data =dat4)

summary(model2)



# Predicting the values for train dataset

training$precticed <- predict(model2, newdata = training, "class")



# Building classification table

ctable <- table(training$Consumption, training$precticed)



# Calculating accuracy - sum of diagonal elements divided by total obs

round((sum(diag(ctable))/sum(ctable))*100,4)

#from the %age of accuracy, multinomial logistic model is not the best,

#we therefore choose the Generalized linear model.

#############################################################################
```

APPENDIX

R CODE:

```
rm(list=ls())

cat("\f")

 #loading required R packages

library(readxl)

library(rcompanion)

library(fBasics)

library(ggplot2)

library(Boruta)

library(gridExtra)

library(car)

library(rsq)



dat = read_xlsx("F:/GeniusBen/Fiverr/Richard/finaldata.xlsx")

View(dat)

nrow(dat)   #no. of observations made



#Changing the type of variables into their respective types

str(dat)  #Viewing the structure of the data

sapply(dat, class) #viewing the class of the variables

col_names <- names(dat)

dat[,col_names] <- lapply(dat[,col_names], factor)

sapply(dat, class)
```

```
#change back to numeric from factors

dat$Consumption  <-  as.numeric(gsub("[\\%,]", "", dat$Consumption))

dat$Age  <-  as.numeric(gsub("[\\%,]", "", dat$Age))

dat$Weight  <-  as.numeric(gsub("[\\%,]", "", dat$Weight))

dat$Height  <-  as.numeric(gsub("[\\%,]", "", dat$Height))



#confirm the new structure and class of the dataset if it is correct

str(dat)

sapply(dat, class)

#check for missing values

sum(is.na(dat))

#data visualization with gg plots (boxplots and scatter plots)

p1 <- ggplot(data = dat,aes(x = Age, y = Consumption, group = 1)) + geom_boxplot() +

  coord_flip() ;p1

p1 <- p1 + geom_point() ;p1

p2 <- ggplot(data = dat,aes(x = Weight, y = Consumption)) + geom_boxplot() +

  coord_flip() ;p2

p2 <- p2 + geom_point() ;p2

p3 <- ggplot(data = dat,aes(x = Height, y = Consumption)) + geom_boxplot() +

 coord_flip() ;p3

p3 <- p3 + geom_point() ;p3

p4 <- ggplot(data = dat,aes(x=Diabetes, y=Consumption,group=1))+geom_boxplot()+
```

```
  coord_flip() ;p4

p5 <- ggplot(data = dat,aes(x=Kidney, y=Consumption,group=1))+geom_boxplot()+

  coord_flip() ;p5

p6 <- ggplot(data = dat,aes(x=Thyroid, y=Consumption,group=1))+geom_boxplot()+

  coord_flip() ;p6

p7 <- ggplot(data = dat,aes(x=Allergy, y=Consumption,group=1))+geom_boxplot()+

  coord_flip() ;p7

p8 <- ggplot(data = dat,aes(x=Asthma, y=Consumption,group=1))+geom_boxplot()+

  coord_flip() ;p8

p9 <- ggplot(data = dat,aes(x=Heart_Disease, y=Consumption,group=1))+geom_boxplot()+

  coord_flip() ;p9

p10 <- ggplot(data = dat,aes(x=Post_traumatic_Stress_Disorder, y=Consumption,group=1))+geom_boxplot()+

  coord_flip() ;p10

p11 <- ggplot(data = dat,aes(x=Irritable_Bowel_Syndrome, y=Consumption,group=1))+geom_boxplot()+

  coord_flip() ;p11

grid.arrange(p1,p2,p3,p4)

grid.arrange(p5,p6,p7,p8)

grid.arrange(p9,p10,p11)



#check for outliers

outlierKD <- function(dt, var) {

  var_name <- eval(substitute(var),eval(dt))

  tot <- sum(!is.na(var_name))
```

```
na1 <- sum(is.na(var_name))

m1 <- mean(var_name, na.rm = T)

par(mfrow=c(2, 2), oma=c(0,0,3,0))

boxplot(var_name, main="With outliers")

hist(var_name, main="With outliers", xlab=NA, ylab=NA)

outlier <- boxplot.stats(var_name)$out

mo <- mean(outlier)

var_name <- ifelse(var_name %in% outlier, NA, var_name)

boxplot(var_name, main="Without outliers")

hist(var_name, main="Without outliers", xlab=NA, ylab=NA)

title("Outlier Check", outer=TRUE)

na2 <- sum(is.na(var_name))

message("Outliers identified: ", na2 - na1, " from ", tot, " observations")

message("Proportion (%) of outliers: ", (na2 - na1) / tot*100)

message("Mean of the outliers: ", mo)

m2 <- mean(var_name, na.rm = T)

message("Mean without removing outliers: ", m1)

message("Mean if we remove outliers: ", m2)

response <- readline(prompt="Do you want to remove outliers and to replace with NA? [yes/no]: ")

if(response == "y" | response == "yes"){

  dt[as.character(substitute(var))] <- invisible(var_name)

  assign(as.character(as.list(match.call())$dt), dt, envir = .GlobalEnv)

  message("Outliers successfully removed", "\n")
```

```
  return(invisible(dt))
 } else{
  message("Nothing changed", "\n")
  return(invisible(var_name))
 } }
dat2 = dat
outlierKD(dat2, dat2$Consumption)
sum(is.na(dat2))
dat2=na.omit(dat2)
str(dat2)
dat3 = dat2[,-c(14,15)]
View(dat3)


#Variable Selection using Boruta package
set.seed(111)
boruta <- Boruta(Consumption~., data = dat3, doTrace=2, maxRuns=500)
print(boruta)
par(mfrow=c(1,1))
plot(boruta, las=2, cex.axis=0.7)
bor<-TentativeRoughFix(boruta)
print(bor)
attStats(boruta)
```

```
## Descriptive statistics of the response variable

Consumption = dat3$Consumption

summary(Consumption)



#Check for normality

basicStats(Consumption)

par(mfrow = c(1,2))

plotNormalHistogram(Consumption,main = "Histogram and Normal

  curve",

                xlab="Consumption in mls")

qqnorm(Consumption, ylab = "Water Consumption")

qqline(Consumption, col="red")



#normally distributed should have a plot similar to

plotNormalHistogram(rnorm(1000,10,5), main="Histogram and normal curve of

normally distributed data",  xlab = "simulations")



#Data Partition

set.seed(222)

ind = sample(2, nrow(dat3), replace = TRUE, prob = c(0.6,0.4))

training = dat3[ind == 1,]

test = dat3[ind == 2,]
```

```
# Model creation

# Initial model with all variables included

model0 <- glm(formula =  Consumption ~ Age + Weight + Height + Hypertension +

              Diabetes + Kidney + Thyroid + Allergy + Asthma + Heart_disease +

              Post_traumatic_stress_disorder + Irritable_bowel_syndrome,

            data = training, family = gaussian())

summary(model0)

model1 = step(model0, direction = "both")

summary(model1)

rsq(model1,adj = TRUE)




#Test for multicollinearity

vif(model1)



### Test of interactions

model2 <- update(model1, ~.+ Age:Weight)

summary(model2)

model3 <- update(model1, ~.+ Age:Height)

summary(model3)

model4 <- update(model1, ~.+ Age:Hypertension)

summary(model4)

model5 <- update(model1, ~.+ Age:Diabetes)
```

```
summary(model5)

model6 <- update(model1, ~.+ Age:Kidney)

summary(model6)

model7 <- update(model1, ~.+ Age:Thyroid)

summary(model7)

#new best model

model5 <- update(model1, ~.+ Age:Diabetes)

summary(model5)

rsq(model5, adj = TRUE)



# This function displays the outlier statistics. Note that there is no limit

# to the number of these so it could blow up if there are a lot of them. You

# can modify the function to deal with this if you like.

show_outliers <- function(the.linear.model, topN) {

 # length of data

 n = length(fitted(the.linear.model))

 # number of parameters estimated

 p = length(coef(the.linear.model))

 # standardised residuals over 3

 res.out <- which(abs(rstandard(the.linear.model)) > 3) #sometimes >2

 # topN values

 res.top <- head(rev(sort(abs(rstandard(the.linear.model)))), topN)

 # high leverage values
```

```
  lev.out <- which(lm.influence(the.linear.model)$hat > 2 * p/n)

  # topN values

  lev.top <- head(rev(sort(lm.influence(the.linear.model)$hat)), topN)

  # high diffits

  dffits.out <- which(dffits(the.linear.model) > 2 * sqrt(p/n))

  # topN values

  dffits.top <- head(rev(sort(dffits(the.linear.model))), topN)

  # Cook's over 1

  cooks.out <- which(cooks.distance(the.linear.model) > 1)

  # topN cooks

  cooks.top <- head(rev(sort(cooks.distance(the.linear.model))), topN)

  # Create a list with the statistics – can't do a data frame as different

  # lengths

  list.of.stats <- list(Std.res = res.out, Std.res.top = res.top, Leverage = lev.out,

                        Leverage.top = lev.top, DFFITS = dffits.out, DFFITS.top = dffits.top,

                        Cooks = cooks.out, Cooks.top = cooks.top)

  # return the statistics

  list.of.stats

}

Outliers <- show_outliers(model5, 10)  ;Outliers



#In practice you often get points that violate more than one criterion. we look for these points.

#We use the function intersect () which gives us the intersection of its two arguments.
```

```
group_of_outliers <- Reduce(intersect,list(Outliers$DFFITS,Outliers$Leverage))



#We extract the outliers from the dataframe and compare the summary statistics

#to that of the original data

Outlier_Values = training[group_of_outliers,]



summary(training)

summary (Outlier_Values)



#data less outliers

Data_less_outliers = training[-group_of_outliers,]



#new model less outliers

model8 <- glm(formula = Consumption ~ Age + Weight + Height +  Hypertension + Diabetes + Kidney + Thyroid + Allergy + Asthma +

Heart_disease + Post_traumatic_stress_disorder +  Irritable_bowel_syndrome +  Age:Diabetes, family = gaussian(), data = Data_less_outliers)

summary(model8)

rsq(model8, adj = "TRUE")



#diagnostic plots for the model

par(mfrow=c(2,2))

plot(model8)
```

```
#compare predicted vs actual on the test dataset

pred = predict(model8, data = test)

forecastVSactual =as.data.frame(cbind(pred, test$Consumption))

names(forecastVSactual)[1] <- "Predicted"

names(forecastVSactual)[2] <- "Actual"

head(forecastVSactual)



library(dplyr)

sample_n(forecastVSactual, 10)

#####################################################################

#Let's try a multinomial logistic model

dat4 = dat3

dat4$Consumption = as.factor(dat4$Consumption)

summary(dat4$Consumption)

str(dat4$Consumption)



library(nnet)

model2 = multinom(Consumption~., data =dat4)

summary(model2)



# Predicting the values for train dataset

training$precticed <- predict(model2, newdata = training, "class")
```

```
# Building classification table

ctable <- table(training$Consumption, training$precticed)



# Calculating accuracy - sum of diagonal elements divided by total obs

round((sum(diag(ctable))/sum(ctable))*100,4)

#from the %age of accuracy, multinomial logistic model is not the best,

#we therefore choose the Generalized linear model.

#############################################################################
```

CURRICULUM VITA

Abdulaziz Alidrees

I have pursued my Master of Industrial Engineering at The University of Texas at El Paso , August 2018 – Present. Thesis title : " Forecasting Water Consumption Using Generalized Linear Models For Better Supply Chain Resilience". In addition , Bachelor of Industrial and System Engineering (Dec 2017) , New Mexico State University. The academic award Outstanding Undergraduate Student, Department of Industrial and System Engineering , New Mexico State University , Spring 2016. I am a current member at Institute of Industrial and System Engineering , Society of Kuwait Engineers and Society of Kuwait Industrial Engineers

Contact Info :- Awalidrees@miners.utep.edu

Thesis was typed by Abdulaziz Alidrees