A MULTIPLE REGRESSION MODEL TO PREDICT TOURISTS' SATISFACTION INDEX

By

ELEANOR SIBORA

17/02447

A RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE IN DATA ANALYTICS IN THE FACULTY OF COMPUTING AND INFORMATION MANAGEMENT AT KCA UNIVERSITY

September, 2020

DECLARATION

I declare that this proposal is my original work and has not been previously published or submitted elsewhere for award of a degree. I also declare that it contains no material written or published by other people except where due reference is made and author duly acknowledged.

X

Eleanor Sibora
Student

Sign: Date: 17/09/2020

Eleanor Sibora **Student No: 17/02447**

I do hereby confirm that I have examined the master's proposal of Eleanor Sibora and have approved it for examination.

X

Lucy Mburu
Supervisor

Sign: Date: 21/09/2020

Name: Dr. Lucy Mburu

Lecturer, Faculty of Computing and Information Management,

KCA University.

ABSTRACT

The tourism industry appears to be one of the fastest growing industry all over the world. This growth can boost a nation’s economy. Nonetheless, much effort is needed for a nation to harness the socioeconomic benefits attributed to the growth of its tourism sector. One of the ways though which a country can capitalize in the growth of tourism is by use of mining Big Data for insight to improve its strategic approach to boosting tourism. In this regard, this paper reviewed several yet relevant past studies about tourism and its socioeconomic implications. Based on the findings in these reviewed literatures, this paper acquired specific socioeconomic data and developed a multiple regression model to predict tourists’ satisfaction. As hypothesized, GDP per capita, social support, health life expectancy, freedom, generosity, and corruption perception, part of a nation’s socioeconomic indicators, can be used to predict a tourist’s satisfaction. The paper concluded that it is possible to predict tourists’ satisfaction with the developed model. Moreover, Big Data can be mined and its insight used to advise tourists, an approach that can boost a nation’s tourism industry.

ACKNOWLEDGEMENT

I am grateful to God Almighty for the provision of good health throughout the time of my study. My gratitude and appreciation also goes to my supervisor Dr. Lucy Mburufor her tireless and enthusiasm to guide whenever required. I wish to also acknowledge the untiring support and inspiration of family and friends. I would also like to thank my fellow students with whom we networked and supported each other throughout the research period.

TABLE OF CONTENTS

DEDICATION

I wish to dedicate this work to my close family, friends, and fellow colleagues for their unwavering support during my undertaking of this study.

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

1.1 Background of the Study

Tourism in Kenya is very important because it is the second-largest source of foreign exchange revenue and it also provides direct and indirect employment opportunities(Maingi, 2019). For this reasons there is need for accurate tourism forecasting since all this will add to the economy of the country. Kenya Tourism Board (2018)released a report on tourism sector performance 2018 which showed hospitality industry is on an upward growth trajectory. For the exception of the past few months where Kenya’s tourism sector, similar to the globe’s tourism sector, has been heavily hampered by Covid-19, Mohammad, Gholipour, Feizi, and Nunkoo, (2020) expect the industry to pick up once the pandemic is contained. There has been a substantial extent of growth in relation to domestic tourism performance and pays (Kenya Tourism Board, 2018). The report seems to agree that there was a domestic bed night’s growth by nine per cent which was attributed due to the following factors: political stability, improved security situations, investor confidence, revitalized marketing strategies and digital marketing(Gwalik, Kabaria, & Kaur, 2015).

Bangwayo-Skeete and Skeete, (2015) recommend that mining big-data should bepart and parcel of an organization's marketing strategy and should not be taken lightly. One of the contributors to big data is social media data, which can depicted two interconnected roles of promoting in a marketplace (Gunter & Onder, 2015). The social media data is critical in establishing a connection between companies and customers and a connection between customers. It is also deemed to be critical to be enabling customers to communicated with their companies (Esfahani, Tavasoli, & Jabbarzadeh, 2019).

In this 21st century, social media and other Big Data are on the rise with billions of users all over the world contributing to the size and value (Esfahani, Tavasoli, & Jabbarzadeh, 2019). It seems that both tourists and nationshave access to more information because through social media and other information systems people generate their own content by sharing their views and experiences through the use of(Onder & Gunter, 2015). Shared social media content also becomes influentialin the provision of information which in turn improves on the reputation and performance of the tourism sector(Procter, Crump, Karstedt, & Voss, 2017).

Big data is defined is mainly composed of a large data set that can be analyzed to capture the trends and relationships which are linked to human behavior and interactions(Ram, Zhang, & Williams, 2015). According to Song and Liu (2017) big data is no merely defined by the 4v'sbut its complexity in essence the focus should be on its details. However, a majority of the definitions are mainly associated with volume, velocity and variety of the data. With this, Tifekci (2015) advises that socioeconomicdata can be mined to give more insight into tourists' satisfaction and or preferences.

These insights, if well addressed, can advise the relevant stakeholders within the hotel and tourism industry on how to attract tourists(Alcantara-Pilar, Crespo, del Barrio-Garcia, & Porcu, 2017). As Alcantara-Pilar, et al. (2017) echo, an increase in tourism ultimately leads to a boost in the local economy. However, Alcantara-Pilar, et al. (2017) and Buhalis & Amaranggana (2015)insist that the journey to achieve such achievements, using big data to model tourists' satisfaction in efforts to capitalize on the derived insight is not an easy task. Data from social media and socioeconomicindicators can be mined and used to train a model that can predict tourists'satisfaction on unseen datasets(Yang, Pan, & Evans, 2017). Nevertheless, Kenya's tourism and hotel sector is far from employing big data as a tool to boost the sector. As such, it is pivotal to look at the areas that pose challenges to the use of big data in boosting tourism by predicting their tastes.

1.2 Problem Statement

There appears to be very little adaptation of mining big data as a tool to understand the tourism sector (Ardito, Cerchione, Del Vecchio, & Raguseo, 2019). It is likely that social media data as well as economic dataon nations collected over the years remains unutilized. This is evident in Ndivo, Waudo, and Waswa (2016)study where they claim that Kenya is yet to discover the full potential of its tourism sector. Also, theyanticipate that the tourism sector will continue to remain economically stagnant.

The tourism industry is linked with the creation of jobs and an increase in the level of economic growth(Yang, Shang, & Kiang, 2015). In the United States, the tourism industry generates a total of 2.2 trillion annually and the creation of a least 101 million jobs(Yang, Shang, & Kiang, 2015).However, there seems to be very little development of new strategies which help in the development of new strategies, products, services and innovations to help boost the East African tourism industry(Sheoran, 2017). The lack of modern and effective strategies affects the quality of promotional informationof travel products and services provided to customers which would help them make informed decisions when choosing their vacation destinations (Xiang, Schwartz, Gerdes, & Uysal, 2015).

Ndivo, Waudo, and Waswa, 2016) did a research on the Kenyan domestic tourism market. Their study attributes the findings that the reasons for the skewed tourism development to poor strategic planning within the sector due to underutilization of insights mined from data. Kenya's tourism sector has not exploited the benefits of Big Data as a source for valuable insight, they claim. However, the stretch of obstacles that impede Kenya's tourism sector can be attributed to the lack of ample studies which have been conducted in Kenya with a focus on harnessing insight from Big Data to advise the tourism sector(Kenya Tourism Board, 2018). They add that very few studies have been conducted which focus on the development of models which would help in the prediction of tourism trends.

The tourism industry in Kenya is receiving heavy investments from both private and public sector which provides a very good opportunity for innovation(Kenya Tourism Board, 2018). However, the lack of inadequate skilled personnel who would advise on the adaptation of modern technology such as the use of Machine learning to model predictions and inform the tourism industry continues to be a challenge (Ndivo, Waudo, & Waswa, 2016). As such, it appears that there are some factors that impedes the nation's ability to capitalize on the potential of the tourism sector. These short comings were critical in advising the next section on how possible solutions.

1.3 Objectives of the Study

1.3.1 Main Objective

The main objective of this research is to develop a multiple regression model to predict tourists' satisfaction index of a nation.

1.3.2 Specific Objectives

The main objective will be achieved through the following specific objectives.

i) To identify some attributes that affect tourism satisfactionindex when vacationing.

ii) To develop a multiple regression model that can predict the tourist satisfaction index.

iii) Evaluate the resulting model against a real world scenario/dataset and highlight its performance.

1.4 Research Questions

In efforts to attain the overall objective, this paper was guided by the following research questions:

a) What are the factors to be considered in relation to predicting the tourist satisfaction index?
b) What should be considered in developing a model for predicting the tourist satisfaction index?
c) How best can someone implement and evaluate a model for predicting the tourist satisfaction index?

1.5 Motivation of Study

The attraction of tourists is mainly dependent on propermarketingan aspect achievablethrough the usage of big data(Ardito, Cerchione, Del Vecchio, & Raguseo, 2019). Big data helps in the identification of tourist's destination, tourist preparation plans and experiences which is inclusive of how they wantto travel. This mainly translates to an intense level of customer-focused marketing which helps ensure the satisfaction of customers(Xiang, Schwartz, Gerdes, & Uysal, 2015).

Many tourism attraction sites like national parts have readily available data but they aren't quite sure what to do with it. This study will transform this data into knowledge that will help increase the number of local tourists in our national parks.

In efforts to attract both local and foreign tourists, this paper acknowledges the recent pandemic, Covid-19 as an impediment to the sector. The Covid-19 pandemic has led to countries restricting travel and shutting down major destinations, a move that has had some detrimental effect on all economies globally(Mohammad, Gholipour, Feizi, & Nunkoo, 2020). However, as of the writing of this paper, parts of the globe are opening up to foreign travel. Tourists destinations, both locally and foreign continue to open up doors per their government advise. As such, this study is timely of offering an avenue through which our nation can employ to boost back the tourism sector.

This research was also inspired by the lack of studies onthe topic big data analytics for the local tourism market in Kenya using data from social media to predict the market. Since we have a lot of social media data laying around, we can transform it to knowledge thatwill create a positive impact on Kenyans to ignite the spirit of being patriotic and proud of our country Kenya by visiting various attraction sites and places.

Unemployment is a big problem in Kenya. This study will help identify the gaps where product and services are needed and key areas to improve as a result more youths will get something to do. This will also apply to the public and private sectors that are offering tourism services on the areas to improve more.

1.6 Significance of Study

This study will help KWS to understand local satisfaction to its stakeholders and performance of its strategy. This will lead to a better relationship between KWS and local citizens in Kenya.

In the hospitality industry tourism takes a huge part and both of those industries cannot become success without each other. Thus an excelling hospitality industry must depend on a better performingtourism management and marketing. Boosting local tourism will increase the number of visitors to hotels hence boosting the hotel industry.

This study will enhance the level of awareness among citizens especially on the importance of tourism and hospitality. It will be critical in understanding the importance of tourism system and the dimensions of hospitality e.g. skills, attitudes and qualities.

Tourist consume other products and services like food transport accommodation etc. This study will help create job opportunities to the citizens who will be providing the above mentioned products and services.

This study will help tourists in making better decisions about a particular attraction place before visiting. Since each individual has different test and preference this study will help derive the exact rating of a particular site/ attraction point making it easier for a tourist.

This study is important as it investigates the influence of big data analytics from social media data to predict local tourist in Kenya, and will come up with the recommendations to turn around this decline, so that the industry can attract more local tourist, that will contribute to the growth of the economy.

This research will help policy makers to create policies that will impact the local tourism industry positively. The main aim of this policies will be to have a growth in the local tourism in the country which will have an impact on the economy.The result may encourage the government on the importance of developing tourism destinations since both the local citizen and the government will benefit. It will create job opportunities to the citizen which will increase the economy. An increase in the number of the local tourists will also be seen in the tourism destination areas.

Kenya Wildlife Service can benefit from data from social media in many ways. That includes pinpointing marketing campaigns,better understanding the tourist’s habits, preferences, behavior patterns, and needs hence offering packages tailored the visitors’ likely interests.This way they can improve the tourist product and service they are selling to local travelers. Starting from marketing the service and finishing with infrastructure improvements.

Future researchers will have an understanding on what has been studied first and enhance it further.They will enhance it by creating new hypothesis in the tourism industry.

1.7 Summary

The rest of this proposal is structured as follows: Chapter 2 reviews the literature review on tourism industry and big data analysis. It will include the challenges and benefits of this study and the different predictive models researched before. Chapter 3 covers the research methodology. It looks at how the specific research objectives will be achieved,research design,the target population, sampling technique, data collection, and data analysis. Finally, the study will present its findings and analysis in Chapter 4, which will be succeeded by the conclusion and recommendations of the study in Chapter 5.

CHAPTER TWO:LITERATURE REVIEW

2.1 Introduction

This section will provide a literature review on Dig Data and how insights drawn from their analysis can be used toapproximate a tourists' satisfaction index of a nation. In addition, it will also capture review severalmachine learning and statistical models which are being used in the forecastingtrends within atourism sector of a country.

2.2 The Tourism Industry

The travel and tourism industry is tremendously important as it contributes USD 2.2 trillion to the world GDP (gross domestic product), and generates 101 million jobs worldwide (Yang, Shang, & Kiang, 2015). The size of this industry and the number of travel related players is likely to encourage competition to attract visitors. Promotion of travel products and services provides customers with information and knowledge in a persuasive manner, hoping to result in sales of the services (Sharma, Tim, & Wong, 2016).

There are many different travel information sources available such as personal sources (e.g. friends and relatives via word of mouth) and market dominated information (e.g. materials from hotels, airlines etc.) as well as socioeconomic data that can advise the tourism sector (Buhalis & Amaranggana, 2015). The internet appears to serves as one of the biggest source of information which is inclusive of data from tourist companies. According to Bangwayo-Skeete and Skeete, (2015) search engines are the starting point of interaction with DMOs and for potential and existing visitors. In this regard, Google is likely to be the best search engine that fits for the online tourism domain (Xiang, Schwartz, Gerdes, & Uysal, 2015).

The use of the internet has significantly been changing since the 1990s with the major reason being the need to ensure reliability and confidence on their queries(Bangwayo-Skeete & Skeete, 2015). In this 21st century, the internet appears to have become a must have necessity needed in the daily life of an individual like a cell phone. A research done to a Pew Research Center showed that the internet is the most important technology that the responders found hard to live without toping with 46% of the responsesfollowed by 44% responses for mobile phones (Park, Ok, & Chae, 2016).

One of the main information sources on the internet is search engines such as yahoo, Google or Bing. Among all the search engines, with Google (67%) has the highest percentage use of them and with about 5.9 billion average searches per day(Zhu, Cui, & Wang, 2015). In addition, they add that Google provides the search data at an aggregate level on its Google Trends page (http://trends.google.com/trends/), where users can identify the trending topics in search results or investigate a search term to find out its popularity in different parts of the world. This data is open, free of charge to Google account holders, can be downloaded in common spreadsheet format and used in different ways for analytical purposes.

2.2.1 Challenges in the Tourism Industry

Tourism is about people traveling for fun which involves movement of people from one place to another for reasons of sightseeing, pleasure and or camping(Watson, 2015). The past decadeshave seen a significant increase in the use of social media adoption and usage(Ardito, Cerchione, Del Vecchio, & Raguseo, 2019). Previous big data studies have shown a great improvement in international tourism industry compared to the local tourist (Sheoran, 2017).However, there are also a significant number of challenges which influence industrySheoran adds. One of these challenges is the security concerns which discourage tourists from visiting some tourist destinations. Similarly, cyber-attacks can also lead to the leakage of sensitive tourist information he which can also prevent the tourists from sharing their location data (Alcantara-Pilar, Crespo, del Barrio-Garcia, & Porcu, 2017).

Travel marketing can also be exaggerated and some tourism advertisements can be deemed false. There is still a need to develop innovative marketing solutions which would be instrumental in luring tourists who are increasingly becoming informed(Buhalis & Amaranggana, 2015).Another significant concern is the quality of tourism infrastructure in the various tourist destinations. In developed countries there are better infrastructure which have been set up to encourage tourism contrary to most developing countries. Issues like faster immigration in airports, efficient hotel checkout and proper transportation are still significant challenges for most countries (Cillo, Rialti, USai, & del Giudice, 2019).

2.3 Big Data Analysis

Big data is mainly characterized by large volumes of data which are hard to store and analyze and especially through the use database technologies(Sheoran, 2017). Big data is mostly indistinct and various process are utilized in order to identify and translate the data in order to capture insights (Sharma, Tim, & Wong, 2016).It is a process that isassociated by data analysis in order to get meaning patterns and relationships which would be of benefit in the tourism industry(Sheoran, 2017). Analyzinga data set to get information has always been considered by researchers and practitioners, but the veracity, big volume, different variety, and high velocity of real-time data streams seem to be pushing the limits of the current processing, management capabilities and storage(Watson, 2015). Using a set of storage and computational resources called cluster, we can analyze big data concerning their amount and speed, satisfying in that way the main principle in big data analytics, the need of high performance in computations (Tifekci, 2017).

2.3.1 Classification of Big Data

Big data is mainly classified into four categories which is inclusive of velocity, value, volume and variety as depicted in Figure 2.1

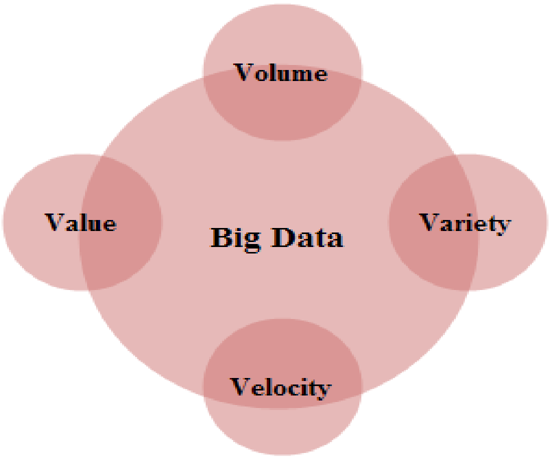


Figure 0.1 Classification of Big Data

2.3.2 Volume

Big data is characterized with large data quantities of data which created and stored in every second which are in turn observed and tracked(Sharma, Tim, & Wong, 2016).The quantity of data that is generated is very important in this context(Wood, Guerry, & Silver, 2015). Facebook has 2.38 billion monthly active users who sends over 10 billion messages per day. Instagram post 67 million posts per day(Williams & Burnap, 2017).

2.3.3 Variety

This refers to the different types of data we can now use collected via sensors, smartphones, or social networks(Williams & Burnap, 2017). They explain that data can come in the form of text posts, images, sound snippets,log data, and many more they. The diversity of this information helps to gain a more complex insight into the customers and their habits but also brings with it the problem of unstructured and hard to handle bundles of data(Wu, Liao, Tseng, & Lim, 2017). More specifically, the data types as depicted in the below.

Data Type	**Description**
Structured data	This is composed of data arranged in a definite pattern and it is composed of all entities like similar attributes, length and a specified format in similar order e.g. Relational Database Management Systems (RDBMS).
Semi-structured	This data type can be supported by using graph data structures and includes scientific data, bibliographic data, etc.
Unstructured data	This is data with data that has no standard data structure, e.g. videos, images, documents, etc.
Multi-structured data	An association of structured, unstructured and semi-structured data, e.g. sensors data, operating system's log files, and customer interaction streams (Makridis2018).

Figure 2.2: Data Types in Big Data

2.3.4 Velocity

This mainly refers to the speed of responsiveness which is usually inclusive of three aspects(Yaqoob, Hasheem, Gani, Mokhatar, & Ahmed, 2016). The first is the level of constant capture, storage and analysis of information of big data. The second aspect deals with timeliness or latency as such it should capture, store and use big data in certain lag time.This is mainly because the data is deemed to be meaningful over a short period of time. The last aspect mainly deals with the speed big data in relation to how is stored and retrieved(Wu, Liao, Tseng, & Lim, 2017). This is inclusive of architecture, analysis and deployment which must be done consistent over thousands of new customers (Xiang, Schwartz, Gerdes, & Uysal, 2015).

2.3.5 Value

Refers to benefits or importance ofthe data which enables an individual or company to collect leverage about a company (Yang, Pan, & Evans, 2017). It helps in the capture of data which can help in the assessment of cost and benefits (Yang, Shang, & Kiang, 2015).

2.4 Benefits of Using Big Data Analysis in Tourism

Big data as a topic in the recent years has such a huge following and its becoming more popular as days does by. Big data analytics and decision-making process are being integrated to provide significant benefits for big companies like Facebook, Google and amazon (Zhu, Cui, & Wang, 2015). There is a widespread belief in executives, managers, and analysts that big data analytics can provide value (Yaqoob, Hasheem, Gani, Mokhatar, & Ahmed, 2016).

Narok County mainly depends on tourism as it is the main source of livelihood for many individuals in the area(Maingi, 2019). In essence, 40% of the population rely on making and hawking in order to have a livelihood according to Maigi. Other individuals also rely on formal employment which are attributed to tourism e.g. working lodges, tour companies and as watchmen(Smeral, 2015).

The hotel and tourism industry mainly thrives on the sale of the right products and services to their customers at the appropriate time and price(Song & Lui, 2017). In essence, they explain that companies can differentiate their services from those of the competitors through use of big data. Big data is also critical in determining the preferences of customers(Gwalik, Kabaria, & Kaur, 2015). Companies can therefore utilize this information in order to acquire valuable information which could be used to maximize sales, ensure customer acquisition and loyalty and ensure profits. Companies mostly use big data in order to maximize their marketing efforts and capture the needs of their customers and provide these services at budget friendlyprices.

Another big benefit of big data analysis is the feedback mechanisms(Gunter & Onder, 2015). In this 21st century customers are leaving reviews in social media platforms and new potential clients are checking those platforms for a review before investing in one. This creates a need for reputation management which can be made possible by using big data analysis(Song & Lui, 2017).

In the tourism industry, feedback is critical in order to capture customer preferences and ensure positive growth as such soliciting feedback from customers helps ensure growth and capture proper customer needs (Gunter & Onder, 2015). This feedback is mostly captured by consumer apps and company sites which is critical in developing proper information (Gunter & Onder, 2015).

2.4.1 Big Data and Socioeconomic Indicators

All businesses mostly thrive on understanding the behaviors and trend of their customers in the greatest form possible. The online platforms and information systems serves as critical ways through which information is collected for use by companies in order to ensure growth and success (Smeral, 2015). It is becoming increasingly important for companies to invest on big data analytics like through the use of social media through sites like Facebook, Twitter and LinkedIn (Sharma et al., 2015). Also, such data can be complimented with other socioeconomic data residing in information systems can be mined for insights (Ndivo, Waudo, & Waswa, 2016).Different strategies are therefore developed in order to gain intelligence e.g. through advertisements, promotion of customer loyalty, keeping tabs on the market trends and those of their competitors (Park, Ok, & Chae, 2016). According to Zhu et al., (2015) most companies research on this information in order to understand their customers and also improve on theiroverall performance.

2.4.2 Tools and Metrics

An increased number of information sources is significantly linked with an increase in the development of tools that would enable the access of information which would enhance visibility on the websites (Esfahani et al., 2019). According to Sharma et al., (2015)gathering metrics like countries, websites and keywords from web pages and websites are critical in improving company performance.

2.4.3 Website Rankings

The ranking of websites is mainly used to specify the popularity of websites in relation to other websites over a specified of time. This is mainly achieved through the use of tools like www.ranking.com and www.alexa.com. These ranks could be used in estimating the popularity of websites and to make a comparison of their competitors(Smeral, 2015).

2.4.3.1 Online Traffic Analytics

This is mostly inclusive of online tools like Google analytics and ww.alexa.com which provide traffic metrics on websites which could in turn be customized by users. This information is created on spread sheets which can be optimized by organization in order to derive meaning (Esfahani et al., 2019). Information that is provided in these tools is mostly composed of the number of visits, unique visitors, number of webpages used average visit durations etc. (Ram et al., 2015).

2.4.3.2 Social Media Monitoring

The social media has significantly been growing over the last two decades and also growing in usage(Ram, Zhang, & Williams, 2015). Similarly, various companies have captured the importance of the social media especially in marketing and ensuring growth of the company. Websites like Facebook, Twitter and Linked have proved to be instrumental in improving the company's connection with their customers. Similarly, organizations can use the social media to interact with their stakeholders and employees e.g. through the use of WhatsApp and Yammer which would improves the level of interaction (Yaqoob et al., 2016). Gwalik et al., (2015) notes that companies can use information from social media to capture information and monitor news on key contributors through online conversations. This information can be measured to monitor the potential problems and business in turn benefits through interactions with the online users.

2.5 Attributes for Predicting Tourist Satisfaction Index

In order to pre-process data for mining it is critical to distinguish both ideal and available attributes within a dataset. This will be important in making a model that is relevant to its course. This section explores some available literature that had studied and documented the socioeconomic data on the use on advising tourists.

One of the success factor in a touring or vacationing is to be happy (Song & Lui, 2017). In their study, they employed a Support vector regression model to forecast the relationship of a nation’s average satisfaction index against the country’s GDP per capita. Based on their findings, Song &Lui, (2017) assert that the GDP per capita of a nation is an indicator of what infrastructures one can expect in a nation. They go ahead to claim that good and reliable infrastrutures contribute to the wellness of a nation’s population as well as it’s foreign visitors. Moreover, Song &Lui, (2017) explain that a good GDP per capita closely implies that the social support functions of the nation’s population are sustainable. This seems to be inline with the assertion that countries with high levels of corruption and lack of public confidence always undermine a nation’s perfomance thus affecting its GDP (Ardito, Cerchione, Del Vecchio, & Raguseo, 2019). Their study used a trained Neural network to explain why coutries cannot attract foreign investments due to fear of unaccountability, a case that undermines affects other aspects of life within the nation inclusing security. So if a poor corruption perception by citizens as well as foreign nations can affect a countries security, Song &Lui, (2017) advise that this undermines tourism in that nation.

Consequently, Song &Lui, (2017) add that nations should invest heaily on their populations especially on matters health. They explain that a healthy nation is a happy nation. Moreover, tourists aspire to be visit countries with a good healthcare system (Buhalis & Amaranggana, 2015). Buhalis & Amaranggana, (2015) exploratory study explain that the medicare is a vital component for tourists because of the health risks involved when travelling to foreign nations. A change of environment and food can affect a tourist’s health. As such, tourists are likely to be happier if the can be assured of a reliable and affordable healthcare service when vacationing (Buhalis & Amaranggana, 2015). Moreover, Buhalis & Amaranggana, (2015) add that free and liberated nation tend to have more tourists. This is due to the fact that freedom is closely associated with fun as highlighted by Buhalis & Amaranggana, (2015). Nations with an oppressed society be it from their government and religion can inhibit the freedom of culture that tourists seek. Because of this, Buhalis & Amaranggana, (2015) explain disperity of oppression within the united arab emirates has left the nations with a higher level of oppression attracting fewer tourists that those perceived to be more liberal.

2.6 Models for Forecasting Tourist Trends

According to Song and Li (2017), in major academic journals between 2010 and 2017, the majority of the tourism forecasting studies are conducted by using quantitative methods. These studies can be divided into 2 categories: non-causal methods (i.e. time series analysis, artificial neural network) and causal methods (i.e. econometric models), which include additional explanatory variables other than past realizations of the tourism demand variable itself. Time-series analysis is usually conducted by using ARIMA, SARIMA, and GARCH models as well as naïve-1, seasonal naïve, naïve-2 and exponential smoothing methods. The Error-Trend-Seasonal or Exponential Smoothing (ETS) model class has recently emerged as a complement to the traditional exponential smoothing methods in tourism demand forecasting(Esfahani et al., 2019).

Concerning econometric models, the most popular methods arethe error correction model (ECM), the vector autoregressive (VAR) model, the time-varying parameter (TVP) model,and autoregressive distributed lag model (ADLM), (Ardito et al., 2019). According to Song and Li (2017) there was only one study that was conducted at city level, which was by Vu and Turner (2016), between 2010 and 2017. Kim and Schwartz (2015) come to a similar conclusion in a recent survey article.

Gunter &Önder (2015) and Smeral (2015) are two recent examples for city level studies using monthly data.

In addition, previous research regarding tourism demand forecasting out of 121 studies only 30 of them used monthly time series (Song & Li, 2017). According to Witt and Witt (2015), 23% of the tourism forecasting studies between 2008 and 2016 also included monthly time series as well. Econometric models in particular have not been frequently applied to monthly tourism demand data so far.

A majority of these models measure specific factors in order to determine the tourism trends from big data e.g. number of visits in a site, page clicks, seasons etc. however, there are still a number of factors which influence the tourism trends which is inclusive of tourism preferences and behaviors.

2.7 Predictive Analytics Models

2.7.1 Overview of predictive analytics models

Predictive models carry out analysis to recognize patterns found in transaction data and historical data to discover various potentials and risks. The model captures the relationship between different attributes to allow the assessment of risks or opportunities associated with a list of specific conditions to guide the decision-making of candidate transactions (S. Poornima& M. Pushpalatha, 2018). Predictive analytics is usually performed with other statistical techniques. The commonly applied techniques are Regression Analysis, Artificial Neural Network, Time series Analysis,Decision trees, Factor Analysis, Naive Bayes etc. In the study, we use Multiple Regression tool for predicting rejection from each process.

Supervised machine learning happens when a predictive model is creating using part of historical data which has the results you wish to predict. Classification and regression modes support the approach of supervised learning that maps input to outputs. Classification techniques identifiesthecategory a newerfield belongs to (i.e., event or customer) based on its inherent characteristics. Regression analysis uses the past values to predict future values by computing the probability relationship between variables for the purpose of predicting and it is mostly used in variance and forecasting analysis. Time-series analysis is very similar to regression but allows the unique attributes of time and calendars to be used to predict the seasonal variances, among other things.

This study used a supervised machine learning approach for prediction. Since regression analysis end product is predicting a number, while classification predicts a category, the study choose to use regression analysis for this reason and also because of the labeled dataset that was used in the study. Multiple regression works well with a dataset that is has a linear factor in it while decision tree regression support the non-linearity in a data by splitting the data into smaller data subset which usually depends with the question asked. Decision tree handles co-linearity better than multiple regression. Larger numberof attributes with less data makes multiple regression outperform decision tree which generally performs average. The distinct nature in which the two models that is multiple regression and decision tree regression work is the reason the study decided to use both methods and analyze the better performing model.

2.7.2 Past predictive analytics models

Karwan andKone (2011) predicted about the expense incurred while delivering liquefied gas to a new customers while making use of the multi-factor linear regression model. The development of a single model, lead to very poor prediction outcome. Therefore, before doing the regression analysis, a supervised learning method was used for grouping the customers who had a similarity in some or the other perceptions. Hyper-boxes were used to denote the different classes on customers, and afterwards, a linear regression model was developed within each and every class. To increase with thecombination of the data regression and classification, the correctness of the prediction isindicated.

Ravi Kanthet al. (2012) did an algorithm for prediction of heartdisease while using case scenario based machine learning methods on non-binary datasets. There was a challenge in mining data-sets in non-binary search space compared to mining in binary search space. At first, the non-binarysearch space needed innovative strategy to calculate the support. As there isan opportunity of expulsion of applicant thing set from non-double information index because pruning applying it at a more elevated level may get continues. Each level of support calculation and candidate generation are performed using a separate mechanism.The final result of the author's was to make a prototype of frequent item sets for the non-binary data set developed.

Rudolph, et al. (2010) did a model which usedspatial-temporal data, becausethe analysts were looking into areas with the aspect oftime and space factorsthat were having unpredictably massiveoccurrences of events. They developed a visually predictive analytics toolkit which assisted theanalysts providing them with the linked statistical and spatiotemporal analytic views.Spatiotemporal events are generated by the system injoining the seasonal trend decomposition and the estimation of event distribution with the addition ofloss smoothing for the reason of temporal predictions.

Bhat et al. (2011) developed a new pre-processing phase with theremoval ofmissing value for both categorical and numerical data. A hybrid combination containing a regression and classification trees, genetic algorithms for removing themissing sequential values and self-organizing feature maps for removing thecategorical values were used in the work.

Yue et al. (2009) highlighted the predictive analytics jobs which werefamiliar to the prediction of future trends andthen introduced an artificial intelligence blackboard based agent leveragingthe interactive visualization and theproblem solving mixed-initiative so as tofacilitate the analysis to look for and preprocess detailed quantities of data which willperforming predictive analytics.

2.8 Conceptual Framework

In relation to the study objectives the prediction of tourism trends (Dependent Variable) will be dependent on the aspects of big data (Independent variable). These are inclusive of factors like the variety of data, page visits and clicks, customer preferences and behaviors, and quality of products and services as shown in
Figure 2.2

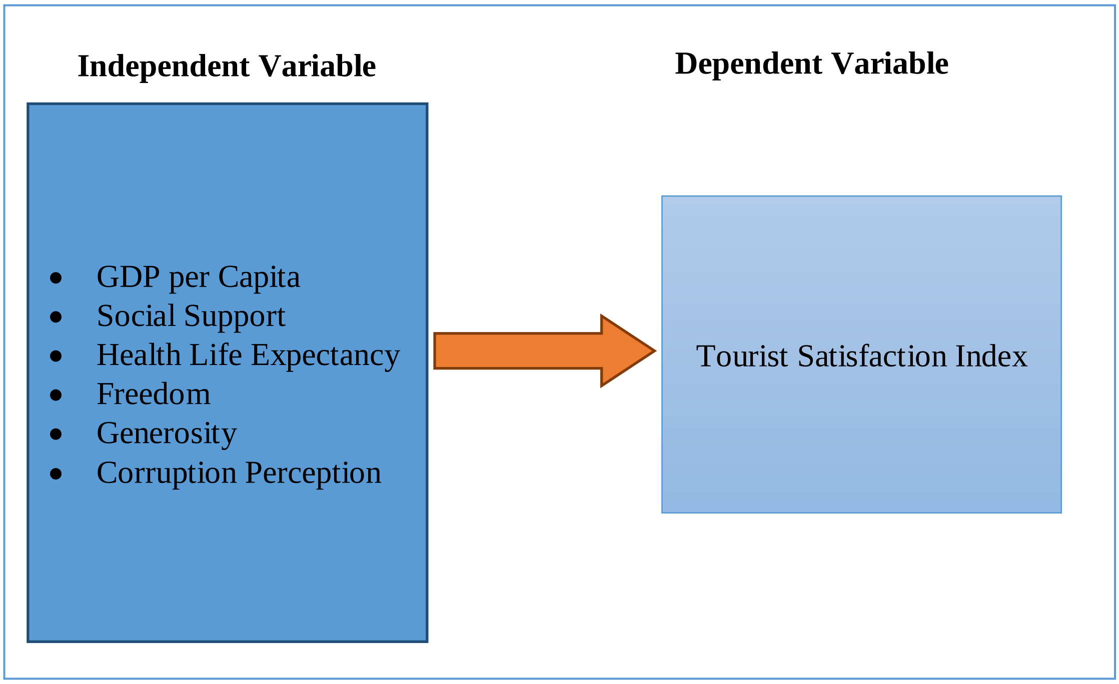


Figure 2.3: Conceptual Framework

2.9 Operationalization of Variables

Conceptual Variables	**Scale (Level of Measurement)**	**Operational Definition**
1. GDP per Capita	Nominal	A metric that provides a representation of a country's economic outcome per individual. GDP is arrived by dividing the Gross Domestic Product of a nation by the size of its population.
2. Social support	Nominal	A national's average to the binary response (either yes/no) to the question, "Do you have friends or family that you can rely on if you were in trouble?"
3. Health Life Expectancy	Nominal	The average number of years that is expected of a new child to live without being hampered by injuries or disabling illness.
4. Freedom	Nominal	The free will to make life choices, an average response by a country to the question, "Are you content or not content with your freedom to pick what to do with your life?"
5. Generosity	Nominal	The end result of regressing a country's average of response to the question, "Have you made any charitable donation in the past month?"
6. Corruption Perception	Nominal	The measure of a country's average to the question "Is corruption within businesses and the government wide-spread?"

7. Tourism satisfaction index	Nominal	The score of the 6 independent variables above. The more the score the higher the index of a particular country.

Figure 0.3: Operationalization of Variables

2.10 Research Gap

In relation to the presented literature on tourism, most studies which focus on the influence of big data only focus its benefits and its influence in the industry. There are few studies which ascertain the effectiveness of these models in forecasting tourism trends. Similarly, there are very few studies which focus on the tourism industry in Kenya and this study will therefore help fill this gap

CHAPTER THREE:RESEARCH METHODOLOGY

3.1 Introduction

This section presents the methodology that was used in the study which includes the methods for achieving the objectives, research design, population, sampling technique and design, data collection and data analysis.

3.2 Method for Achieving the Objectives

*Objective one: to identify some attributes that affect tourism satisfaction index when vacationing.*The study was guided by relevant yet in depth reviews of several literatures in efforts to identity the attributes, independent variables that are measurable and test their significance toward the dependent variable.

Objective two:to develop a multiple regression model to predictivea tourism satisfaction index using socioeconomic indicators. This was achieved through an analysis of the readily available datasets from public academic or research repositories. The study ensured that the dataset used was in line with its main objective in efforts to answer the research questions. The identified dataset underwent some data preprocessed. The processed dataset was divided into a training andtesting dataset. The training data set was used to train a predictive model that was used to predict a predefined tourism Satisfaction index. The resulting trained multiple regression model in this stage wastested against a section of the test dataset for validation checks.

Objective three:Evaluate the resulting model against a real world scenario/dataset and highlight its performance. Testing of the model was done by using the test dataset on the trained model to help point out any inconsistencies and/or ambiguities that might have been captured during the model development. In addition, a proof of concept/pilot was conducted to ascertain the viability of the model against a real world scenario.

3.3 Research Design

The selection of Design Science as a methodology in the development of the tool is designed to offer a solution to the research problem. This is informed by the versatility it offers as a knowledge elicitor through a well-developed mapping theory enabling creation of artefacts that satisfy given functional requirements parameters. This offers a good foundation for formation of constructs techniques and methods.

Designing of artefacts is crucial in solving observed problems, offering contributions to research its design evaluation ability and presenting the findings to audiences of interests. This makes design science an all rounded methodology that assures of thoroughness in knowledge discovery. Artifacts in design science are like models, methods, social innovations, instantiations and either social or informational resources.

This study employed a quantitative research design to address the hypothesized research questions. They, Mitchell & Janina, (2015) and Creswell & David, (2018) agree that a research design constitutes a map for the collection, measurement, analysis of data and the presentation of the resulting findings.A quantitative design employs mathematical, statistical, and or logical techniques on collected data. Additionally, this method limits the amount of resources that will be used in the study.

3.4 Study Population and Size

Many research tasks intend to examine varying characteristics of populations. A population is a group of individual bodies that possess a common feature, commonality (Burke & Christensen, 2019). This study intended to classify socioeconomic data on nations to predict tourists' enjoyment.

A sampling frame is a list of elements within a study population from which a sample is drawn(Burke & Christensen, 2019). In efforts to be thorough, this study considered the entire dataset as the sapling frame.

3.5 Sampling Procedure

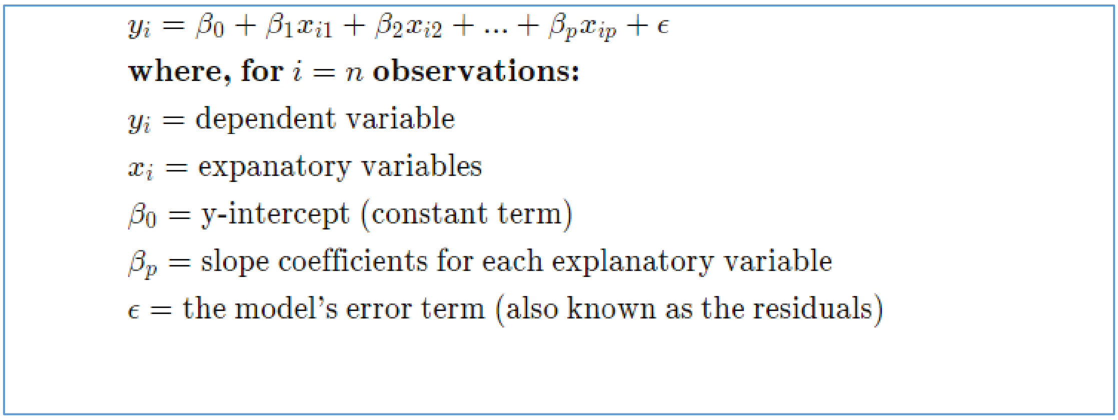
A sample is a representation of a given study population(Burke & Christensen, 2019). Thisresearch study used a probability sampling technique. This technique is feasible because it uses randomization to ensure that all members of a population have an equal chance of being selected. In addition, the study employed a stratified sampling approach as a way to shuffle the data when training the resulting model.

3.6 Data Collection

In research, data collection instruments are the devices that can be employed by a study to gather data(Creswell & Creswell, 2018).The research was reliant on primary data which was collected from publically available dataset from www.kaggle.com. The data collection instruments needed were a computer with an internet access.

3.7 Building and Training Predictive Model

The resulting models were developed using Scikit Learn, a machine learning library. In efforts to be thorough, this study developed and tested two models, Multiple Regression and Decision Tree Regression. A multiple regression model was used because it employs several explanatory variables to determine the outcome of a dependable variable (Goodfellow, Bengio, & Courville, 2016).



$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

x_i = expanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

Figure 0.1: Multiple Regression Formulae

Conversely, the Decision Tree Regression approach builds classification models in a tree line structures (Goodfellow, Bengio, & Courville, 2016). They explain that in such a model, the data set is broken down into smaller subsets whilst an association decision tree is built to form the resulting model.

3.8 Data Preprocessing

The efforts of preprocessing the dataset was to prepare it for training the resulting model. The entire dataset was checked for any missing values and none was found. Secondly, for all the columns that had a wide range within their values were normalized. Specifically, some columns (features) were normalized, a scaling technique where feature values are rescaled and shifted to a range between 0 and 1. Also, standardization was used on some features to prepare the dataset for the model. Standardization is a technique where feature values are centered on the mean (Burke & Christensen, 2019). Also, in efforts to understand the data, the study introduced a continent column. All preprocessing was done using the Pandas and Numpy Python libraries.

3.9 Model Assessment

After the models were created and trained using the preprocessed data, the remaining bit was to measure their performance on unseen data. This evaluation determined how well the models could make accurate predictions in the real world. Model evaluation can be achieved by calculating some metrics such as the R-Squared andMean-Squared-Error. The primary measurement metric for the regression model was the R-Squared, a figure with a value between 0 and 1. This metric is a measure of the extent to which changes of the dependent variable, Tourism satisfaction index, can be appraised by changes in the independent variables(Burke & Christensen, 2019). They also define Mean-Squared-Error as the measure of the difference of the expected value from the predicted value, a metric that affirms the predictive score of a model.

3.10 Ethical Consideration

All respondents will be required to consent to the study before participating in the study. The respondents will be assured of their confidentiality and no form of identification will be used required in the study.

CHAPTER FOUR: DATA ANALYSIS FINDINGS, AND DISSCUSSIONS

4.1 Introductions

In efforts to train a binary classifier to predict if a tourist will enjoy their visit, this paper adopted the Multi Regression to achieve its desired goals. Therefore, this chapter provides a detailed data analysis and the findings.

4.2 Descriptive Statistics

The study made use of a publically available dataset, https://www.kaggle.com as source of training and testing data. Also, the dataset used was vital due to its domain content, socioeconomic data on nations Satisfaction. The data contains information of various socioeconomic indicators. The entire dataset has 26,327 rows. A part of the dataset, 80% was used for training the resulting model and 20% used for validation and testing.

4.3 Demographic Information

The dataset was based on the survey of the state of the globe. Precisely, the socioeconomic indicators, features, represent a total of 155 countries globally. Based on the features of the dataset, versus we calculated an average Satisfaction score and discovered that of all the continents, Australia has the highest average of all features after Europe and America. Conversely, Africa and Asia are seen to have the lowest tallies in all fields.

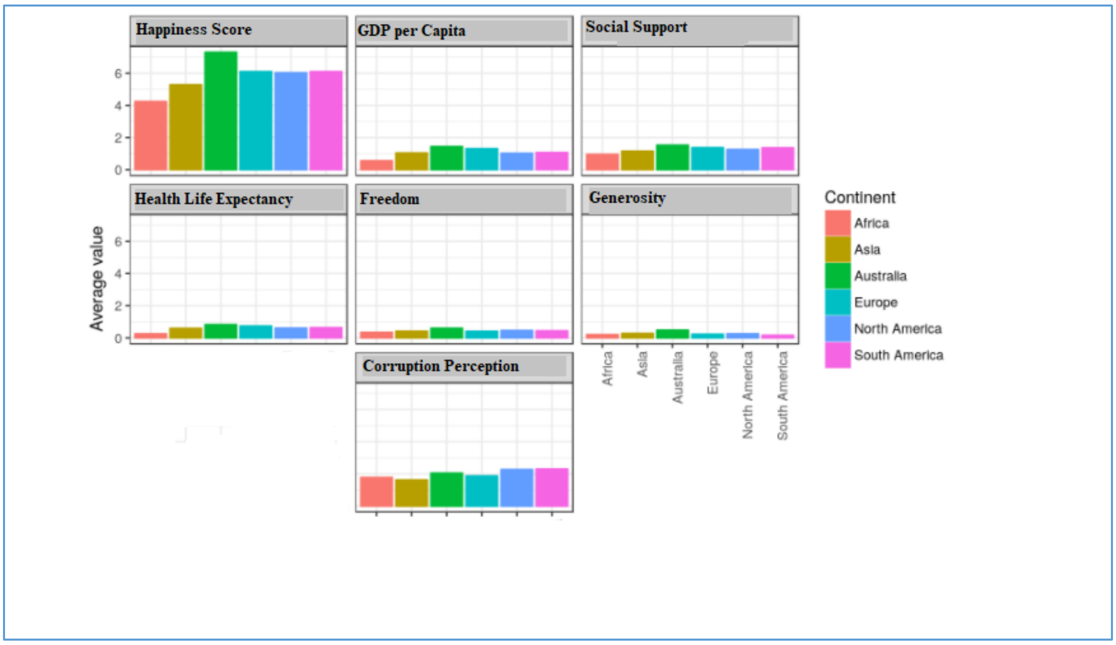


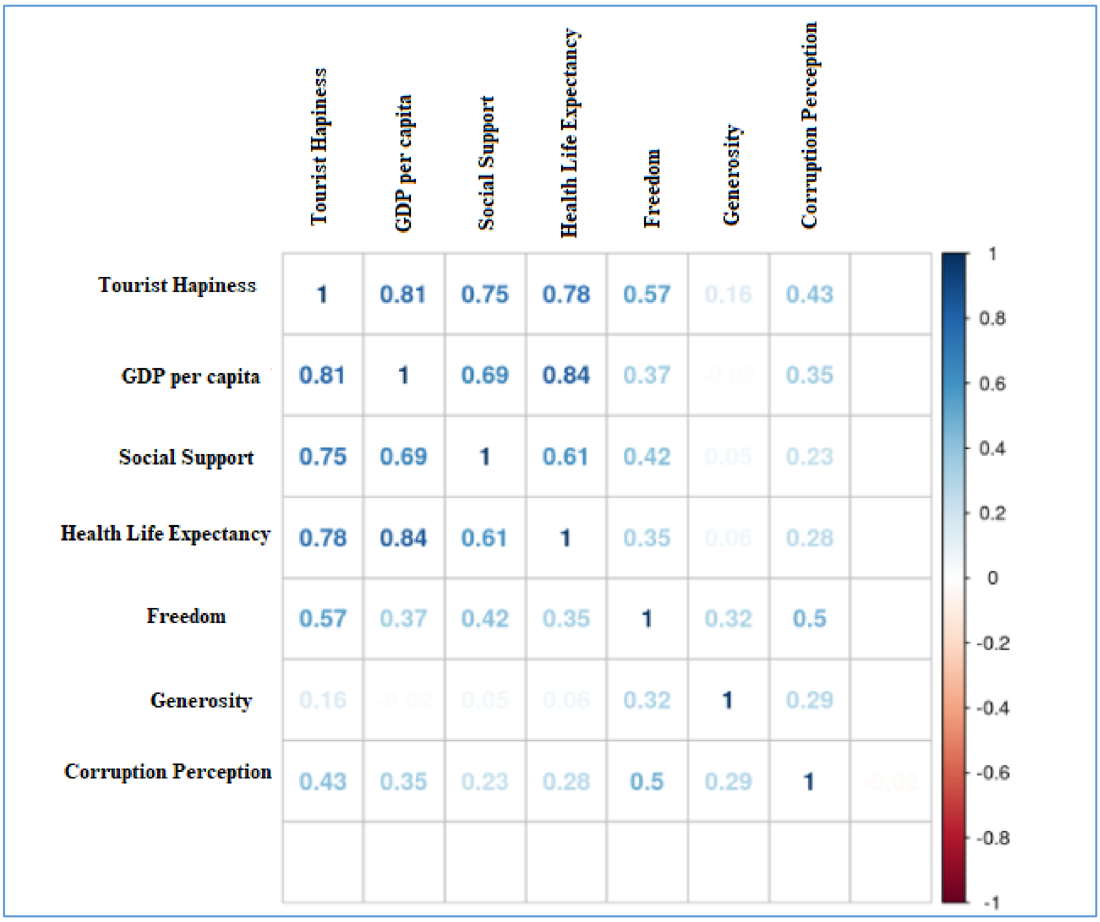
Figure 0.1: Average Value of Satisfaction for Every Continent.

4.4 Study Variables

In order to answer the research questions, this study identified 6 general independent variables which were defined in the operational definition of variables section. These variables are GDP per capita, social support, health life expectancy, freedom, generosity, and corruption perception were tested against the dependent variable, tourist will enjoy visit. Ideally, each of these variables was measurable as highlighted in the literature review section.

4.5 Diagnostic Tests

It was important to explore some diagnostic statistics of the data in order to visualize more insight about the data prior to training the model. One of the insights drawn was from the correlation between numeric features in the data. By dropping the dependent variable feature, Satisfaction index, we saw that GDP per capita, health life expectancy, and social support were of a higher significance when it came to contributing to the tourism Satisfaction score. However, from the same results we saw that Generosity and corruption perception had the lowest impact.



	Tourist Hapiness	GDP per capita	Social Support	Health Life Expectancy	Freedom	Generosity	Corruption Perception	
Tourist Hapiness	1	0.81	0.75	0.78	0.57	0.16	0.43	
GDP per capita	0.81	1	0.69	0.84	0.37	-0.02	0.35	
Social Support	0.75	0.69	1	0.61	0.42	0.05	0.23	
Health Life Expectancy	0.78	0.84	0.61	1	0.35	0.06	0.28	
Freedom	0.57	0.37	0.42	0.35	1	0.32	0.5	
Generosity	0.16	-0.02	0.05	0.06	0.32	1	0.29	
Corruption Perception	0.43	0.35	0.23	0.28	0.5	0.29	1	-0.02

Figure 0.2: Correlation Plot between Satisfaction Index and Independent Variables

Similarly, the study mapped the tourist Satisfactionindex distribution for all the continents in the dataset. These results were grouped per continent and their mean and median of the Satisfaction score was calculated.

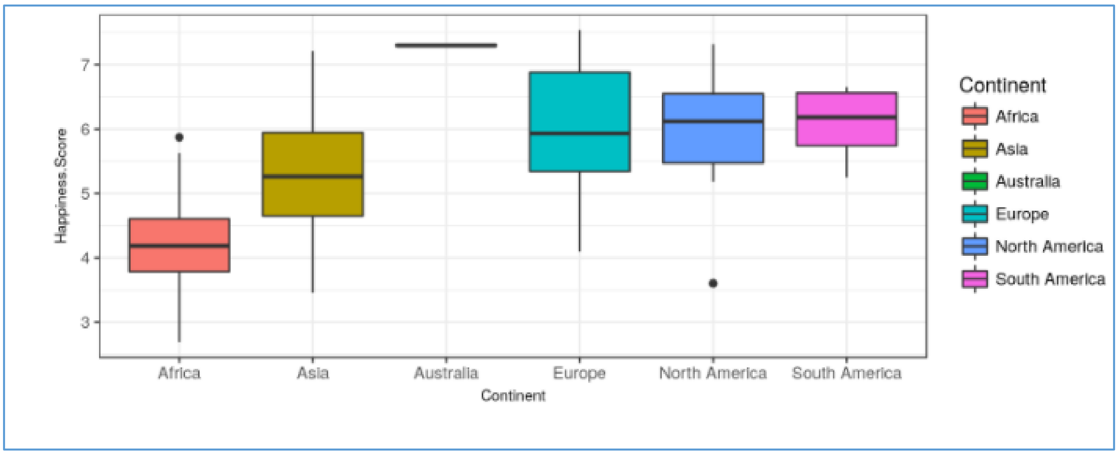
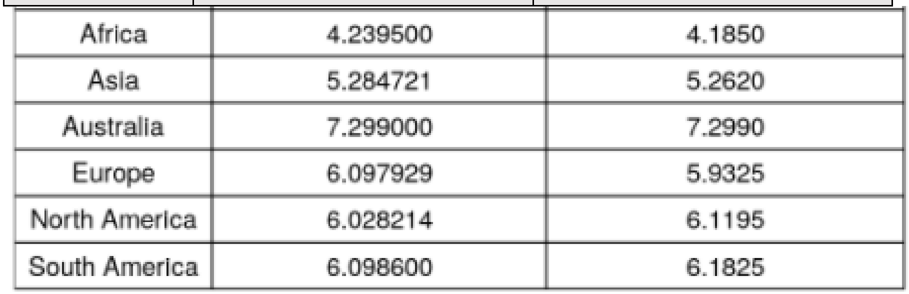


Figure 0.3: Satisfaction Index Distribution Per Continent



Continent	Mean of Satisfaction Index	Median of Satisfaction Index
Africa	4.239500	4.1850
Asia	5.284721	5.2620
Australia	7.299000	7.2990
Europe	6.097929	5.9325
North America	6.028214	6.1195
South America	6.098600	6.1825

Figure 0.4: Mean vs Median Values of Satisfaction Index per Continent

These diagnostic results appear to be consistence because we saw that Australia had the highest median score for the dependent variable, tourist Satisfaction. This was followed by Europe then America. On the lower end, we had Africa and Asia respectively

4.6 Model Training and Evaluation

The dataset was trained and tested against a Decision Tree Regression model based on Scikit-Learn machine learning library. The resulting model evaluation metrics was compared to the proposed Multiple Regression model.

4.6.1 Model 1: Multiple Regression

In efforts to predict a tourist Satisfaction index, the dependent variable, this model was based on the assumption that there is a linear relationship between the independent variables, and the predictor. Moreover, preliminary analysis showed that the independent variables were not too tightly correlated to one another. All the observations were randomly and independently selected when training the model. Also, all residuals were normally distributed. Lastly, the model was evaluated using the coefficient of determination, R-Squared. This metric was used to explain the variation in outcomein relation to the independent variable.

4.6.2 Evaluation of the Model (Multiple Regression Model)

We can see from the model that all independent variables are of significance when it comes to predicting the tourist Satisfaction index. Also, the adjusted R-squared was 0.81. With this we can conclude that there exists a linear correlation between our study variables and the dependent variable. Below is an extract of the model outcome and metrics.

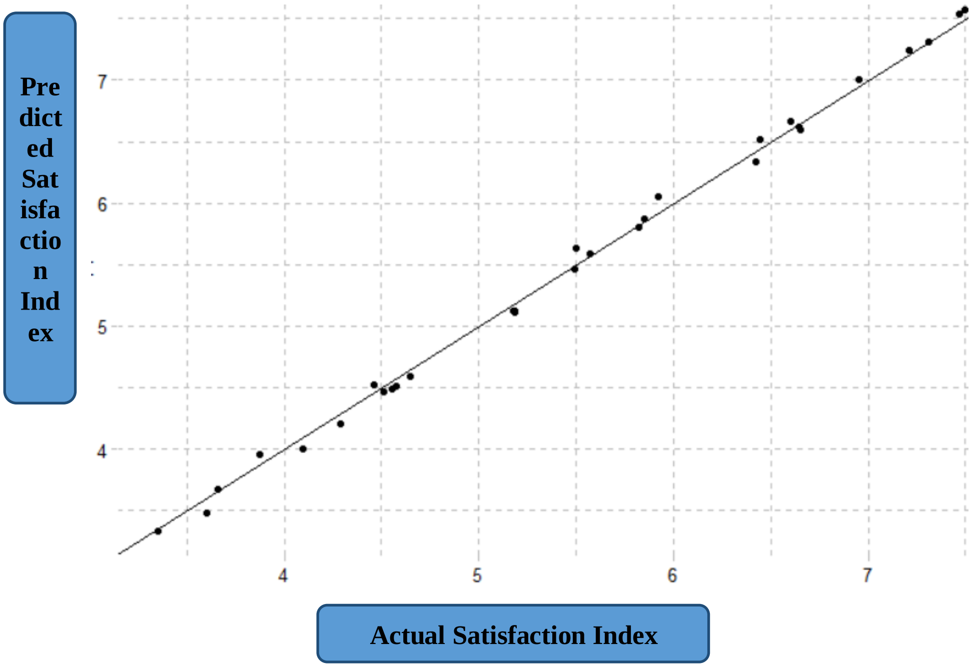


Figure 0.5: Actual vs Predicted Satisfaction Index Multiple Regression Model

4.6.3 Model 2: Decision Tree Regression

This study went ahead and adopted a Decision Tree Regression model and trained the model using the same dataset. The resulting model metrics were compared to the study's proposed Multiple Regression model. Its approach, the Decision Tree Regression aims at breaking down a dataset into smaller subsets and incrementally develop a related decision tree. By plotting the actual Satisfaction score and verses those predicted by the Decision Tree Regression model it is clear that this model is not ideal for the dataset.

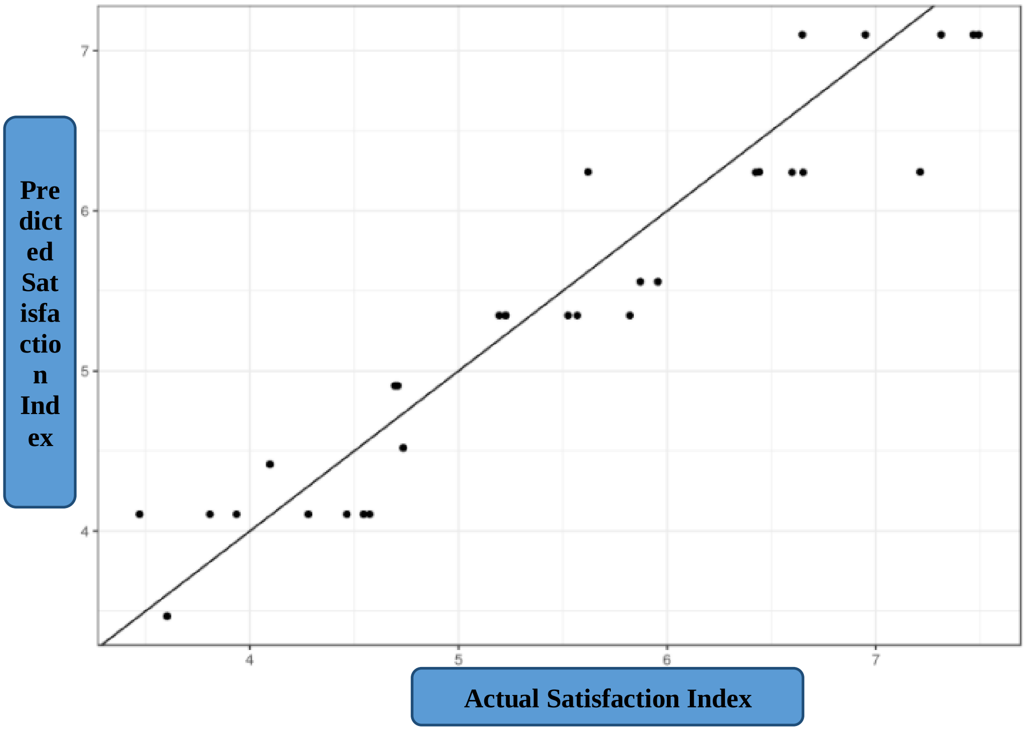
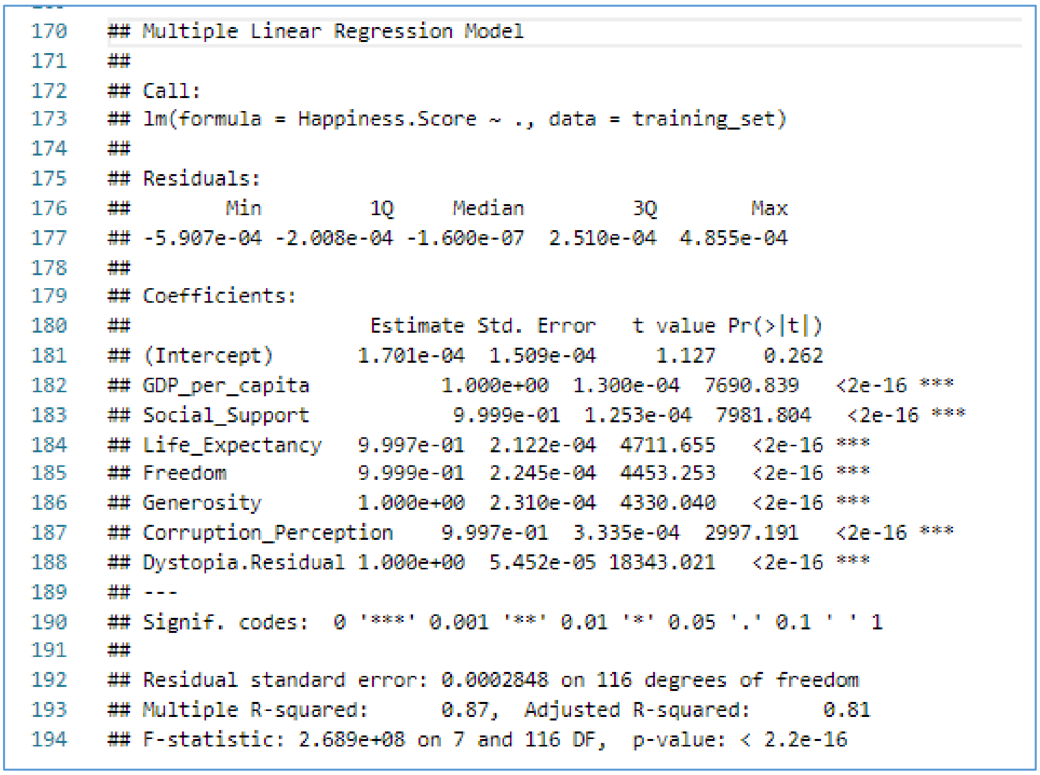


Figure 0.6: Actual vs Predicted Satisfaction Index (Decision Tree Regression Model)

4.7 Discussion of Results



```
## Multiple Linear Regression Model
##
## Call:
## lm(formula = Happiness.Score ~ ., data = training_set)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -5.907e-04 -2.008e-04 -1.600e-07  2.510e-04  4.855e-04
##
## Coefficients:
##                    Estimate Std. Error   t value Pr(>|t|)
## (Intercept)       1.701e-04  1.509e-04     1.127    0.262
## GDP_per_capita           1.000e+00  1.300e-04  7690.839   <2e-16 ***
## Social_Support            9.999e-01  1.253e-04  7981.804   <2e-16 ***
## Life_Expectancy   9.997e-01  2.122e-04  4711.655   <2e-16 ***
## Freedom           9.999e-01  2.245e-04  4453.253   <2e-16 ***
## Generosity        1.000e+00  2.310e-04  4330.040   <2e-16 ***
## Corruption_Perception    9.997e-01  3.335e-04  2997.191   <2e-16 ***
## Dystopia.Residual 1.000e+00  5.452e-05 18343.021   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0002848 on 116 degrees of freedom
## Multiple R-squared:      0.87,  Adjusted R-squared:      0.81
## F-statistic: 2.689e+08 on 7 and 116 DF,  p-value: < 2.2e-16
```

Figure 0.7: Model Results (Multiple Regression)

Metric	**Results**
Residual standard error	0.0002848 on 116 degrees of freedom
Multiple R-squared	0.87
Adjusted R-squared	0.81
F-statistics	2.689 on 7 and 116 degrees of freedom
P-value	< 2.2

Figure 0.8: Tabulated Model Results

In order to predict a tourist Satisfaction index, this study utilized a publically available dataset to train and evaluate the chosen model. Specifically, the dataset was selected because it was a representation of 155 countries national wide and it contained some significant socioeconomic data that had been used to predict Satisfaction index. With this, the study was determined to predict the Satisfaction index of a tourist who'd visit a nation based on the socioeconomic indicators.

The dataset underwent a rigorous pre-processing task to clean the data in preparation for model training. During this stage, data cleaning was employed to identify any missing or noisy data. Gladly, there were no missing or noisy values within the dataset. No data was transformed during the pre-processing stage that would result in additional or fewer features. The Scikit-Learn python library was the machine learning library chose for this endeavor. With it, the study accomplished its desired preprocessing, model building training and evaluation. The adopted Multiple Regression model was also available as a function within the Scikit Learn library. The resulting regression model performance in predicting the desired values was as follows.

The p-value in the multiple regression model was less than the significance level (usually 0.05) which meant that the model fits the data well. It also showed that there was a statistically significant association between the dependent and independent variables.This shows us that the model can indeed be accepted for prediction as a result of this yield observation.The adjusted R^2 are measures of goodness of fit of the model. The higher it is the better it fits to our model.The model had a high adjusted R-square of 0.81 and low p-value of 2.2e-16 which meant that the model explains a lot of variation within the data and was statistically significant (best scenario). The study used a normal probability plot of residuals to verify the assumption that the residuals werenormally distributed. The normal probability plot of the residuals should approximately follow a straight line.

Finally, yet importantly, supervised learning appears to be the dominant approach in making predictions (Ardito, Cerchione, Del Vecchio, & Raguseo, 2019). So much so that they employed two supervised learning models in forecasting the Satisfaction index of countries using a Random Forest Regression and a Support Vector Regression. These resulting models were able to predict the dependent variable with a mean R-squared score of 0.81. Although their findings could not be evaluated against other studies, they reported that their models could be used to advise individuals willing to tour the world on making informed decisions when picking their desired destinations.

Both these models in the highlighted research draw similar approaches to the Multiple Regression model adopted in this paper. A supervised learning approach was used to train a Multiple Regression model to predict a tourist Satisfaction index whenever one toured a country. Nonetheless, this study's model has some differences in the diversity of data features and dependent variable that was to be predicted. Ardito, Cerchione, Del Vecchio, & Raguseo, (2019) study did use some comon features to those in this study, GDP per capita, social support, generosity and corruption perception, but their dependent variable was on a nation's Satisfaction score. Nonetheless, this study was keen to look at other socioeconomic features that were seen to be attributes that would have an effect on tourism. Finally, yet importantlyArdito, Cerchione, Del Vecchio, & Raguseo, (2019) study forecast on predicting the Satisfaction index within a nation, whereas this study explored the novelity of predicting a tourist Satisfaction index after visiting a nation. In this regard, it seems apparent that machine learning can be used to advise tourists on the destinations that will yeild more value for money, a happy tour.

Esfahani, H., at el (2019) on their study on big data and social media stated that there were fewer studies that provided a comprehensive review on the literature associated with big data in social media. The study noted that globally researchers in USA, UK, China and India were among themost popular. This is similar to the limitation of this study saying that very little research regarding the use of modern technology, to help derive insight from big data.

4.8 Summary

This study had hypothesized that some socioeconomic features within a nation can be used to predict the expected Satisfaction index of tourists after they had visited a foreign nation on vacation. Also, the study sort to probe if these features can be used to predict the tourist Satisfaction index. With efforts to undertake this research, the study collected relevant datasets for mining and model training purposes. This datasetwas preprocessed and readied for training a Multiple Regression model that was able to predict a tourist Satisfaction index with a Residual standard error: 0.0002848 on 116 degrees of freedom, an R-squared and Adjusted R-squared of 0.87, and 0.81 respectively. However, the study did encounter some challenges and insights that will be detailed out in the next section.

CHAPTER FIVE: CONCLUSSION AND RECCOMENDATION

5.1 Introduction

This section will give a summary of the entire research endeavor. Consequently, it will make some recommendation based on what was learned from undertaking this study.

5.2 Conclusion

It was observed that the tourism sector remains an unexploited sector that can boost our country's economy. Despite effort from key stakeholders to boost the sector, there seems to be little adaptation of new technology that can advise the sector on howit can strategically achieve its full potential. With this, it seems feasible for the tourism sector to employ modern technology such as machine learning and AI in efforts to harness insight from the ever-growing Big Data. Despite the lack of adequate research on the use of such technology to boost the tourism sector, most of the reviewed literature tend to advise that the use of AI to mine insight appears to be a promising tool for the trade, promote and empower local and foreign tourism in Kenya.

A countries' values and culture are important given that the analysis gives us an image of what people in a particular country value the most and what policies they should come-up with so as to improve the living standards.This article should not be used as a conclusion for any country, because each of the countries has a huge cultural heritage, and that is what we should value first.

5.3 Contributions of the Study

This study did an analysis on the attributes that determined a higher tourism satisfaction index. For this reason countries are able to take into account the different factors that will increase the score and help improve the living standards of its citizens. This in return has a huge positive impact on the hospitality industry given that it goes hand to hand with the tourism industry. An increase in opportunities and revenue in the tourism industry will also led to an increase in the hospitality industry.

The tourism sector seems to be growing in a slow pace. This study attributed part of this slow growth to the little adaptation of modern technology such as machine learning in mining the abundance of big data to advise the tourism sector on its strategic planning. Shirishjebel at el, (2016), agreed with this by concluded that we can't afford to ignore the use of predictive analytics and big data in providing insightful knowledge that can further assist in addressing the various research gaps in the tourism industry. With the use of data from different sources, predictive analytics could bring new correlations and patterns that were previously unknown. Maximizing value of the data is an executive decisions that comes as an advantage of understanding these correlations.

This study has identified that readily available data, under the umbrella of big data, can be mined for insight. Specifically, this study demonstrated that a Multiple Regression model can be trained to advise tourism of destinations that are likely to result into a satisfactory vacationing endeavor.

This research identified the different attributes that determined a tourism satisfaction index. With the help of the attributes, the study will help policy makers to create policies that will impact the local tourism industry positively. The main aim of this policies will be to have a growth in the local tourism in the country which will have an impact on the economy.The result may encourage the government on the importance of developing tourism destinations since both the local citizen and the government will benefit. It will create job opportunities to the citizen which will increase the economy. An increase in the number of the local tourists will also be seen in the tourism destination areas.

Finally, yet importantly, the study has added to the existing body of knowledge in predictive analytics.This will help future researchersto have an understanding on what has been studied first and enhance it further.

5.4 Limitation of the Study

The undertaking of this research work was not without its hurdles and these limitations are addressed below. Despite the abundance of big data which is readily available, there seems to be little data shared by organizations or bodies within the tourism sector. This limits researchers from conducting extensive studies of such information or data. Conversely, the lack of relevant data impedes the amount of research conducted, which happens to be another limitation observed by this study. There was very little research regarding the use of modern technology, data mining, to derive insight from big data that would otherwise advice on strategic planning for the tourism sector. Also, the other challenge observed was the limited financial resources available for this study. The limited budget meant restricted the extent to which resources would be used to better this study. Finally, yet importantly, the study was limited by the recent global pandemic, Covid-19, which ultimately affected the time available to extent the scope of this study.

Finally, due to the unique and diverse nature of the tourism sector, further research on other suitable predictive modelling techniques can be done, so as to increase the options for tourismpredictions. Despite these limitations, this study was carried out which utilized all the available resources in efforts to be thorough and knowledgeable.

5.5 Recommendations for Future Research

During the course of this research work, the study identified that there were few studies that focused on the harnessing big data insights to advise or promote the tourism sector. As such, this study wishes to recommend that more research be carried out on the potential use of such technology in respect to boosting the tourism sector.

Also, there seems to be very little data shared by organizations that otherwise would have been used for research work. This is seen by the adaptation of publically available datasets in this research as opposed to using data specifically generated or availed from local, Kenya, institutions or bodies. As such, this study recommends that additional data can be availed and used to train the model in efforts to make it more practical for a real world scenario.

Different countries have different cultural heritage, which is valued first by its citizen. This study recommends an additional attribute that takes into account the different indigenous cultures that different countries possess. This will bring out the uniqueness of different cultures.

With the ever changing 21st century society, new approaches are required to measure progress and well-being that contributes to a higher satisfaction index. Due to this there is a need for future researchers to understanding the complex attributes that really makes one satisfied. This will indeed need further research.

REFERENCES

Alcantara-Pilar, M. S., Crespo, A., del Barrio-Garcia, E., & Porcu, L. (2017). Toward understanding of online information processing in e-tourism: Does national culture matter? *Journal of Travel & Tourism Marketing*, 1128–1142.

Ardito, L., Cerchione, R., Del Vecchio, P., & Raguseo, E. (2019). Big data in smart tourism: challenges, issues and opportunities. Current Issues in Tourism. 1805–1809.

Bangwayo-Skeete, F., & Skeete, W. (2015). Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tourism Management*, 454–464.

Buhalis, D., & Amaranggana, A. (2015). Smart tourism destinations enhancing tourism experience through personalisation of services. *Information and Communication Technologies in Tourism*, 377-389.

Burke, R. J., & Christensen, L. B. (2019). *Educational Research Quantitative, Qualitative, and Mixed Approaches.* London: Sage.

Cillo, V., Rialti, R., USai, A., & del Giudice, M. (2019). Niche tourism destinations' online reputation management and competitiveness in big data era: Evidence from three Italian cases. *Current Issues in Tourism*, 2-23.

Creswell, J., & Creswell, D. (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches.* London: SAGE Publications.

Esfahani, H., Tavasoli, K., & Jabbarzadeh, A. (2019). Big data and social media: A scientometrics analysis. *International Journal of Data and Network Science*, 145-164.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning (Adaptive Computation and Machine Learning).* London: The MIT Press .

Gunter, U., & Onder, I. (2015). Forecasting international city tourism demand for Paris: Accuracy of uni- and multivariate models employing monthly data. *Tourism Management*, 23-135.

Gwalik, E., Kabaria, H., & Kaur, S. (2015). Predicting tourism trends with Google Insights. *Tourism Management*, 5-11.

Kenya Tourism Board. (2018). *Tourism Sector Perfomance Report 2018.* Nairobi: Tourism Research Institute.

Maingi, S. W. (2019). Sustainable tourism certification, local governance and management in dealing with overtourism in East Africa. *Worldwide Hospitality and Tourism Themes*, 3-17.

Mitchell, M., & Jolley, J. (2015). *Research Design Explained.* Liverpool: Cengage Learning.

Mohammad, R., Gholipour, H., Feizi, M., & Nunkoo, R. (2020). International Tourism and Outbreak of Coronavirus (COVID-19): A Cross-Country Analysis. *Sage Journals*, 1-10.

Ndivo, R., Waudo, N., & Waswa, F. (2016). Examining Kenya's Tourist Destinations' Appeal: the Perspectives of Domestic Tourism Market. *Tourism Planning & Development*, 17-30.

Onder, I., & Gunter, U. (2015). Forecasting tourism demand wih Google Trends for a major European City Destination. *Modul University*, 50-62.

Park, B., Ok, M., & Chae, K. (2016). Using twitter data for cruise tourism marketing and research. *Journal of Travel & Tourism Marketing*, 85–98.

Procter, R., Crump, J., Karstedt, S., & Voss, A. (2017). Reading the riots: What were the police doing on Twitter? *In Policing Cybercrime*, 5-28.

Ram, S., Zhang, W., & Williams, M. (2015). Predicting asthma-related emergency department visits using big data. *IEEE J. Biomedical and Health Informatics*, 1216–1223.

Sharma, S., Tim, S., & Wong, J. (2016). A brief review on leading big data models. *Data Science Journal*, 138-157.

Sheoran, S. (2017). Big data: A big boon for tourism sector. *International Journal of Research in Advanced Engineering and Technology*, 10-13.

Smeral, E. (2015). Forecasting the City Hotel Market. *Tourism Analysis*, 339–349.

Song, H., & Lui, H. (2017). *Predicting Tourist Demand Using Big Data: Analytics in Smart Tourism Design.* Switzerland: Springer International Publishing.

Tifekci, Z. (2017). Engineering the public: Big data, surveillance and computational politics. *First Monday*, 3-17.

Watson, J. (2015). Tutorial: Big data analytics: Concepts, technologies, and applications. *Communications of the Association for Information Systems*, 34-65.

Williams, L., & Burnap, P. (2017). Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patterns. *The British Journal of Criminology*, 320-340.

Wood, S., Guerry, A., & Silver, J. (2015). Using social media to quantify nature-based tourism and recreation. *Scientific Reports*, 3-24.

Wu, K., Liao, L., Tseng, L., & Lim, K. (2017). Toward sustainability: using big data to explore the decisive attributes of supply chain risks and uncertainties. *Journal of Cleaner Production*, 663-676.

Xiang, Z., Schwartz, Z., Gerdes, J., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management*, 121-130.

Yang, M., Shang, W., & Kiang, M. (2015). Filtering big data from social-media– Building an early warning system for adverse drug reactions. *ournal of Biomedical Informatics*, 33-41.

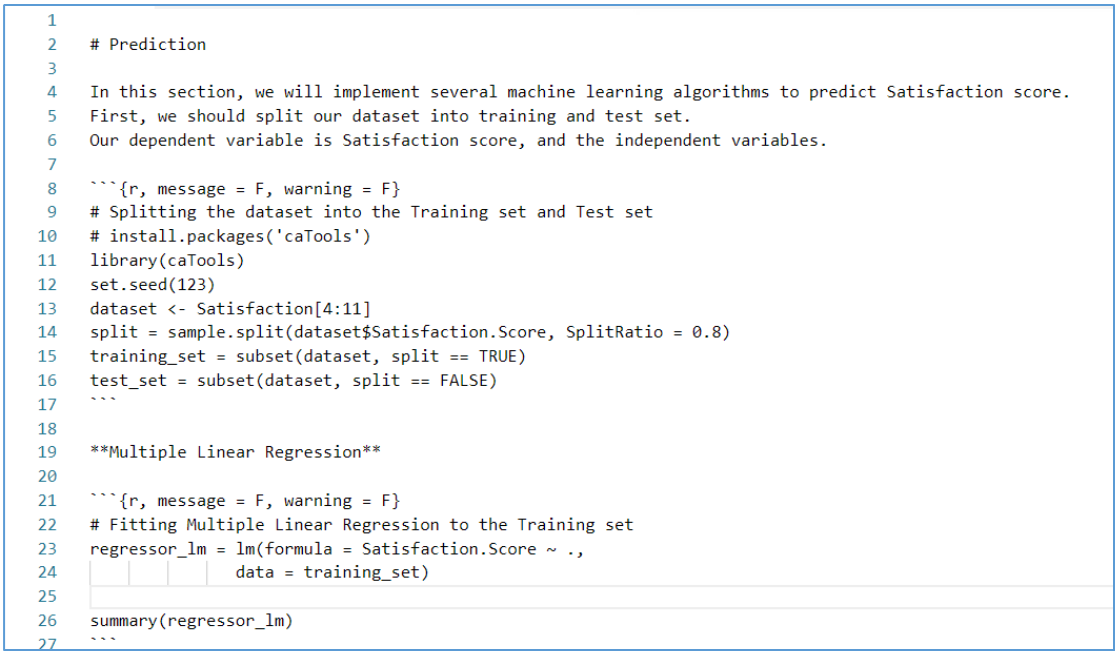
Yang, X., Pan, B., & Evans, A. (2017). Forecasting of Chinese tourist volume with search engine data. *The Tourism Management*, 77-85.

Yaqoob, I., Hasheem, T., Gani, A., Mokhatar, S., & Ahmed, E. (2016). Big data: From beginning to future. *International Journal in Information Management.*, 31-47.

Zhu, W., Cui, P., & Wang, Z. (2015). Multi media big data computing. *IEEE Multimedia*, 3-18.

APPENDIX

7.1 Python Extracts



Prediction

In this section, we will implement several machine learning algorithms to predict Satisfaction score.
First, we should split our dataset into training and test set.
Our dependent variable is Satisfaction score, and the independent variables.

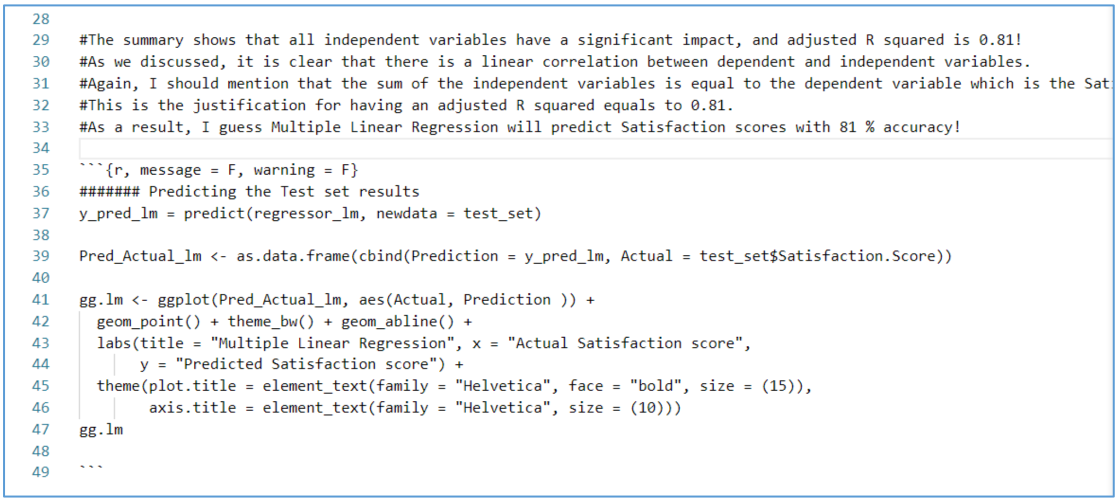
```{r, message = F, warning = F}
# Splitting the dataset into the Training set and Test set
# install.packages('caTools')
library(caTools)
set.seed(123)
dataset <- Satisfaction[4:11]
split = sample.split(dataset$Satisfaction.Score, SplitRatio = 0.8)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)
```

Multiple Linear Regression

```{r, message = F, warning = F}
# Fitting Multiple Linear Regression to the Training set
regressor_lm = lm(formula = Satisfaction.Score ~ .,
               data = training_set)

summary(regressor_lm)
```
```

**Figure 0.1.1: Multiple Regression Model (Python extract)**



```

#The summary shows that all independent variables have a significant impact, and adjusted R squared is 0.81!
#As we discussed, it is clear that there is a linear correlation between dependent and independent variables.
#Again, I should mention that the sum of the independent variables is equal to the dependent variable which is the Sat
#This is the justification for having an adjusted R squared equals to 0.81.
#As a result, I guess Multiple Linear Regression will predict Satisfaction scores with 81 % accuracy!

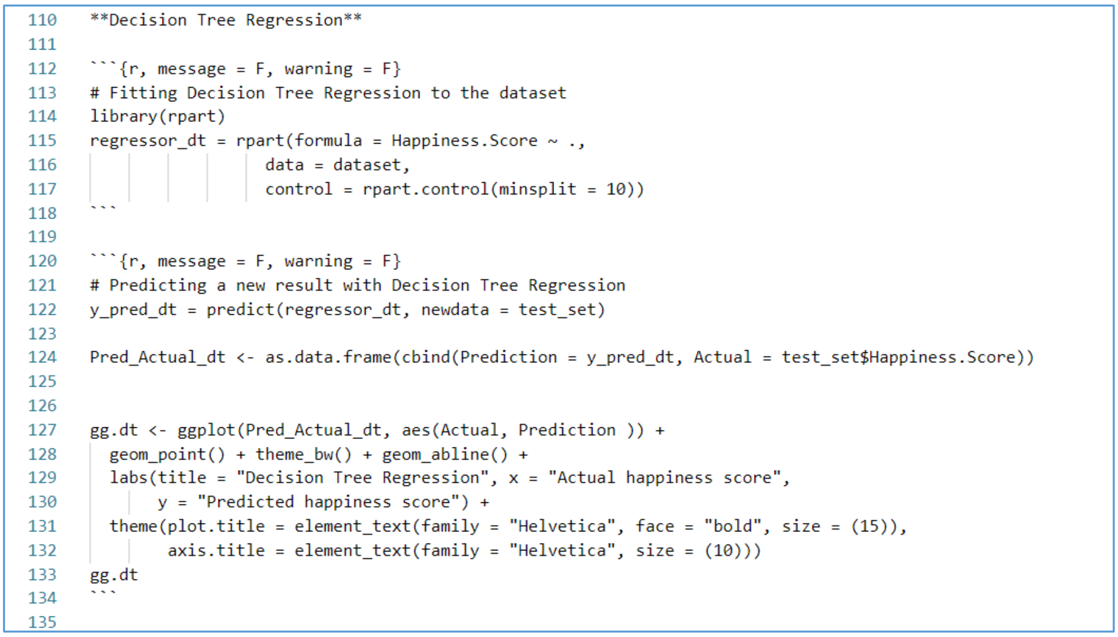
```{r, message = F, warning = F}
####### Predicting the Test set results
y_pred_lm = predict(regressor_lm, newdata = test_set)

Pred_Actual_lm <- as.data.frame(cbind(Prediction = y_pred_lm, Actual = test_set$Satisfaction.Score))

gg.lm <- ggplot(Pred_Actual_lm, aes(Actual, Prediction )) +
  geom_point() + theme_bw() + geom_abline() +
  labs(title = "Multiple Linear Regression", x = "Actual Satisfaction score",
       y = "Predicted Satisfaction score") +
  theme(plot.title = element_text(family = "Helvetica", face = "bold", size = (15)),
        axis.title = element_text(family = "Helvetica", size = (10)))
gg.lm

```
```

Figure 0.1.2: Multiple Regression Model (Python Extract 2)
```



~~~
**Decision Tree Regression**

```{r, message = F, warning = F}
# Fitting Decision Tree Regression to the dataset
library(rpart)
regressor_dt = rpart(formula = Happiness.Score ~ .,
                  data = dataset,
                  control = rpart.control(minsplit = 10))
~~~

```{r, message = F, warning = F}
# Predicting a new result with Decision Tree Regression
y_pred_dt = predict(regressor_dt, newdata = test_set)

Pred_Actual_dt <- as.data.frame(cbind(Prediction = y_pred_dt, Actual = test_set$Happiness.Score))


gg.dt <- ggplot(Pred_Actual_dt, aes(Actual, Prediction )) +
  geom_point() + theme_bw() + geom_abline() +
  labs(title = "Decision Tree Regression", x = "Actual happiness score",
       y = "Predicted happiness score") +
  theme(plot.title = element_text(family = "Helvetica", face = "bold", size = (15)),
        axis.title = element_text(family = "Helvetica", size = (10)))
gg.dt
```

```

**Figure 0.1.3: Decision Tree Regression Model (Python Extract)**

## 7.2 Research Budget

| **Item** | **Estimated Cost (Kshs)** |
|---|---|
| Computer expenses | 45,000 |
| Data Research, Collection and Analysis | 30,000 |
| Stationery, printing | 25,000 |
| Internet and airtime | 10, 000 |
| Hard Cover Binding | 4,000 |
| Travel expense | 15,000 |
| Reproduction Copies | 1,000 |
| Miscellaneous | 20, 000 |
| Total | 150,000 |
```

7.3 Work Plan

Task	Start Date	End Date	Duration (Days)
Research concept, supervisor assignment	01/07/2019	14/09/2019	14
Chapter one	15/08/2019	14/09/2019	30
Chapter two	15/09/2019	14/10/2019	30
Chapter three	15/10/2019	30/10/2019	15
Get final sign-off from supervisor and prepare and present proposal document to school of graduate studies and research	01/11/2019	15/11/2019	15
Proposal presentation	16/11/2019	18/11/2019	3
Correction on proposal presentation	19/11/2019	20/12/2019	30
Pre-test data collection instrument	01/02/2020	14/02/2020	14
Finalize sampling plan	15/02/2020	28/02/2020	14
Dataset identification	01/03/2020	15/03/2020	15
Carry out data collection	16/03/2020	31/03/2020	15
Preprocess data for analysis	01/04/2020	15/04/2020	14
Analyze data	16/05/2020	16/06/2020	30

Draw conclusions/ recommendations	17/06/2020	07/07/2020	21
Prepare final draft (chapter four and five) of report	08/07/2020	09/08/2020	31
Review draft with supervisor	10/08/2020	11/09/2020	30
Get final sign-off from supervisor and prepare and present report document to school of graduate studies and research	12/09/2020	20/09/2020	8
Final defense presentation	21/09/2020	24/09/2020	3
Final editing	25/09/2020	02/10/2020	6
Printing, binding and final submission	03/10/2020	09/10/2020	6