**A MODEL FOR PREDICTING SALES IN A SUPERMARKET**

KORIR PHYLIS JEPCHUMBA

SC212/1256/2017

**A report submitted in partial fulfillment of the requirements for the award of Bachelor’s degree in Software Engineering at the department of Computer Science ,School of Computing and Information Technology, Murang’a University of Technology**

**2021**

## **DECLARATION**

This project is my original work and it has not been presented before to the school of computer science and information technology for the award of bachelor’s degree in software engineering of Murang’a University of technology. No part of this report shall be duplicated without my prior consent.

…………………………… ……………………………..

SIGN DATE

NAME……………………………………

REGISTRATION NUMBER………………………………

SUPERVISOR

……………………………. ………………………….

PETER MWANGI DATE

Department of Computer Science school of Computing and Information Technology

Murang’a University of Technology

## **DEDICATION**

I am a complacement to my friends, lecturers, family members for their support, whether informational, financial, educational and physical or in any way.

This report courtesy of mentioned role players and I would love to dedicate my findings, experience and achievements to them

## **ACKNOWLEDGEMENT**

This would have not been successful without cooperation and support from a number of people who gave me a total support.

First, I would like thank almighty God for the charitable time; good healthy, continuous grace and strength that enabled me complete my research.

Secondly, my gratitude goes to my supervisor for valuable guidance he gave me and assessing my progress during my research.

Finally, I would like to thank my loving parents for their support.

## **ABSTRACT**

Sales forecasting is an important field in supermarkets, and it has recently got immense popularity to boost market operations and productivity due to new technologies. The industry has focused on conventional statistical model but in recent years, Machine learning techniques have received more attention.

The use of traditional statistical method to forecast supermarket sales has left a lot of challenges unaddressed and mostly result in the creation of predictive models that perform poorly.

The era of big data coupled with access to massive compute power has made machine learning a goto for sales forecast.

The objective of this project is to develop a model for predicting sales in supermarkets keeping in view sales and the amount used to advertise.

Using regression analysis product variables such as supermarket type, product price and supermarket opening year are used to predict the sales.

## **TABLE OF CONTENTS**

[**DECLARATION** ii](#_Toc83630553)

[**DEDICATION** iii](#_Toc83630554)

[**ACKNOWLEDGEMENT** iv](#_Toc83630555)

[**ABSTRACT** v](#_Toc83630556)

[**TABLE OF CONTENTS** vi](#_Toc83630557)

[**LIST OF FIGURES** viii](#_Toc83630558)

[**LIST OF TABLES** ix](#_Toc83630559)

[**ACRONYMS AND ABREVIATIONS** x](#_Toc83630560)

[**CHAPTER 1: INTRODUCTION** 11](#_Toc83630561)

[**1.1: BACKGROUND INFORMATION** 11](#_Toc83630562)

[**1.2: PROBLEM STATEMENT** 3](#_Toc83630563)

[**1.3: OBJECTIVES** 3](#_Toc83630564)

[**1.3.1: General Objectives** 3](#_Toc83630565)

[**1.3.2: Specific Objectives** 4](#_Toc83630566)

[**1.4: Significance of the study** 4](#_Toc83630567)

[**1.5: Scope of the study** 4](#_Toc83630568)

[**1.6: Limitations** 4](#_Toc83630569)

[**CHAPTER 2: LITERATURE REVIEW** 5](#_Toc83630570)

[**2.1: INTRODUCTION** 5](#_Toc83630571)

[**2.2: EXISTING SYSTEMS** 5](#_Toc83630572)

[**2.2.1: Time series forecasting using Artificial Neural Networks Methodologies** 5](#_Toc83630573)

[**2.2.2:** **Time series sales forecasting for short shelf-life food products based on ANN and evolutionary computing** 7](#_Toc83630574)

[**2.2.3: A survey of machine learning techniques for food sales prediction** 9](#_Toc83630575)

[**2.2.4: Sales prediction for a pharmaceutical distribution company: A data mining based approach** 11](#_Toc83630576)

[**2.2.5: Proposed System** 13](#_Toc83630577)

[**2.3: Existing software design and development tools** 13](#_Toc83630578)

[**2.3.1: Python Programming Language** 13](#_Toc83630579)

[**2.4: Justification** 14](#_Toc83630580)

[**2.5: Conclusion** 15](#_Toc83630581)

[**CHAPTER 3: RESEARCH METHODOLOGY** 15](#_Toc83630582)

[**3.1: Introduction** 15](#_Toc83630583)

[**3.2: Data Collection Techniques** 15](#_Toc83630584)

[**3.2.1: Interview** 15](#_Toc83630585)

[**3.2.2: Questionnaires** 17](#_Toc83630586)

[**3.2.3: Observation** 17](#_Toc83630587)

[**3.2.4: Documents and records** 18](#_Toc83630588)

[**3.2.5: Justification** 18](#_Toc83630589)

[**3.3: Software Development Techniques** 18](#_Toc83630590)

[**3.3.1: Waterfall Methodology** 19](#_Toc83630591)

[**3.3.2: Rapid Application Development Methodology** 19](#_Toc83630592)

[**3.3.3: Agile Methodology** 21](#_Toc83630593)

[**3.3.4: Justification** 22](#_Toc83630594)

[**3.4: System Requirements** 23](#_Toc83630595)

[**3.4.1: Software Requirements** 23](#_Toc83630596)

[**3.4.2: Hardware Requirements** 23](#_Toc83630597)

[**3.4.3: Functional Requirements** 23](#_Toc83630598)

[**3.4.4: Non-Functional Requirements** 23](#_Toc83630599)

[**3.5: Conclusion** 24](#_Toc83630600)

[**Chapter 4: System design, Implementation and Testing.** 25](#_Toc83630601)

[**4.1: Introduction.** 25](#_Toc83630602)

[**4.2: System design.** 25](#_Toc83630603)

[**4.2.1: Logical Design** 25](#_Toc83630604)

[**4.2.2: User Interface Design.** 26](#_Toc83630605)

[**4.2.3: Data Design.** 27](#_Toc83630606)

[**4.2.4: Process Design.** 28](#_Toc83630607)

[4.3: Implementation Approaches 32](#_Toc83630608)

[4.3.1: Multiple Linear Regression Algorithm. 32](#_Toc83630609)

[**4.3.2: Flask framework.** 33](#_Toc83630610)

[**4.4 Coding Details and Code Efficiency.** 33](#_Toc83630611)

[**4.5: Testing Approach** 38](#_Toc83630612)

[**4.6. Modifications and Improvements.** 40](#_Toc83630613)

[**Chapter 5** 41](#_Toc83630614)

[**5.1. Test Reports** 41](#_Toc83630615)

[**5.2: User Documentation** 46](#_Toc83630616)

[**Chapter 6: Conclusions and Future Works.** 49](#_Toc83630617)

[**6.1. Conclusion** 49](#_Toc83630618)

[**6.2: Future Works** 49](#_Toc83630619)

[**APP 1: Budget** 50](#_Toc83630620)

[**APP2: Schedule** 50](#_Toc83630621)

[References 51](#_Toc83630622)

## **LIST OF FIGURES**

[Figure 1: process to develop the model 12](#_Toc83585177)

[Figure 2: Waterfall model 19](#_Toc83585178)

[Figure 3: Rapid application Methodology 20](#_Toc83585179)

[Figure 4: Agile methodology 21](#_Toc83585180)

[Figure 5 User Interface Design 26](#_Toc83585181)

[Figure 6 Data Design 27](#_Toc83585182)

[Figure 7 Process Design 27](#_Toc83585183)

[Figure 8 Functional Testing 38](#_Toc83585184)

[Figure 9 Dataset 40](#_Toc83585185)

[Figure 10 Outlier Analysis 41](#_Toc83585186)

[Figure 11 Model Deployment 44](#_Toc83585187)

## **LIST OF TABLES**

[Table 1: Budget 39](#_Toc74765581)

[Table 2: Schedule 39](#_Toc74765582)

## **ACRONYMS AND ABREVIATIONS**

ANN- Artificial Neural Network

ERNN-STNN- Elman Recurrent Neural Network, Stochastic Time Neural Network

BPNN- Back-Propagation Neural Network

SVR-Support Vector Regression

SVM- Support Vector Machines

PDC- Pharmaceutical Distribution Company

RAD-Rapid Application Development

## 

## **CHAPTER 1: INTRODUCTION**

## **1.1: BACKGROUND INFORMATION**

Sales prediction is an estimation of sales volume that a company can expect to attain within the plan period based on historical data and industry trends [1]. It’s also the determination of a firms share in the market under a specified future.

Earlier companies used to produce goods without considering the number of sales and demand. For any manufacturer to determine whether to increase or decrease the production of several units, data regarding the demand for products on the market is required. Therefore the companies used to face losses while competing in the market since they don’t know how much to sell.

Managers used to make sales predictions randomly. Professional managers however, become hard to find and not always available.

In today’s highly competitive and ever changing consumer landscape, accurate and timely forecasting of future revenue or sales can offer a valuable insight to companies engaged in manufacture and distribution of retail goods. Short tern forecasts help with production planning and stock management while long term forecasts can deal with business growth and decision making.

Sales prediction can be assisted by computer systems to play the qualified managers role when they are not available. One way of implementing such a method is to try and model professional manager’s skill inside a computer program for a company to gain better results for the progress of current society.

In this project, we propose a predictive model using linear regression technique for predicting sales in a supermarket. The major of this machine learning project is to build a predictive model and also search out sales of each of the products at a particular selected supermarket. Using machine learning model, supermarket prediction tries to understand the properties of products and stores which play a key role in increasing sales of products. Python is used as a programming language and Jupyter Notebook is used as tools. To build this application, regression task aspect is used to predict sales of a given store in the future

Various processes used are; Data Preprocessing, Feature engineering, creating model, Evaluation and supervised learning helps understand the flow of data and knowing sales prices.

Regression task includes data visualization, cleaning and transformation. Linear regression algorithm will be used in the proposed system

The approach of using machine learning to predict sales is accurate, simpler and flexible. Linear regression model is important in that it can be used to understand all kinds of patterns that occur in data.

The aim of developing a sales prediction system is to enable companies efficiently allocate resources for future growth and manage cash flow. Also to help businesses to estimate their cost and revenue accurately based on which they are able to predict their short-term and long-term performance. The motivation for this project lies within a natural passion for market research

## **1.2: PROBLEM STATEMENT**

The problem proposed in this project is sales prediction, where information about the items sold and stores in which those items are exhibited will be used to predict sales that items would make when sold in new stores.

Regression is an important machine learning model for this kind of a problem. Predicting sales of a company needs time series data of the company and based on that data the model can predict future sales of a supermarket or product.

For this kind of project of sales prediction, linear regression will be applied to evaluate the result based on training, testing and validation set of the data. The main aim of linear regression is to find the best fit line to target variable and independent variables of the data.

According to Grigorios tsoumakas [2] they used machine learning techniques to perform a survey on forecasting food sales. They addressed data analyst design decisions such as output variable and input variable in their survey. The authors experimented by taking point of sale as internal data and even external data by considering different environments to enhance efficiency of demand forecasting. They used algorithms such as boosted decision tree regression and Bayesian linear regression.

Most of the recent studies focused on sales modeling without considering the relationship between training and test data, they used training data directly. This causes many errors which led to a reduction in accuracy.

Clustering techniques have been suggested to separate the entire forecasting data into several clusters of predictable data before designing predictable models to minimize computational time and achieve effective evaluating performance.

## **1.3: OBJECTIVES**

## **1.3.1: General Objectives**

To develop a model that can predict sales of products from different supermarkets based on amount used to advertise the items.

## **1.3.2: Specific Objectives**

* To gather and analyze existing sales prediction system
* To design the proposed sales prediction system
* To implement the developed sales prediction system
* To test and validate the newly developed sales prediction system

## **1.4: Significance of the study**

The proposed system aims to help supermarkets identify benchmarks and determine incremental impacts of new initiatives, plan resources in response to expected demand and project failure budgets

## **1.5: Scope of the study**

The project aims at providing an efficient prediction system to the supermarkets for managing their inventory. The system analyzes the sales, compares it with the past sales and predict future sales.

The proposed system uses linear regression model of machine learning to make predictions of sales in supermarket using python programming language

## **1.6: Limitations**

A sales history or past records are essential for a sound forecast plan. If past data are not available, then the forecast is made on a guess work without a base and this may lead to failure.

Since customer’s attitude may change at any time, the forecast may not be able to predict exactly the behavior of customers

## 

## **CHAPTER 2: LITERATURE REVIEW**

## **2.1: INTRODUCTION**

Literature review is a survey of scholarly sources on a specific topic that provides an overview of current knowledge, allowing you to identify relevant theories, methods and gaps in existing research [3].

Due to importance of forecasting in many fields, many prominent approaches have been developed. Statistical methods, machine learning methods and hybrid models have been practiced.

### **2.2: EXISTING SYSTEMS**

## **2.2.1: Time series forecasting using Artificial Neural Networks Methodologies**

Time series is a general problem of great practical interest in many disciplines since it allows you discover with some margin error, future of values of series from its past values.

The project studied advances in time series forecasting models using artificial neural networks methodologies in a systematic literature review using manual search of published papers. Also it applied the research methodology LSR in context of software engineering. The methodology promotes use of systematic strategy for defining the research questions, declaring the search strategy, identifying primary studies, data synthesizing and data analysis [4].

The objective of this LSR is to identify the most important theoretical contributions in development of neural network models for forecasting non-linear time series performed in the period between 2006 and 2016 and also identify new research problems originated from published proposal.

The search process consisted of manual search of articles published in journals serials using the largest bibliographic system called SCOPUS which includes one of the largest collections of abstracts, bibliographic references and indexes. Two criteria were used; the first criterion was non-linear neural model for forecasting and the second on were neural networks and non-linear time series modeling using the search string.

Although there are very high numbers of publications on ANN, there are few studies that propose new models with an appropriate theoretical support. According to Ahmed Teelab, several quality criteria were used to analyze the best ANN models that can be used in forecasting.

The research project proposed the following models for ANN;

ERNN-STNN- a model based on Elman recurrent networks and stochastic time effective. The empirical results show that proposed neural network displays the best performance between linear regression, complexity invariant distance, multi scale complexity invariant distance compared to back propagation neural network in financial time series forecasting [5].

Application of novel neural network technique in financial time series forecasting, support vector machine SVM to examine the feasibility of SVM in financial time series forecasting and proposed that SVMs machines achieve an optimum network structure by implementing the structural risk minimization principle which seeks to minimize an upper bound of generalization error rather than minimize the training error. SVMs have also extended to solve non- linear regression estimation problems [6].

They also made an attempt with ensembles aiming for the improvement of prediction performance and recognized ensembles as one of the most ambitious forms for solving predictive tasks and conventional in reducing the variance and bias components of forecasting seeeror by taking advantage of diversity and amid models. They compared bagging and ARIMA and positive results are achieved showing that the approach can be used as an alternative for forecasting time series.

Financial time series forecasting is inevitably a center point for the practitioner for its available data and for its profitability [7]. Ensemble algorithms are substantial in improvising performances of base learners in financial time series forecasting. The research was experimented using SVR support vector regression, BPNN back-propagation neural network, RBFNN radial basis function neural network, bagging for comparison and evaluation research.

The authors also experimented financial time series forecasting by using intelligent hybrid models to overcome the issue of capturing the non-stationary property and identify the accurate movements. Empirical mode description and support vector regression are used to evaluate performance

Advantages of time series forecasting using neural networks

* Neural networks have the advantage that can approximate non-linear functions
* Time series analysis allows you analyze major patterns such as trends, seasonality, cyclicity and irregularity.
* Neural networks are data driven

Disadvantages

* It was observed that original pattern of time series of the index is not stationary

## **2.2.2:** **Time series sales forecasting for short shelf-life food products based on ANN and evolutionary computing**

In retail food industry, the main cause of wasted products and stock outs is the inaccuracy of sales forecasting leading to incorrect orders. More specifically in fresh food industry, including refrigerated ones such as dairy, fruit and juice segments and the need to maintain quality in storage and distribution process makes sales forecasting accuracy and important factor for planning and minimizing wastage.

They presented a framework that can be used to develop non-linear time series sales forecasting models comprising two artificial intelligence technologies namely radial basis function neural network and a specially designed genetic algorithm. The methodology was applied successfully to sales data of fresh milk provided by a major company of dairy products [8].

Hybrid system of non-linear methods; genetic algorithm for variable selection and adaptive radial basis function (RBF) artificial network were used to model the relationship between variables and sales volume. To integrate linear and non-linear models they used ARMA for linear auto regression and neural network for modeling of forecasting moving average errors.

RBF networks are non-linear modeling structures that unveil the mathematical relationships between the hidden node and output node. RBF has a special structure that has a certain advantages including faster training algorithms and more successful capabilities.

Genetic algorithms are machine learning procedures which derive their behavior from the process of evolution in nature and are used to solve complicated optimization problems.

The combined GA-RBF method was applied on sales data of fresh milk. It selects appropriate factors that are going to be used as inputs to the models.

They obtained the following results; the problem under study is evaluation of forecasting performance of the GA-RBF methodology on the daily sales of fresh milk in area of Athens, Greece and more specifically on 11 pack. Daily sales data of 11 pack for the first few months of the year were provided by leading manufacturer of dairy products. Effect national holidays have on sales were analyzed and arranged.

Past sales data were also utilized in order to exploit information they contain. Past sales data from current year contain the changes that have meanwhile occurred in the market and have affected the level and trend of sales.

The change in trend could be fed into a model by providing it with percentile change in sales between the current year and previous year.

**Advantages of using the model**

* Accuracy in fresh food forecasting improves efficiency of order and inventory management enabling retailers reduce their disposal by about 40%
* Disposal detoriation is avoided.
* Minimizes lost sales due to lack of products, reducing returns due to proximity of expiration dates.

**Disadvantages of the model**

* GA-RBF utilizes only historical data therefore does not show how additional information like price, promotions can be explicitly taken into account in development of the time series model.
* The type of non-linearity is not known in advance hence the model produces about 28.2% of errors.
* For time series forecasting to be carried out historical data for a long time period is needed to capture seasonality. In this case when a new product is launched, maybe a perishable good and they have a time series for a similar product they may assume that the new product will have a similar sales pattern.

## **2.2.3: A survey of machine learning techniques for food sales prediction**

Food sales prediction is concerned with estimating future sales of companies in the food industry, such as supermarkets, groceries, restaurants, bakeries and patisseries. Accurate short-term sales prediction allows companies to minimize stocked and expired products inside the stores at the same time.

This survey reviewed existing machine learning approaches for food sales prediction. They discussed important design decisions of a data analyst working on food sales prediction, such as temporal granularity of sales data, input variables to use for predicting sales and the representation of sales output variable [2].

It reviews machine learning algorithms that have been applied to food sales prediction and appropriate measures for evaluating accuracy. And also discusses the challenges and opportunities for applied machine learning in the domain of food sales prediction.

The author experimented by taking point of sale as internal data and even external data by considering different environments to enhance the efficiency of demand forecasting. They considered different machine learning algorithms such as Boosted Decision Tree Regression, Bayesian Linear Regression and Decision Forest Regression for evaluation.

The author had also researched interestingly about customers coming to the restaurants using Random Forests, k-nearest neighbor and XGBoost. They chose two real world data sets from different booking sites and also made different input variables from restaurant features. They found the XGBoost is most appropriate for dataset.

They had observed that regular restaurants sales are influenced by weather. They considered two algorithms; XGBoost and neural network and the results showed that XGBoost is more accurate and the performance of their system improved. To improve accuracy, they had considered numerous variables such as date characteristics, sales history and weather factors [9].

However the study focused on sales without considering the relationship between the training and testing data. They used training data directly hence causing many errors which led to reduction in accuracy. Recent studies suggest clustering techniques to separate entire data into several clusters of predictable data before assigning predictable models to minimize computational time ach achieve effective evaluating performance.

## **2.2.4: Sales prediction for a pharmaceutical distribution company: A data mining based approach**

For pharmaceutical distribution companies it is essential to obtain good estimates of medicine needs, due to short shelf life of many medicines and the need to control stock levels so as to avoid excessive inventory costs while guaranteeing customer demand satisfaction and thus decreasing the possibility of loss of customers due to stock shortage.

They explored the use of time series data mining technique for sales prediction of individual products of pharmaceutical distribution company in Portugal [10].

Through data mining techniques, the historical data of product sales are analyzed to detect patterns to make prediction based on the experience contained in the data.

The results they obtained with the technique as well with proposed method suggested that the performed modeling maybe considered appropriate for the short term product sales prediction.

They examined the role of data prescription and pharmacies sales mining in pharmaceutical industry and various type of techniques that be used.

They found that most Pharmaceutical distribution companies (PDC) in Portugal still use heuristic or simple statistical models for their sales forecasting. With the access to past sales data and by use of data mining techniques, almost all companies and especially pharmaceuticals distribution centers can make accurate and reliable prediction for future sales. Since sales prediction should be performed with high accuracy and in short time, it is impossible to do it with manual or traditional methods. Data mining techniques enhance accuracy and speed up the process.

They collected the required data from a large PDC that dispenses medicine to customers in a number of provinces in Iran. After receiving the orders the company is committed to supplying drugs to provinces within 24hours, cities within 48hours and remotes areas within 72 hours. In keeping with its market leading position, this company needs to have large product inventories in order to meet customers demand, as a shortage of drugs is not acceptable.

The company keeps inventories for about 2months. This fact causes many excessive costs and investments for Iranian PDCs. Thus this gap causes undesired expenses, monthly and precise sales prediction would shorten or even eliminate the gap.

According to restrictions on sales of medicines such as existing new items with short numbers of past sales records and having a great diversity of medicines their objective was concerned with development of a novel and accurate sales forecasting method for pharmaceutical products by means of one of the related data mining approaches to overcome the problem of having numerous kinds of medicine and not having enough past sales records of each medicine.

To predict sales of company, past sales records were collected. The company provided the sales data of nearly 1200 kinds of medicine which were sold to different provinces or centers in Iran during three years. Database of the company included name, code of medicines, sales number, name and code of centers, name of manufacturers and price and monthly date of sales.

To approach their objective, code, date and number of products sold were selected from the database. Three-year monthly sales data were gathered and from PDC , in preprocessing phase raw data was prepared to suit the research objectives, exploratory analysis was performed to specify nature of data and also a comprehensive graph based analysis was performed to find clique sets and group members and visualize the network of drugs.

Sales forecasting models were built in 3 different approaches;

* ARIMA methodology for time series forecasting
* Hybrid neural network approach for forecasting by means of each drugs past records
* Hybrid neural network for time

Their research verified that by applying data mining approaches forecasting performance can be considerably improved since the approach captured different patterns in data.

Disadvantages of Data Mining Approach in prediction

Data mining is not perfectly accurate. Therefore if inaccurate information is used in prediction it will cause serious consequences.

Data mining may violate user privacy. Data mining collects information about people using the pharmaceutical products.

## 

## **2.2.5: Proposed System**

In this project, linear regression will be trained and tested for dataset. The raw data from the source data cleaning to make the data smooth, feature extraction and selection is applied to select best features out of available which are influencing the result more. Machine learning regression model are applied for training dataset to train the model. The train model is then tested and test dataset and validation dataset for checking accuracy of the model.

Figure 1: process to develop the model

ML model for classification

Testing and validation of model

Feature extraction and selection

Data cleaning

Raw sales data

## **2.3: Existing software design and development tools**

## **2.3.1: Python Programming Language**

Python is an interpreted, object-oriented, high level programming language with dynamic semantics. Its high level built in data structures, combined with dynamic typing and dynamic binding makes it very attractive for Rapid application Development, as well as for use as a scripting or glue language to connect existing components together [11]

Python libraries include;

Pandas

It’s an open source python package used for data science and machine learning tasks. It provides support for multi-dimensional array.

It makes it simpler to do the following tasks associated with the working data; Data exploration, data cleaning and data visualization [12]

Plotly

It’s an open source tool used for data visualization and understanding data simply and easily. It supports various types of plots like line charts, scatter charts, histograms and cox plots

Plotly will be used to generate graphs in sales prediction [13]

Scikit-Learn

It’s a python tool that provides supervised and unsupervised learning algorithm

It contains efficient tools for machine learning and statistical modelling including regression, clustering and classification

Proposed system will use regression analysis which is supervised learning algorithm [14]

## **2.4: Justification**

Literature review summarizes and synthesizes the arguments and ideas of existing sales prediction systems and also other prediction system without adding any contributions. With profound knowledge of the gaps exposed in the existing systems proposed system will overpower them.

Python programming will be used to develop the prediction model because its selection of machine learning-specific libraries and frameworks simplify development process and cut development time. Python has a simple syntax and its readability promote rapid testing of complex algorithms

## **2.5: Conclusion**

According to the presented literature review, numerous prediction methods have been offered and each method has its specific advantages and disadvantages in comparison with other techniques. However, none of the accomplished studies described the applications of linear networks in forecasting. They also did not offer novel technique for handling the problem of not having enough past records for prediction.

This motivates the evolution of regression analysis to make precise sales prediction. Regression analysis is used in determining the strength of predictors, forecasting an effect and also trend forecasting

With traditional methods not being of much help to the business organization in revenue growth, use of machine learning approaches prove to be an important aspect for shaping business strategies keeping into consideration the purchase patterns of the customers. Prediction of sales with respect to various factors including sales of previous years helps business adopt suitable strategies for increasing sales and set their foot undaunted in the competitive world

## **CHAPTER 3: RESEARCH METHODOLOGY**

## **3.1: Introduction**

Research methodology is a way to systematically solve a research problem following specific procedures and techniques. Methodology allows one to critically evaluate study’s overall validity and reliability [15].

It discusses how data is collected or generated, and how data is analyzed. I obtained data from both primary and secondary sources. Primary sources were more reliable and enabled me have confidence on decision making.

## **3.2: Data Collection Techniques**

## **3.2.1: Interview**

Interview is a qualitative research technique which involves asking open-ended questions to converse with respondents and collect elicit data about a subject [16].

Type of interviews include;

Personal interview where questions are asked personally directly to the respondent it gives a higher response rate

Telephonic interviews are widely used and easy to combine with online surveys to carry out research effectively.

Email or web-page interview; since online research is growing and more consumers are migrating to more virtual world e-mail and web-page interviews are efficient [16].

**Advantages of using Interviews**

* I was able to gain valuable insights based on the depth of the information gathered and the wisdom.
* Interviews require only simple equipment and build on conversation skills which researchers already have.
* Interviews are more flexible
* Direct contact at the point of interview means data can be checked for accuracy and relevance are they are collected

**Disadvantages of using interviews**

* Data analysis and preparation can be difficult and time consuming.
* Consistency and objectivity are hard to achieve
* Identity of researcher may affect the statements of the interviewee
* Some people may not show up for the interview

## **3.2.2: Questionnaires**

Questionnaire is the main instrument for collecting data in survey research. It’s a set of standardized questions, often called items, which follow a fixed scheme in order to collect individual data about one or more specific topics [17].

I have used both open-ended and closed-ended questions.

**Advantages**

* Result into wide range of views from customers
* Questionnaires are the most affordable ways to gather quantitative data.
* It’s easy and quick to collect results
* When data has been quantified it can be used to compare and contrast other research and maybe used to measure change.

**Disadvantages**

* There is a chance that some questions will be ignored and left unanswered
* Differences in understanding and interpretation
* Questionnaire cannot fully capture emotional responses and feelings

## **3.2.3: Observation**

It’s a technique that involves systematically selecting, watching, listening, reading, touching and recording behavior and characteristics of living beings, objects or phenomena [18].

**Advantages**

* Data can be collected at the time they occur
* Observation study describe observed phenomena as they occur in natural setting
* Offers an opportunity for longitudinal analysis

**Disadvantages**

* Difficulties in quantification
* Sample size observed is usually small
* There is no opportunity to study the past when using observation method

## **3.2.4: Documents and records**

It’s examining existing data from databases, reports and financial records that relate to your area of research [19].

Some companies still record their sales history in books. Therefore I obtained from their sales records. The records contained sales for every month of the year.

The data obtained was useful to predict the sales of the next year for the company

**Advantages of using Documents and Records**

* Easy to obtain historical data
* It’s an inexpensive way to gather information
* Document and record study offers an opportunity for longitudinal analysis.

## **3.2.5: Justification**

Since data collection is essential in research, to gather information in the proposed system two methods will be used; interviews and use of documents and records.

Interviewing specific persons in supermarket will enable one obtain information such as how much sales they make weekly, quarterly and monthly, factors affecting increase and low sales and also how prediction system may help utilize resources if implemented.

Through interviews one is exposed to first-hand information and also helps in gaining more insights into current systems

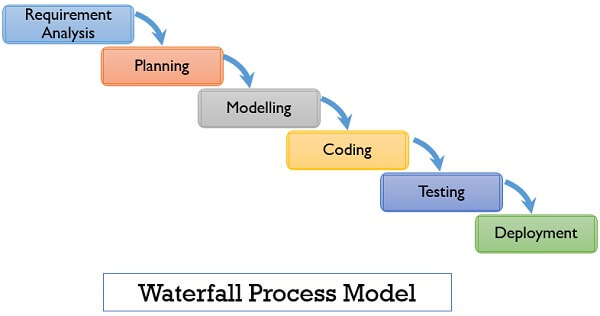
Since prediction system involves use of historical data to obtain

## **3.3: Software Development Techniques**

## **3.3.1: Waterfall Methodology**

Waterfall model is a linear application development that uses rigid phases: when one phase ends, next begins. Steps occur in sequence, and if unmodified, the model does not allow developers to go back to previous steps [20]

It’s also referred as linear-sequential lifecycle model [21]. It follows a structured sequential path from requirements to maintenance, setting out milestones at each steps before next step begins [21].

 Figure 2: Waterfall model

[21]

**Advantages of Waterfall Model**

* Waterfall model divides the entire process of software development into finite independent stages making controlling of each stage easier.
* Requirements are stable and known to the developer at the starting point of the project
* Only one stage is processed at a time thus avoiding confusion
* It’s simple and easy to implement [22]

**Disadvantages of Waterfall Model**

* It’s difficult to implement in complex project
* It’s difficult to state all requirements explicitly at the starting which causes natural uncertainty at the beginning of the project
* A strict waterfall model doesn’t allow going back once the stage is completed. [22]

## **3.3.2: Rapid Application Development Methodology**

RAD is an agile software development approach that focuses more on ongoing software projects and user feedback and less on following a strict plan [23].

RAD develops software via the use of prototypes, dummy, backend databases and its goal is to meet the business need of the system and customer is heavily involved in the process [24].

It consists of four phases [25];

Requirement analysis- Developers, clients and team members communicate to determine the goals and expectations for the project

User Design- involves building out user design through various prototype iterations

Rapid construction- Takes the prototypes and beta systems from design phase and converts them into a working model.

Cutover – implementation phase where finished product is launched

 [26]

Figure 3: Rapid application Methodology

**Advantages of using RAD Methodology**

* RAD lets you break the project into smaller and more manageable tasks
* Task oriented structure allows project managers to optimize their team’s efficiency by assigning tasks according to members specialist and experience.
* Clients get a working product delivered in a shorter time frame
* Regular communication and constant feedback between team members and stakeholders increases the efficiency of design and build process

**Disadvantages of RAD**

* Needs strong team collaboration
* Needs highly skilled developers
* Only suitable for projects which have a small development time
* Only systems which can be modularized can be developed using RAD

## **3.3.3: Agile Methodology**

Agile methodology is a type of project management process, mainly used for software development, where demands and solutions evolve through the collaborative effort of self-organizing and cross-functional teams and their customers [27].

It is used to deliver complex projects due to its adaptiveness. It emphasizes on collaboration, flexibility, continuous improvement and high quality results.

The five phases are;

Project initiation which is about discussing project vision and ROI justification. Team members, time and work resources required are determined.

Planning- it is where the team gets together with their sponsor or product owner and identifies exactly what they are looking for.

Development –once requirements have been defined actual work begins

Production –a handover with relevant training should take place between the production and support teams

Retirement – it is the final stage. Customers are notified and informed about migration to newer releases or alternative options

Figure 4: Agile methodology

Fig 3 [28]

It has several frameworks such as;

Scrum used to implement the ideas behind agile software development

Kanban is a visual method used to paint picture of the workflow process, with an aim to identify any bottlenecks early in the process

FDD- Is a lightweight iterative and incremental software development process with an objective to deliver tangible, working software in timely manner.

Agile methodology has the following benefits [29];

* Better product quality- agile methods have excellent safeguards to make sure that quality is as high as possible
* Higher customer satisfaction- by keeping customers involved and engaged.
* High team morale-being part of self-managing team allows people to be creative, innovative and acknowledged for their expertise.
* Increased collaboration and ownership- development team, product owner and scrum master work closely together on a daily basis

## **3.3.4: Justification**

In this project I have used both agile methodology and waterfall methodology because;

Agile methodology is suitable for projects which comprise multiple iterations of understanding a business problem by asking questions, data acquisition from multiple sources, data cleaning, feature engineering and modelling.

Waterfall methodology is easy to implement and doesn’t need a lot of resources and effort

## **3.4: System Requirements**

## **3.4.1: Software Requirements**

* OPERATING SYSTEM: Windows 10 and higher version, Linux or MacOS
* PROGRAMMING LANGUAGE: Python

## **3.4.2: Hardware Requirements**

* PROCESSOR: Intel Core I 7 and above
* RAM: minimum of 16gb
* Laptop
* Printer

## **3.4.3: Functional Requirements**

Are function or features that must include in any system to satisfy the business needs and be acceptable to the users. The developed system has the following functional requirements;

* The system is able to generate and approximate sales
* The system can collect accurate data from supermarket database in a consistent manner
* Database is updated by the latest values

## **3.4.4: Non-Functional Requirements**

It’s a description of features, characteristics and attributes of the system as well as any constraints that may limit the boundaries of the proposed system. They are based on performance, information, control and security efficiency and services. Based on the developed system, non-functional requirements include;

* The system provides better accuracy
* The system has a simple interface for users to use.
* Perform efficiently in short amount of time.

## **3.5: Conclusion**

Sales forecasting plays a vital role in the business sector in every field. With the help of the sales forecasts, sales revenue analysis will help to get the details needed to estimate both revenue and the income. Linear regression has been evaluated on supermarket sales to find critical factors that influence sales to provide a solution for forecasting sales.

## **Chapter 4: System design, Implementation and Testing.**

## **4.1: Introduction.**

System design is the process of defining the architecture, product design, modules, interfaces, and data for a system to satisfy specified requirements [30].

## **4.2: System design.**

The proposed system is intended to build a model which predicts monthly sales based on the money spent on different platforms for marketing in a supermarket.

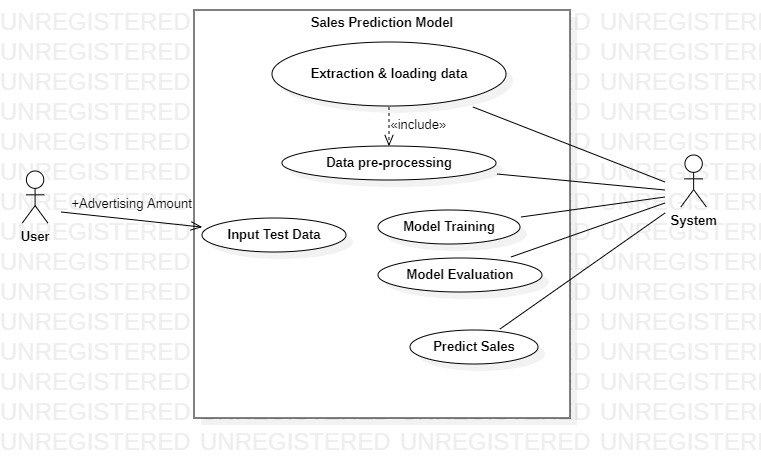
To develop the proposed system the following process of defining the architecture will be followed.

This process is iterative in nature as it trains the model to get the best-suited information for business purposes in this case to predict the amount sales based on money spent.

## **4.2.1: Logical Design**

The logical design of a system pertains to an abstract representation of the data flows, inputs and outputs of the system [31]. This is often conducted via modelling, using an over-abstract (and sometimes graphical) model of the actual system.

Figure Use Case Model

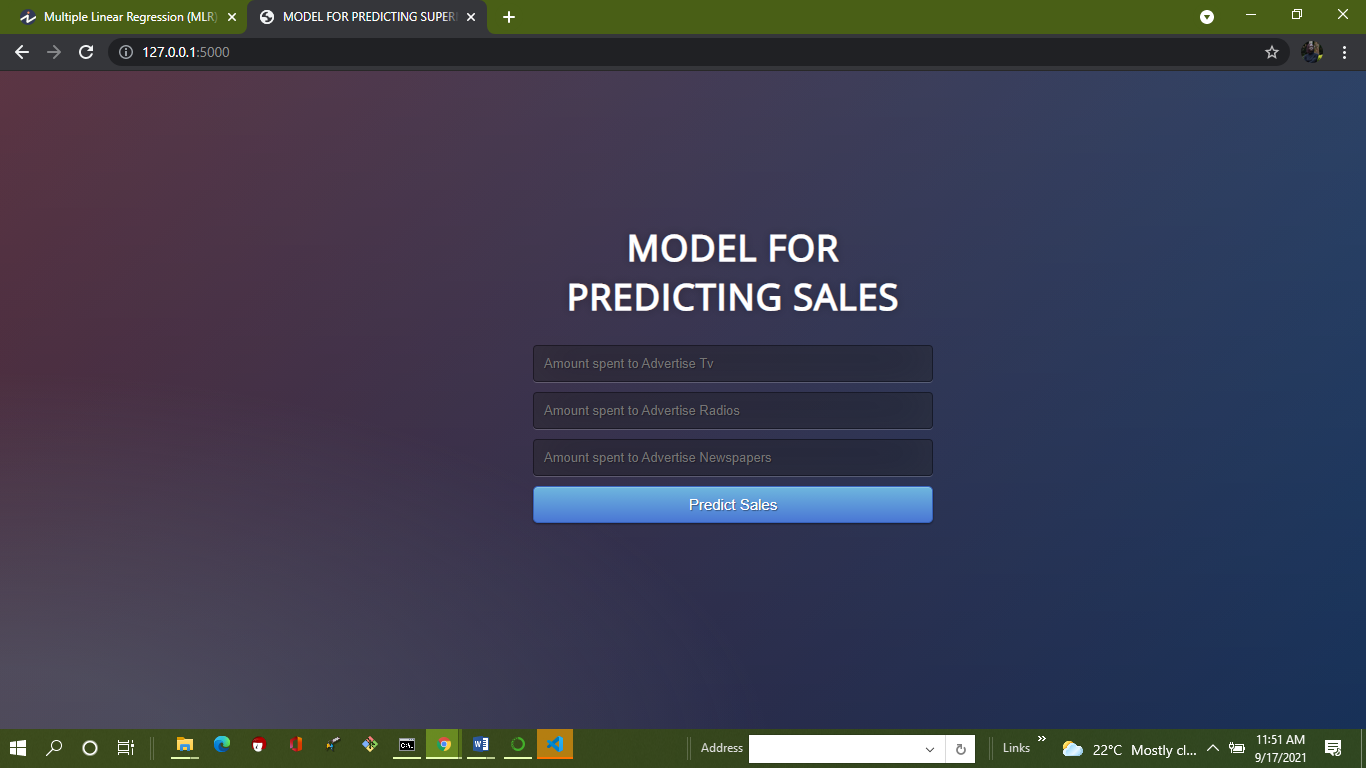


## **4.2.2: User Interface Design.**

User interface design is the visual layout of the elements that a user might interact with in a system.

Sales prediction model will have the following layout where the user can enter amount spent to advertise on TV, Radio and Newspapers so as to predict future sales.

Figure 6 User Interface Design

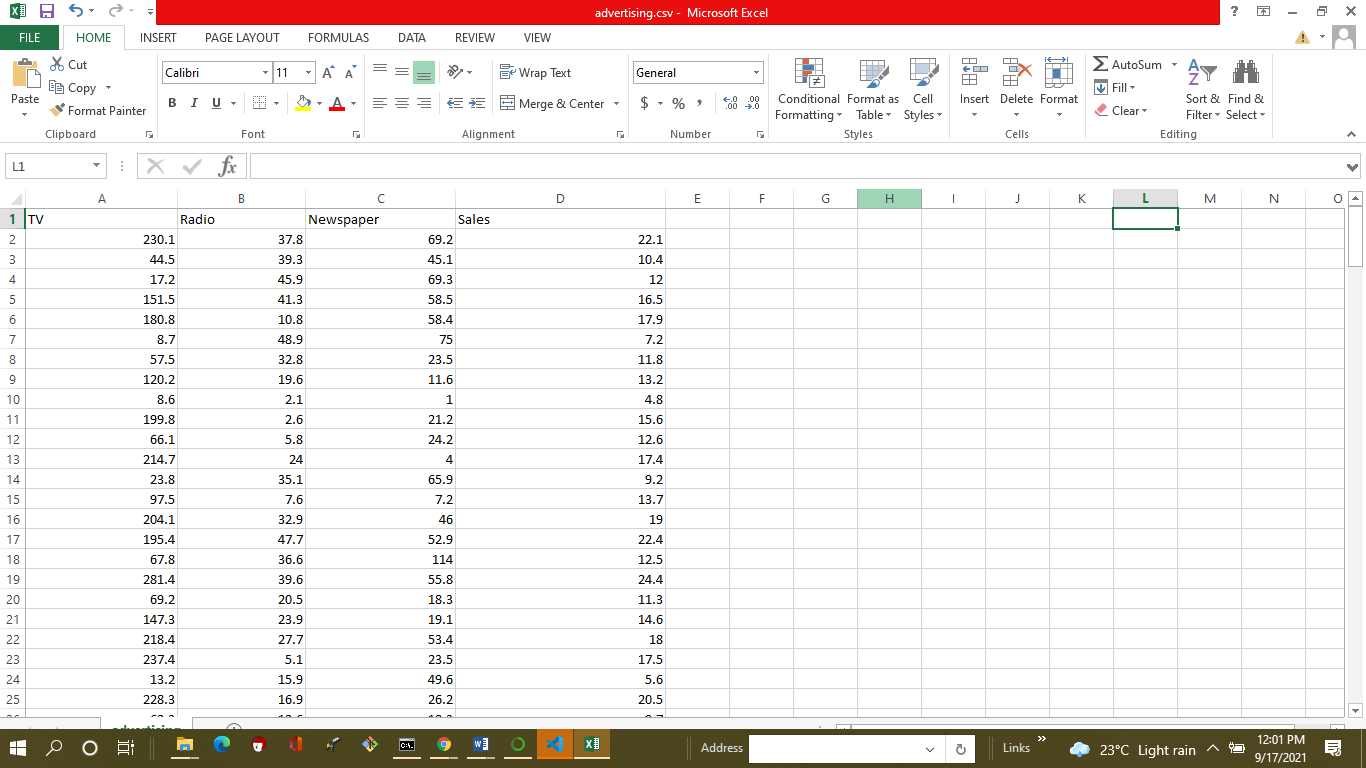


## **4.2.3: Data Design.**

Data design is concerned with how the data is represented and stored within the system.

The dataset used in the model is in tabular form and is stored in database as follows.

Figure Data Design

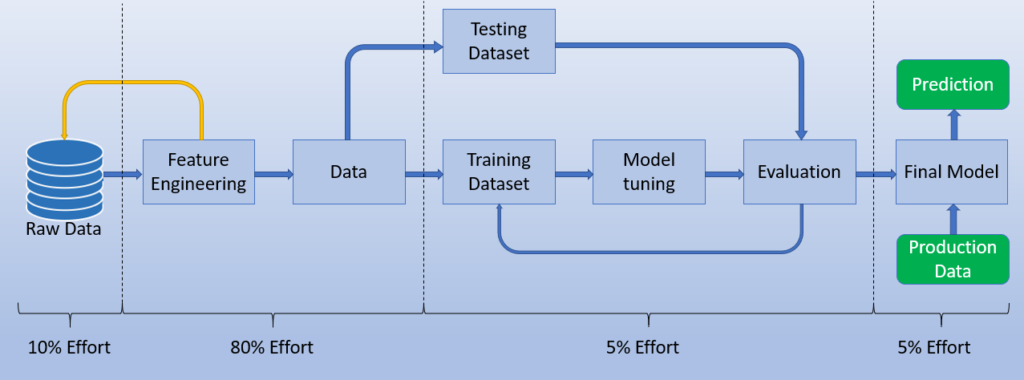


## **4.2.4: Process Design.**

 Process Design is concerned with how data moves through the system, and with how and where it is validated, secured and/or transformed as it flows into, through and out of the system.

To develop the proposed system the following process of defining the architecture will be followed.

Figure Process Design



**Data Collection and Cleaning.**

In the proposed system we will use the advertising dataset given in ISLR and analyze the relationship between TV, Radio and Newspaper and sales using multiple regression model.

Once the data has been collected it is cleaned .Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

**Feature Engineering.**

Feature engineering is the process of using domain knowledge to extract features (characteristics, properties, attributes) from raw data. In our proposed system, the useful features are; TV, Radio, Newspaper and sales.

In order to test a feature’s usefulness, we will proceed to split the data, create some models, and check its efficiency by **setting the values for independent (X) variable and dependent (Y)variable.** X= dataset [['TV', 'Radio', 'Newspaper']]

y = dataset ['Sales']

#### **Split Train/Test**

Once the useful features have been identified, we must split our dataset into a Train and Test dataset.

In the proposed system, we will train the model into the Train dataset and test it in the Test dataset.

The split can be done taking 70% and 30% of the data for train and test respectively.

As shown;

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size = 0.3, random\_state = 100)

**Model Tuning.**

The proposed model uses Multiple Linear regression algorithm to predict the sales.

**Multiple linear regression** (MLR) algorithm is used to estimate the relationship between**two or more independent variables**and**one dependent variable.**

Advantages of Regression Analysis algorithm.

Simple implementation

Linear Regression is a very simple algorithm that can be implemented very easily to give satisfactory results.

Performance on linearly separable datasets

Linear regression fits linearly separable datasets almost perfectly and is often used to find the nature of the relationship between variables.

Overfitting can be reduced by regularization

**Overfitting** is a situation that arises when a machine learning model fits a dataset very closely and hence captures the noisy data as well.

**Regularization** is a technique that can be easily implemented and is capable of effectively reducing the complexity of a function so as to reduce the risk of overfitting.

**Evaluation of the Model.**

Model evaluation aims to estimate the generalization accuracy of a model on future (unseen/out-of-sample) data.

The proposed model will use the following evaluation metrics to measure how good a model performs and how well it approximates the relationship.

Mean Squared Error (MSE)

It is the most common metric for regression tasks. It has a convex shape. It is the average of the squared difference between the predicted and actual value.

Mean Absolute Error (MAE)

This is simply the average of the absolute difference between the target value and the value predicted by the model.

R-squared or Coefficient of Determination

This metric represents the part of the variance of the dependent variable explained by the independent variables of the model. It measures the strength of the relationship between your model and the dependent variable.

Root Mean Squared Error (RMSE)

This is the square root of the average of the squared difference of the predicted and actual value.

#### **Final Model**

The last step the proposed system will undergo is getting the final model. Once we have obtained the best tuning for a model, we train that model into the full dataset (Train andTest) in order to train the model with all the available data.

Finally, the model is prepared to predict future sales, so we can introduce future sales and start showing the predictions.

The purpose of the System Design process is to provide sufficient detailed data and information about the system and its system elements to enable the implementation consistent with architectural entities as defined in models and views of the system architecture.

Using the proposed system design we will be able to implement the stated steps to come up with our model.

## 4.3: Implementation Approaches

In order to successfully achieve our intended goal of developing a model for predicting sales, we need to have and implementation plan.

Implementation plan is designed to document, in detail, the critical steps necessary to put your solutions into practice.

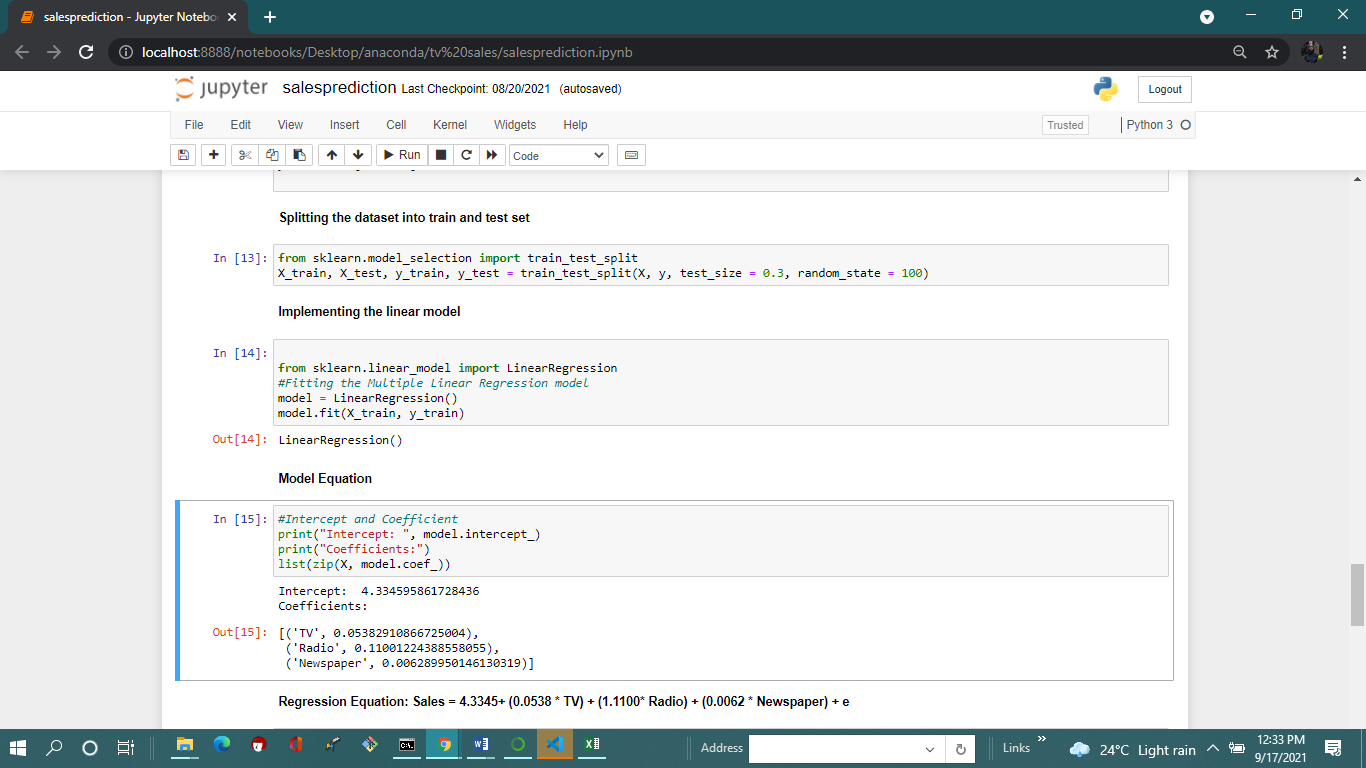
To implement the steps identified in proposed system design the following approaches have been used.

## 4.3.1: Multiple Linear Regression Algorithm.

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable [32].

The main goal of regression is the construction of an efficient model to predict the total sales from a bunch of attribute variables that is money spent to advertise TV sales, Radio sales and Newspaper sales.

Multiple linear regression algorithm functions as follows;



## **4.3.2: Flask framework.**

To develop and implement the user interface design the model uses flask framework for frontend design.

Flask is a micro web framework written in python [33]

## **4.4 Coding Details and Code Efficiency.**

**Importing necessary packages and reading the dataset**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

**numpy**: NumPy stands for numeric Python, a python package for the computation and processing of the multi-dimensional and single-dimensional array elements.

**pandas**: Pandas provide high-performance data manipulation in Python.

**matplotlib**: Matplotlib is a library used for data visualization. It is mainly used for basic plotting. Visualization using Matplotlib generally consists of bars, pies, lines, scatter plots, and so on.

**seaborn**: Seaborn is a library used for making statistical graphics of the dataset. It provides a variety of visualization patterns. It uses fewer syntax and has easily interesting default themes. It is used to summarize data in visualizations and show the data’s distribution.

**Loading/Reading the Dataset.**

#Reading the dataset

dataset = pd.read\_csv("advertising.csv")

dataset

**Data Inspection**

dataset.tail(10)

**Data Cleaning**

# Checking Null values

dataset.isnull().sum()\*100/dataset.shape[0]

# There are no NULL values in the dataset, hence it is clean.

**Outlier Analysis to identifying the anomalous observation in the dataset**

# Outlier Analysis

fig, axs = plt.subplots(3, figsize = (5,5))

plt1 = sns.boxplot(dataset['TV'], ax = axs[0])

plt2 = sns.boxplot(dataset['Newspaper'], ax = axs[1])

plt3 = sns.boxplot(dataset['Radio'], ax = axs[2])

plt.tight\_layout()

**Exploratory Data Analysis**

Exploratory data analysis (EDA) is used to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

It can also help determine if the statistical techniques you are considering for data analysis are appropriate.

**Splitting datasets.**

Setting the values for independent (X) variable and dependent (Y) variable

X= dataset[['TV', 'Radio', 'Newspaper']]

y = dataset['Sales']

Splitting the dataset into train and test set

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size = 0.3, random\_state = 100)

**Data Visualization.**

It is the graphical representation of information and data. By using [visual elements like charts, graphs, and maps](https://www.tableau.com/learn/articles/data-visualization/glossary), data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

**Scatter plot**

Let's see how Sales are related with other variables using scatter plot.

sns.pairplot(dataset, x\_vars=['TV', 'Newspaper', 'Radio'], y\_vars='Sales', height=4, aspect=1, kind='scatter')

plt.show()

**Boxplot**

sns.boxplot(dataset['Sales'])

plt.show()

**Heatmap**

# Let's see the correlation between different variables.

sns.heatmap(dataset.corr(), cmap="YlGnBu", annot = True)

plt.show()

**Implementing the Linear model**

from sklearn.linear\_model import LinearRegression

#Fitting the Multiple Linear Regression model

model = LinearRegression()

model.fit(X\_train, y\_train)

**Model Equation**

#Intercept and Coefficient

print("Intercept: ", model.intercept\_)

print("Coefficients:")

list(zip(X, model.coef\_))

**Predicting test set**

#Prediction of test set

y\_pred= model.predict(X\_test)

#Predicted values

print("Prediction for test set: {}".format(y\_pred))

**Evaluating the Model.**

#Model Evaluation

from sklearn import metrics

meanAbErr = metrics.mean\_absolute\_error(y\_test, y\_pred)

meanSqErr = metrics.mean\_squared\_error(y\_test, y\_pred)

rootMeanSqErr = np.sqrt(metrics.mean\_squared\_error(y\_test, y\_pred))

print('R squared: {:.2f}'.format(model.score(X,y)\*100))

print('Mean Absolute Error:', meanAbErr)

print('Mean Square Error:', meanSqErr)

print('Root Mean Square Error:', rootMeanSqErr)

**Saving the model using pickle.**

import pickle

pickle.dump(model, open('model.pkl','wb'))

**Deploying the model using flask**

from flask import Flask, request, jsonify, render\_template

import pickle

import numpy as np

app=Flask(\_\_name\_\_)

model = pickle.load(open('model.pkl', 'rb'))

@app.route('/')

def home():

    return render\_template ('index.html')

@app.route('/predict',methods=['POST'])

def predict():

    '''

    For rendering results on HTML GUI

    '''

    int\_features = [int(x) for x in request.form.values()]

    final\_features = [np.array(int\_features)]

    prediction = model.predict(final\_features)

    output = round(prediction[0], 2)

    return render\_template('index.html', prediction\_text='Total sales $ {}'.format(output))

if \_\_name\_\_=='\_\_main\_\_':

    app.run(debug=True)

**User interface Design**

<!DOCTYPE html>

<html >

<!--From https://codepen.io/frytyler/pen/EGdtg-->

<head>

  <meta charset="UTF-8">

  <title>MODEL FOR PREDICTING SUPERMARKET SALES</title>

  <link href='https://fonts.googleapis.com/css?family=Pacifico' rel='stylesheet' type='text/css'>

<link href='https://fonts.googleapis.com/css?family=Arimo' rel='stylesheet' type='text/css'>

<link href='https://fonts.googleapis.com/css?family=Hind:300' rel='stylesheet' type='text/css'>

<link href='https://fonts.googleapis.com/css?family=Open+Sans+Condensed:300' rel='stylesheet' type='text/css'>

<link rel="stylesheet" href="{{ url\_for('static', filename='css/style.css') }}">

</head>

<body>

 <div class="login">

  <h1> MODEL FOR PREDICTING SALES</h1>

     <!-- Main Input For Receiving Query to our ML -->

    <form action="{{ url\_for('predict')}}"method="post">

      <input type="text" name="TV" placeholder="Amount spent to Advertise Tv" required="required" />

        <input type="text" name="Radio" placeholder="Amount spent to Advertise Radios" required="required" />

    <input type="text" name="Neswpaper" placeholder="Amount spent to Advertise Newspapers" required="required" />

        <button type="submit" class="btn btn-primary btn-block btn-large">Predict Sales</button>

    </form>

   <br>

   <br>

   {{ prediction\_text }}

 </div>

</body>

</html>

## **4.5: Testing Approach**

Software testing has the power to point out all the defects and flaws during development. . Different kinds of testing allow us to catch bugs that are visible only during runtime.

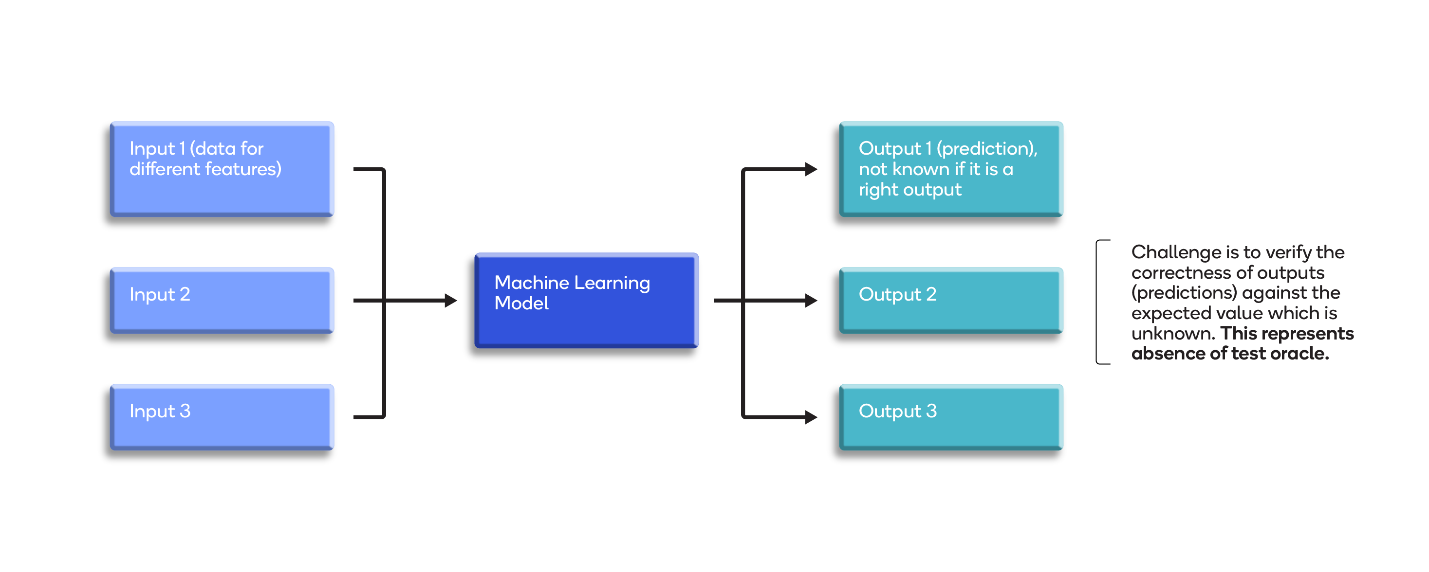
The purpose of machine learning testing is to ensure that this learned logic will remain consistent, no matter how many times we call the program.

Functional Testing

It is a type of software testing that validates the software system against the functional requirements/specifications.

The purpose of Functional tests is to test each function of the software application, by providing appropriate input, verifying the output against the Functional requirements.

Figure Functional Testing



Functional testing mainly involves;

**Black-box testing** of machine learning (ML) models refers to testing with no knowledge about the internal details of the model, such as the algorithm used to create it and the features in it. The main objective of black-box testing is to ensure the quality of the models in a sustained manner.

**Unit tests**. The program is broken down into blocks, and each element (unit) is tested separately

It involves testing individual units of the source code, such as functions, methods, and class to ascertain that they meet the requirements and have expected results.

Each piece of code has been tested individually and results executed.

**Regression tests**. They cover already tested software to see if it doesn’t suddenly break and also ensures quality of the user experience along with the new changes.

**Integration tests**

These tests aim to determine whether modules that have been developed separately work as expected when brought together. In terms of a data pipeline, these can check that:

* The data cleaning process results in a dataset appropriate for the model
* The model training can handle the data provided to it and outputs results (ensuring that code can be refactored in the future)
* The data is consumable by the model (a label exists for every input, the types of the data are accepted by the type of model chosen)
* We are able to refactor our code in the future, without breaking the end to end functionality.

## **4.6. Modifications and Improvements.**

Attempt performance metrics such as time while predicting the sales.

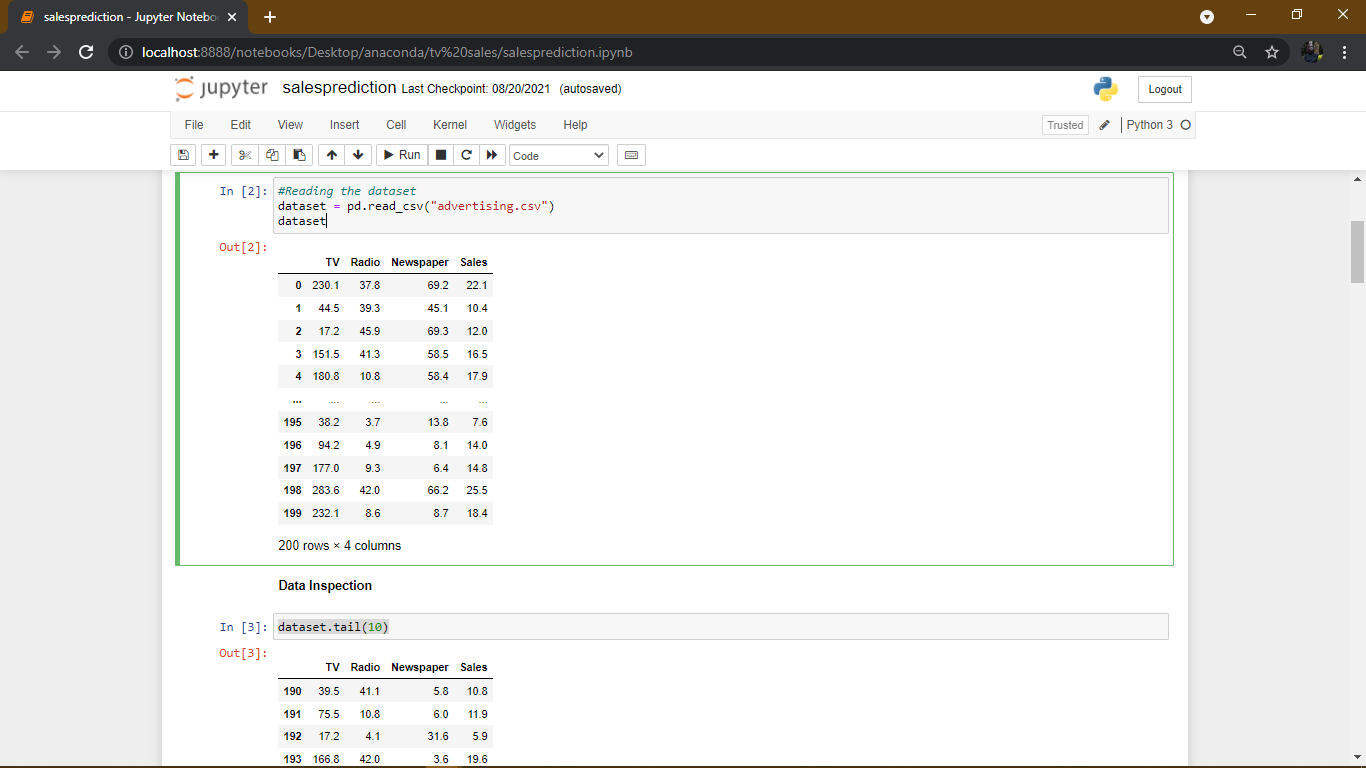
## **Chapter 5**

## **5.1. Test Reports**

Test report is a document which contains a summary of all test activities and final test results of a testing project [34].

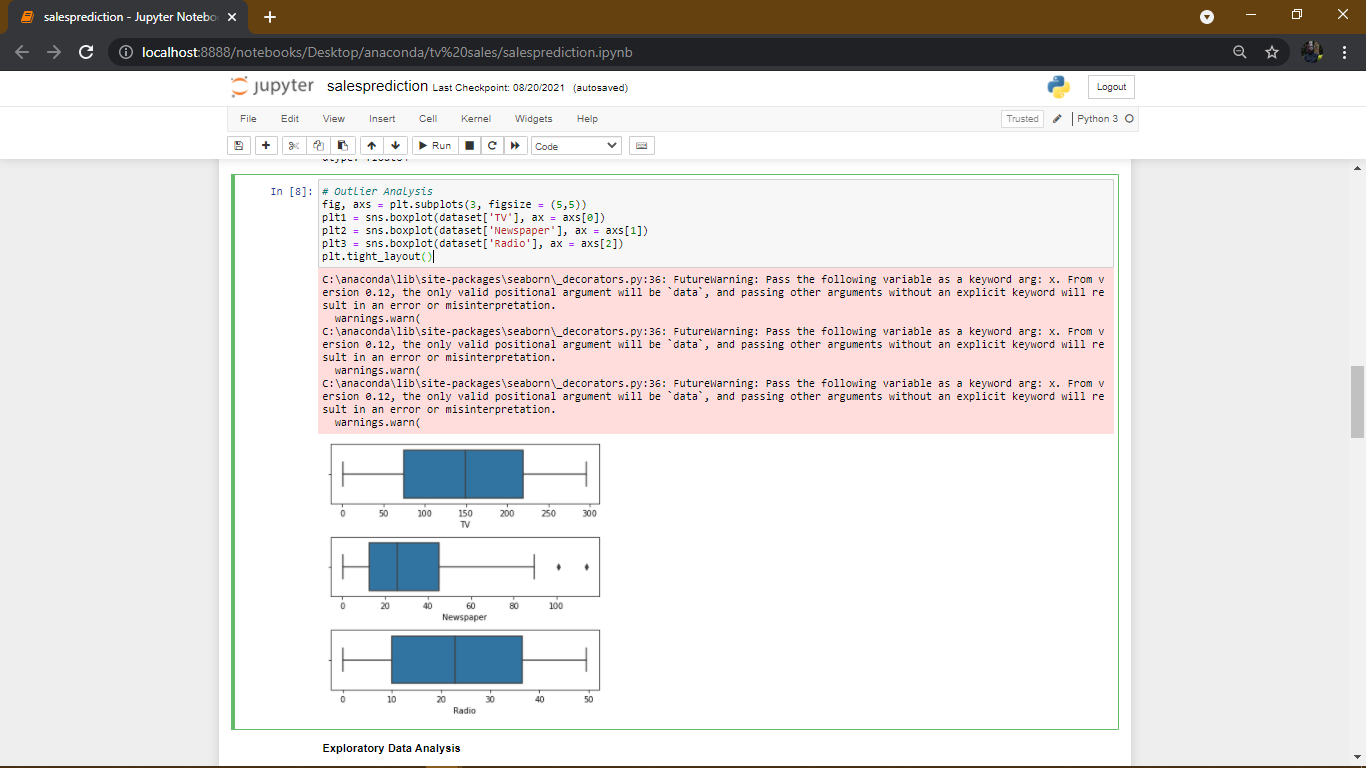
Reading the dataset (Output)

Figure Dataset

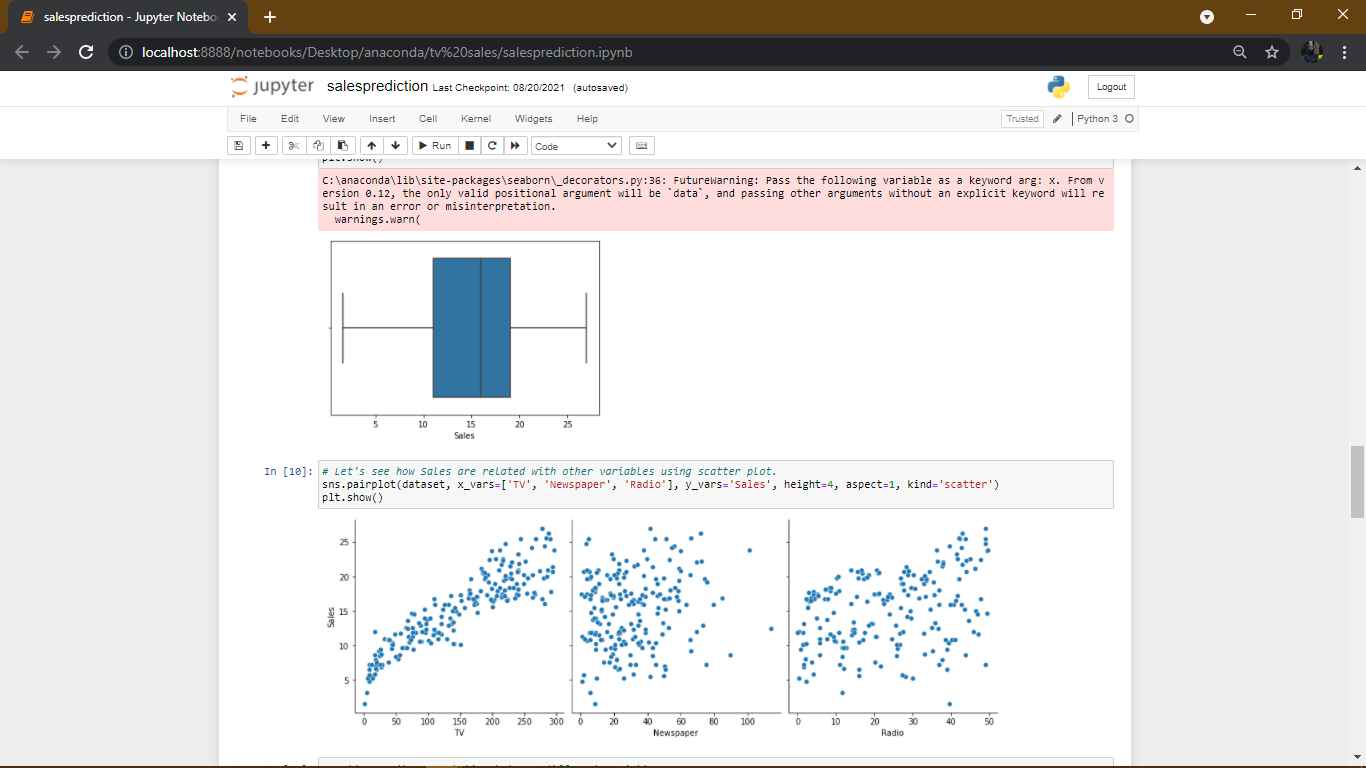


**Outlier Analysis**

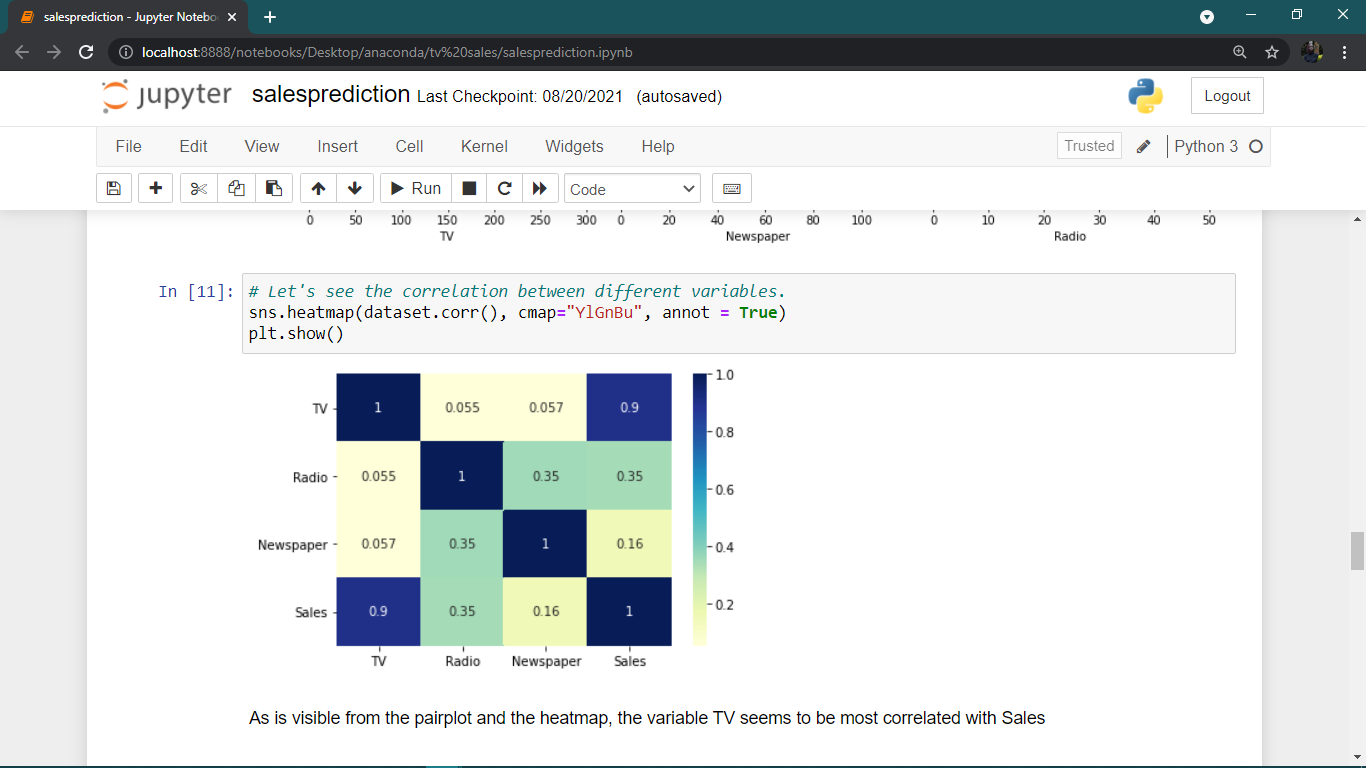
Figure Outlier Analysis

****

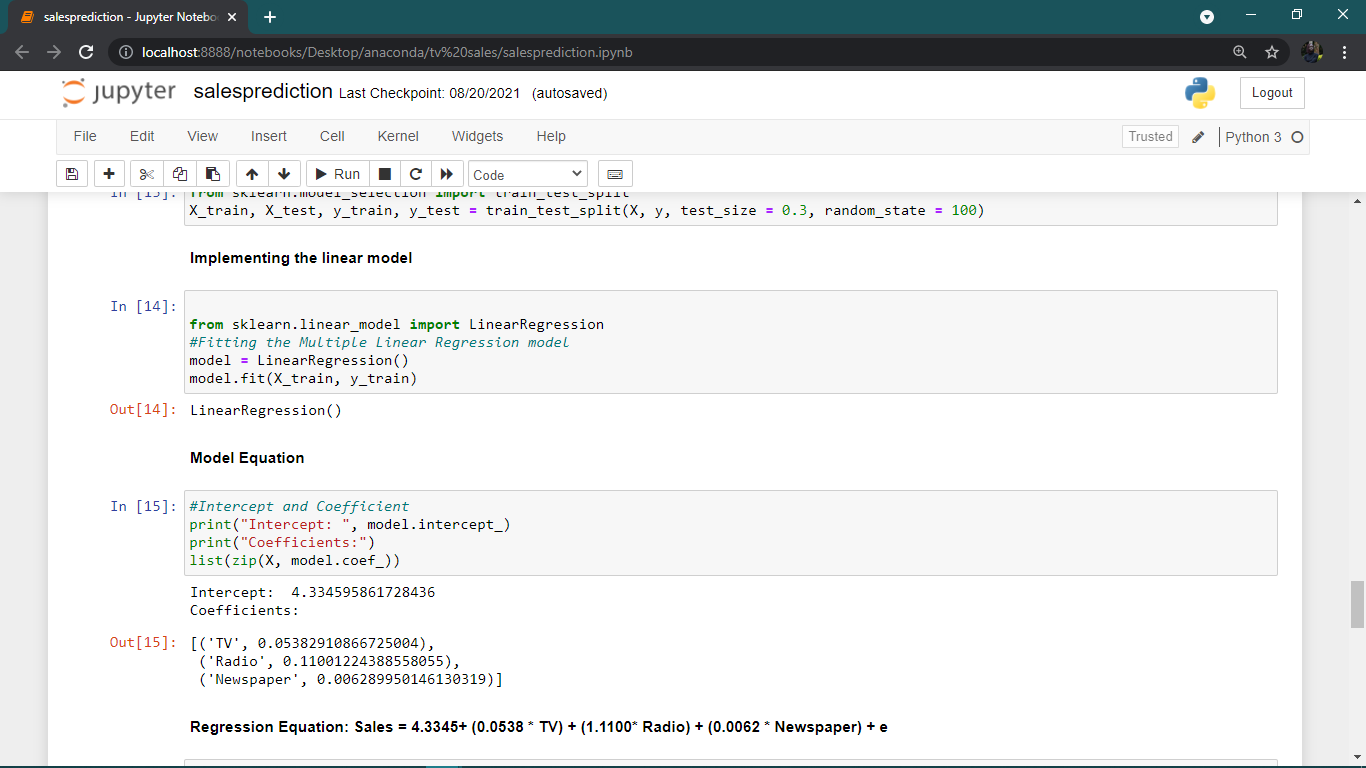
**Relationship between sales and other Variables.**

****

**Correlation between different variables**

****

**Model Equation**

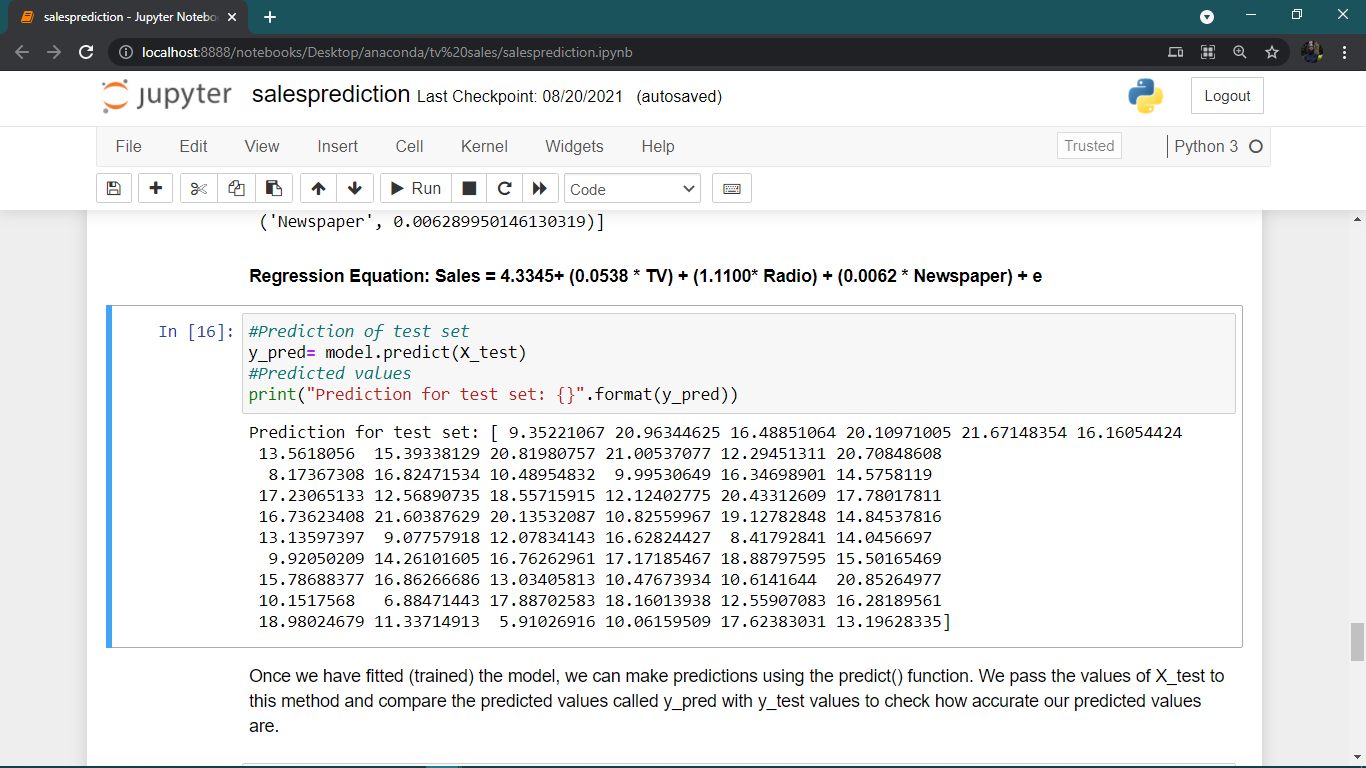
****

**Model Evaluation Results.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | R squared | Mean Absolute Error | Mean Square Error | Root Mean Square Error |
| **Linear Regression** | 90.11 | 90.11 | 2.6360765623280673 | 1.6235998775338913 |

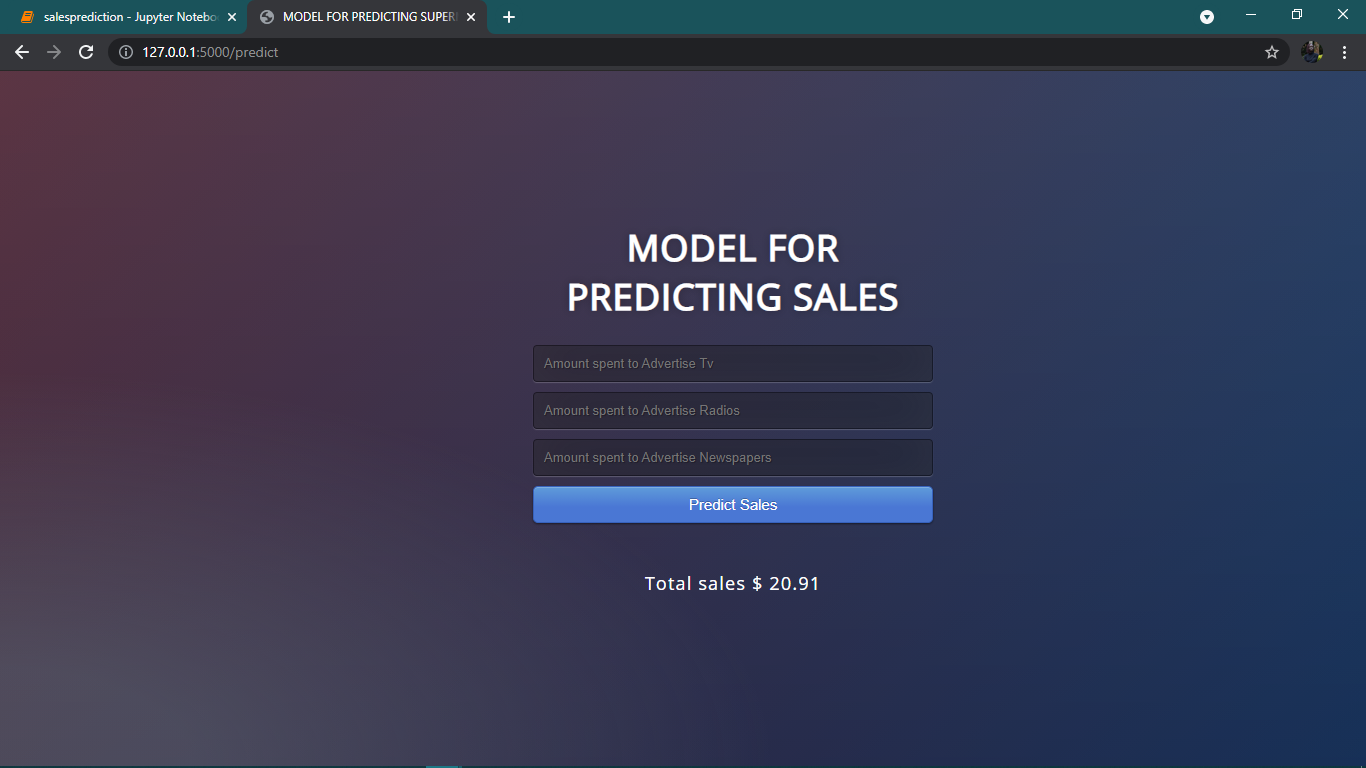
From the above results, Multiple Linear Regression model performs well as 90.11% of the data fit the regression model. Also, the mean absolute error, mean square error, and the root mean square error are less

**Prediction of test set.**



**Deploying the model using Flask and a sample prediction**

Figure Model Deployment



## **5.2: User Documentation**

Sales prediction system is a system that predicts total monthly sales based on money spent to advertise Tv sales, Radio sales and Newspaper sales.

Dataset used in this project is from Kaggle.com. You can also create your dataset also.

The project uses the following tools

Anaconda- it is a scientific python distribution that comes with all necessary packages needed to build the model. The packages include pandas, numpy, sklearn and Jupyter notebook which is an interactive, open source web application for creating and sharing documents that integrate live code.

Jupyter notebook is used to perform task such as data cleaning, data transformation, exploratory data analysis, statistical modelling, machine learning and data visualizations.

Visual studio code- it is a code editor redefined and optimized for building and debugging modern web and cloud applications. In this project the user interface design has been designed using the flask framework.

The interface has fields that enables users enter the test data. After entering the test data the system is able to predict the sales.

The model has been trained using the multiple linear regression algorithm

## **Chapter 6: Conclusions and Future Works.**

## **6.1. Conclusion**

Sales forecasting is a pivotal part of the financial planning of business for any organization. It can be said as a self-assessment tool which uses the statistics of the past and the current sales in order to predict future performance.

Sales forecasting plays an important role in optimizing the supermarket sales process. Financial and Sales planning with the help of the sales forecasts helps to get the information needed to predict the revenue as well as the profit.

Thus, in finding such solution for sales forecasts Linear Regression algorithm have been evaluated on sales data which can forecast the short term sales and help the organization in making the key decisions. After performing the various statistical tests and performance metrics, it is found that Linear Regression is a suitable algorithm in accordance to the chosen dataset and thus accomplishing the aim of this project.

## **6.2: Future Works**

In future work one can attempt performance metrics such as time while predicting the sales. These metrics can play a crucial role in evaluating multiple Machine Learning algorithms.

And also one can attempt to implement more accurate data in the continued study. Machine Learning has the advantage of analyzing data and key variables so that you can aim to develop a systematic approach using a variety of Machine Learning techniques.

**APPENDICES**

## **APP 1: Budget**

Table 1: Budget

|  |  |  |  |
| --- | --- | --- | --- |
| ITEM | QUANTITY | UNIT PRICE | TOTAL(Ksh) |
| Printing and binding |  | 5000 | 5000 |
| Laptop | 1 | 40000 | 40000 |
| Software | 3 | 40000 | 120000 |
| Internet |  | 6000 | 6000 |
| Miscellaneous |  | 15000 | 15000 |
| TOTAL (ksh) |  |  | 186,000 |

Table 2: Schedule

## **APP2: Schedule**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ACTIVITY | MARCH | APRIL | MAY | JUNE | JULY | AUGUST | SEPTEMBER |
| Project identification |  |  |  |  |  |  |  |
| System analysis |  |  |  |  |  |  |  |
| System Design |  |  |  |  |  |  |  |
| Coding and Testing |  |  |  |  |  |  |  |
| Implementation |  |  |  |  |  |  |  |
| Documentation |  |  |  |  |  |  |  |
| Project submission |  |  |  |  |  |  |  |

# References

|  |  |
| --- | --- |
| [1] | K. Bishop, "Sales Hacker," 4 December 2020. [Online]. Available: https://www.saleshacker.com/sales-forecasting-101/. [Accessed 7 june 2021]. |
| [2] | G. tsoumakas, "Survey Of machine learning techniques for food sales techniques," *Artificial intelligence review,* vol. 52, pp. 441-447, 2018. |
| [3] | S. McCombes, "How to write a Literature Review," 22 February 2019. |
| [4] | A. Teelab, "Time Series Forecating using Artificial Neural networks," *Future computing and informatics journal,* vol. 3, no. 2, pp. 334-340, 2018. |
| [5] | W. F. H. N. Jun Wang, "Financial Time Series Prediction Using Elman Recurrent Random Neural Networks," *Computational Intelliegence and Neuroscience,* vol. 2016, p. 14, 2016. |
| [6] | F. E. H. T. a. L. Cao, "Application of support Vector Machine in Financial Time series forecasting," *Neural Computings and Applications,* vol. 29, no. 4, pp. 184-192, 2015. |
| [7] | October 2019. [Online]. Available: https://www.diva-portal.org/smash/get/diva2:1366957/FULLTEXT02. |
| [8] | A. H. P.Doganis, "Forecasting for shelf life food using AI and evolutionary computing," *Food Engineering,* vol. 75, no. 2, pp. 196-204, 2006. |
| [9] | M. &. H. P. Holmberg, *Abstract Machine Learning for Restaurant Forecast,* 2018. |
| [10] | I. Ribeiro, "Sales Prediction for a pharmaceutical distribution company," in *11th Iberian Conference on Information systems and technologies*, Las Pamas, 2016. |
| [11] | P. S. foundation, "Python," [Online]. Available: https://www.python.org/doc/essays/blurb/. [Accessed 8 june 2021]. |
| [12] | "Pandas and numpy fundamentals," Daquest labs, 15 November 2018. [Online]. Available: https://www.dataquest.io/course/pandas-fundamentals/. [Accessed 8 June 2021]. |
| [13] | "ML Regression," plotly, 2020. [Online]. Available: https://plotly.com/python/ml-regression/. [Accessed 8 June 2021]. |
| [14] | "scikit learn: Machine learning in python," Dataquest Labs, 15 November 2018. [Online]. Available: https://www.dataquest.io/blog/sci-kit-learn-tutorial/. [Accessed 8 june 2021]. |
| [15] | M. Patel and N. Patel, "Exploring Research Methodology," *International Journal of Research and Review,* vol. 6, no. 3, March 2019. |
| [16] | "Question pro: Types and methods of interview in Research," Questionpro survey sofware, [Online]. Available: https://www.questionpro.com/blog/types-of-interviews/. [Accessed 10 june 2021]. |
| [17] | P. J.Lavkaras, "Sampling,Survey Research, Data Collection, Response Rates, Random Sampling," *Encyclopedia of Survey Research methods,* 1 January 2011. |
| [18] | "Observation method of Data Collection," iEduNote, [Online]. Available: https://www.iedunote.com/observation-method-of-data-collection. [Accessed 8 June 2021]. |
| [19] | S. M. S. Kabir, Methods of Data Collection, Bangladesh: Book Zone Publication, July, 2016. |
| [20] | E. Conrad, Eleventh Hour CISSP, Elsevier B.V, 2011. |
| [21] | P. S. Ganney and E. Claridge, Clinical Engineering, UK: Elsevier ltd, 2020. |
| [22] | N. T, "Binary Terms," Affiliate Labs, 19 February 2020. [Online]. Available: https://binaryterms.com/waterfall-process-model.html. [Accessed 10 June 2021]. |
| [23] | S. Idesis, "Rapid Application Development: Why RAD and Why Now," 9 October 2020. |
| [24] | J. Feldman, CISSP Study Guide, UK: Elsevier, 2016. |
| [25] | "4 Phases of RAD," LucidChart, [Online]. Available: https://www.lucidchart.com/blog/rapid-application-development-methodology. [Accessed 11 June 2021]. |
| [26] | "Rapid Application Development: Changing How Developers Work," Kissflow, 31 March 2021. [Online]. Available: https://kissflow.com/low-code/rad/rapid-application-development/. [Accessed 12 June 2021]. |
| [27] | D. Muslihat, "Agile Methodology: An Overview," Zenkit, 2 March 2018. [Online]. Available: https://zenkit.com/en/blog/agile-methodology-an-overview/. [Accessed 14 June 2021]. |
| [28] | G. Windsor, "5 stages of agile system," 28 February 2020. |
| [29] | T. Bunsiri, "Benefits of Agile Project Management," *APHEIT JOURNAL,* vol. 5, no. 1, pp. 23-29, 2016. |
| [30] | "Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Systems\_design. [Accessed 17 September 2021]. |
| [31] | S. Link, "The Logic of Design as a Conceptual Logic of Information," *Minds and Machines ,* no. 27, p. 495–519, 14 June 2017. |
| [32] | A. HAYES, "Multiple Linear Regression," 30 March 2021. |
| [33] | M. Grinberg, Flask Web Development, O'Reilly Media, Inc., 2018. |
| [34] | T. Hamilton, "Test Summary Reports Tutorial: Learn with Example & Template," Guru 99, 27 August 2021. [Online]. Available: https://www.guru99.com/how-test-reports-predict-the-success-of-your-testing-project.html. [Accessed 17 September 2021]. |
| [35] | S. m. |