A PROJECT

on

WALMART HOLIDAY SALES ANALYSIS AND SALES PREDICTION.

Submitted By

1.Rutika Sanjay Patil (69)

2.Sharayu Mahendra Sutar (74)

Under the Guidance of

Prof. Madhumati Lotkar

Department of Information Technology

Saraswati Education Society’s

SARASWATI COLLEGE OF ENGINEERING

Kharghar,Navi Mumbai

(Affiliated to University of Mumbai)

Academic Year :-2021-22

Saraswati College of Engineering, Kharghar

Vision:

To become center of excellence in Engineering education and research.

Mission:

To educate students to become quality technocrats for taking up challenges in all facets of life.

Department of Information Technology

Vision:

To create technically qualified talent through research to take up challenges in industries.

Mission:

1. To impart quality education.

2. To develop technical and managerial skills through training and modern teaching-learning process.

SARASWATI Education Society's

SARASWATI College of Engineering

Learn Live Achieve and Contribute

Kharghar, Navi Mumbai - 410 210.

CERTIFICATE

This is to certify that the requirements for the synopsis entitled "   "

Have been successfully completed by the following students:

1)Rutika Sanjay Patil(69)

2)Sharayu Mahendra Sutar(74)

In partial fulfillment of Sem –VI **Bachelor of Engineering of Mumbai University in Information Technology** of Saraswati college of Engineering, Kharghar during the academic year 2021-22.

Internal Guide

Prof. Madhumati Lotkar

Project coordinator	**Head OF Department**
Prof.Madhumati L.	Prof. Mahi K

Acknowledgement

A project is something that could not have been materialized without cooperation of many people. This project shall be incomplete if I do not convey my heartfelt gratitude to those people from whom I have got considerable support and encouragement.

It is a matter of great pleasure for us to have a respected **Prof. Madhumati L** as my project guide. We are thankful to her for being constant source of inspiration.

We would also like to give our sincere thanks to **Prof. Mahi K., H.O.D, Information Technology** Department**, Prof. Madhumati L., Project co-ordinator** for their kind support.

We would like to express our deepest gratitude to **Dr. Manjusha Deshmukh,**our principal of Saraswati college of Engineering, Kharghar, Navi Mumbai

Last but not the least I would also like to thank all the staffs of Saraswati college of Engineering (Information Technology Department) for their valuable guidance with their interest and valuable suggestions brightened us.

1)Rutika Sanjay Patil(69)
2)Sharayu Mahendra Sutar(74)

WALMART HOLIDAY SALES ANALYSIS AND SALES PREDICTION.

ABSTRACT

Information technology in this 21st century is reaching the skies with large-scale of data to be processed and studied to make sense of data where the traditional approach is no more effective. Now, retailers need a 360-degree view of their consumers, without which, they can miss competitive edge of the market. Retailers have to create effective promotions and offers to meet its sales and marketing goals, otherwise they will forgo the major opportunities that the current market offers. Many times it is hard for the retailers to comprehend the market condition since their retail stores are at various geographical locations. Big Data application enables these retail organizations to use prior year's data to better forecast and predict the coming year's sales. It also enables retailers with valuable and analytical insights, especially determining customers with desired products at desired time in a particular store at different geographical locations. In this paper, we analysed the data sets of world's largest retailers, Walmart Store to determine the business drivers and predict which departments are affected by the different scenarios (such as temperature, fuel price and holidays) and their impact on sales at stores' of different locations. We have made use of Scala and Python API of the Spark framework to gain new insights into the consumer behaviours and comprehend Walmart's marketing efforts and their data-driven strategies through visual representation of the analysed data.

Keywords—Big Data Analytics; R, Multiple Linear regression

INDEX

1.INTRODUCTION

Predicting future sales for a company is one of the most important aspects of strategic planning. I wanted to analyse how internal and external factors of one of the biggest companies in the US can affect their Weekly Sales in the future. This module contains complete analysis of data, includes time series analysis, identifies the best performing stores, performs sales prediction with the help of multiple linear regression. The data collected ranges from 2010 to 2012, where 45 Walmart stores across the USA were included in this analysis. It is important to note that we also have external data available like CPI, Unemployment Rate and Fuel Prices in the region of each store which, hopefully, help us to make a more detailed analysis.

2.PROBLEM STATEMENT

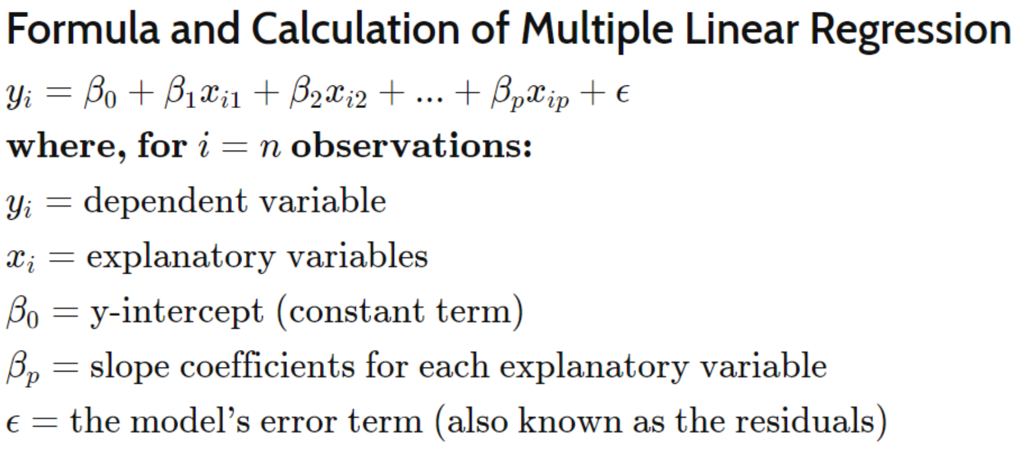
- Retailer’s first priority is usually to understand their customers to be able to satisfy their needs so that these customers will return to the store for future needs, thus increasing the product demands and adding to the business value. These businesses want this information to plan where and when to invest profitably. Walmart uses data mining to discover patterns in point of sales data. Data mining helps Walmart find patterns that can be used to provide product recommendations to users based on which products were bought together or which products were bought before the purchase of a particular product. Effective data mining at Walmart has increased its conversion rate of customers. A familiar example of effective data mining through association rule learning technique at Walmart is – finding that Strawberry pop-tarts sales increased by 7 times before a Hurricane.
- After Walmart identified this association between Hurricane and Strawberry pop-tarts through data mining, it places all the Strawberry pop-tarts at the checkouts before a hurricane. Another noted example is during Halloween, sales analysts at Walmart could look at the data in real-time and found that thought a specific cookie was popular across all walmart stores, there were 2 stores where it was not selling at all. The situation was immediately investigated and it was found that simple stocking oversight caused the cookies not being put on the shelves for sales. This issue was rectified immeadiately which prevented further loss of sales.Walmart tracks and targets every consumer individually. Walmart has exhaustive customer data of close to 145 million Americans

3.PROPOSED SYSTEM

3.1 ALGORITHM

- **Multiple Linear Regression**

Multiple linear regression is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. The independent variables can be continuous or categorical (dummy coded as appropriate).There are 3 major uses for multiple linear regression analysis. First, it might be used to identify the strength of the effect that the independent variables have on a dependent variable. Second, it can be used to forecast effects or impacts of changes. That is, multiple linear regression analysis helps us to understand how much will the dependent variable change when we change the independent variables. Third, multiple linear regression analysis predicts trends and future values. The multiple linear regression analysis can be used to get point estimates.



Formula and Calculation of Multiple Linear Regression

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

where, for $i = n$ observations:

y_i = dependent variable

x_i = explanatory 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)

The formula used in this model of Multiple linear Regression is as follows.

- ***y=a+XTemperature*x1+XFuel_Price*x2+XMarkDown1*x3+XMarkDown2*x4+XMarkDown3*x5+XMarkDown4*x6+XMarkDown5*x7+XCPI*x8+XUnemployment*x9***

WE CAN PREDICT THE WEEKLY SALES BY PUTTING VALUES in x1 …. x9 and obtain a value for weekly sales.

- **Pearson correlation coefficient**

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier *product-moment* in the name.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:

- cov is the covariance
- ϭx is the standard deviation of X
- ϭy is the standard deviation of Y

In this analysis for the pearson correlation coefficient **X=sales and Y=Holiday**.

3.2 FLOWCHART

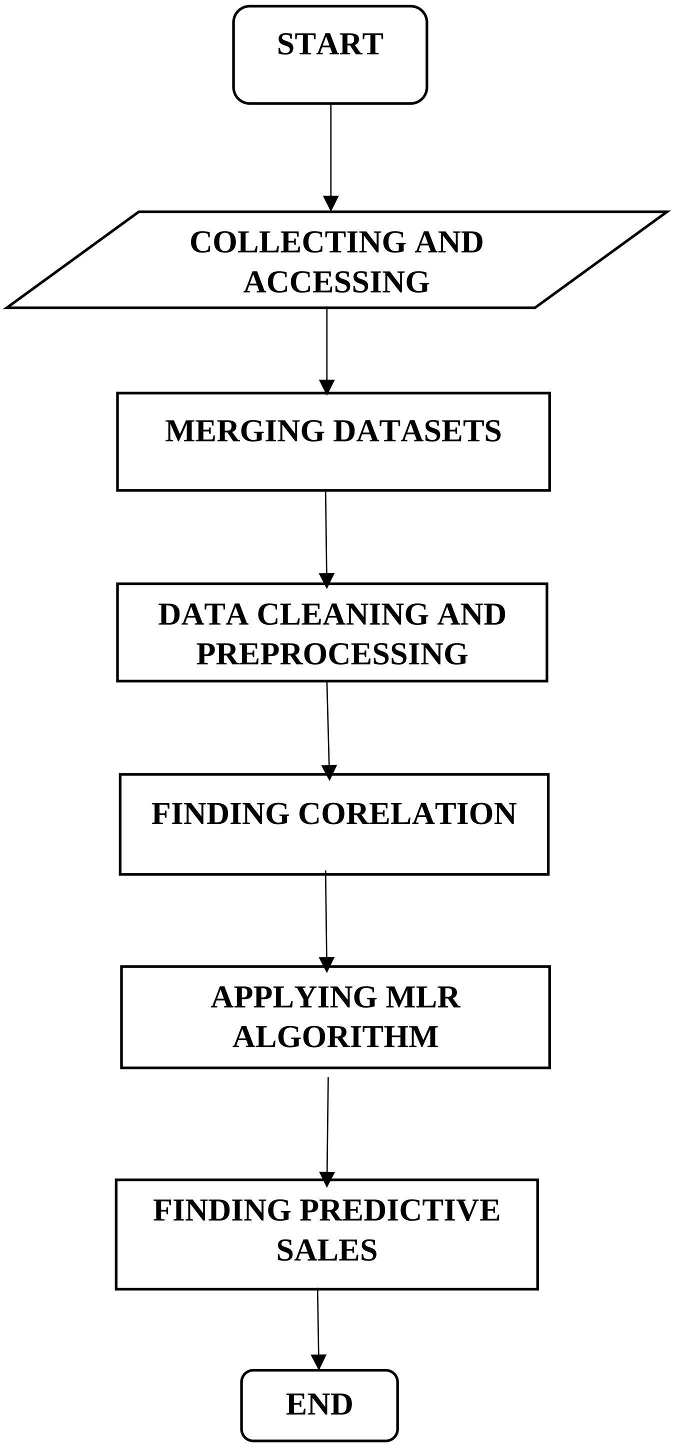


Fig 3.2.1: Steps for finding predictive sales from a dataset.

4.CODE

```
stores<-read.csv(file.choose())

View(stores)
stores_df <-stores
sales <-read.csv(file.choose())
sales_df <- sales
View(sales)
test1 <- read.csv(file.choose(),header = TRUE, check.names = TRUE)
features_df <- test1
pre_final_df <- merge(stores_df,sales_df,by = "Store")
head(pre_final_df)
final_df <- merge(pre_final_df,features_df,by= c("Store","Date","IsHoliday"))
head(final_df)
final_df$IsHoliday [final_df$IsHoliday == "true"] <- 1
final_df$IsHoliday [final_df$IsHoliday == "false"] <-0
head(final_df)
final_df[is.na(final_df)] <- 0
final_df
subset1 <-subset(final_df$Date,final_df$Weekly_Sales<0)
subset2 <-subset(final_df,select =
c("Size","Weekly_Sales","Temperature","Fuel_Price","MarkDown1","MarkDown2","MarkDown3","MarkDown4","MarkDown5","CPI","Unemployment"))

cor(final_df$Weekly_Sales,final_df$IsHoliday,use="everything",method="pearson")
subset1 <- subset(final_df$Date,final_df$Weekly_Sales<0)
mean_markdown1 <- mean(final_df$MarkDown1)
mean_markdown2 <- mean(final_df$MarkDown2)
mean_markdown3 <- mean(final_df$MarkDown3)
mean_markdown4 <- mean(final_df$MarkDown4)
```

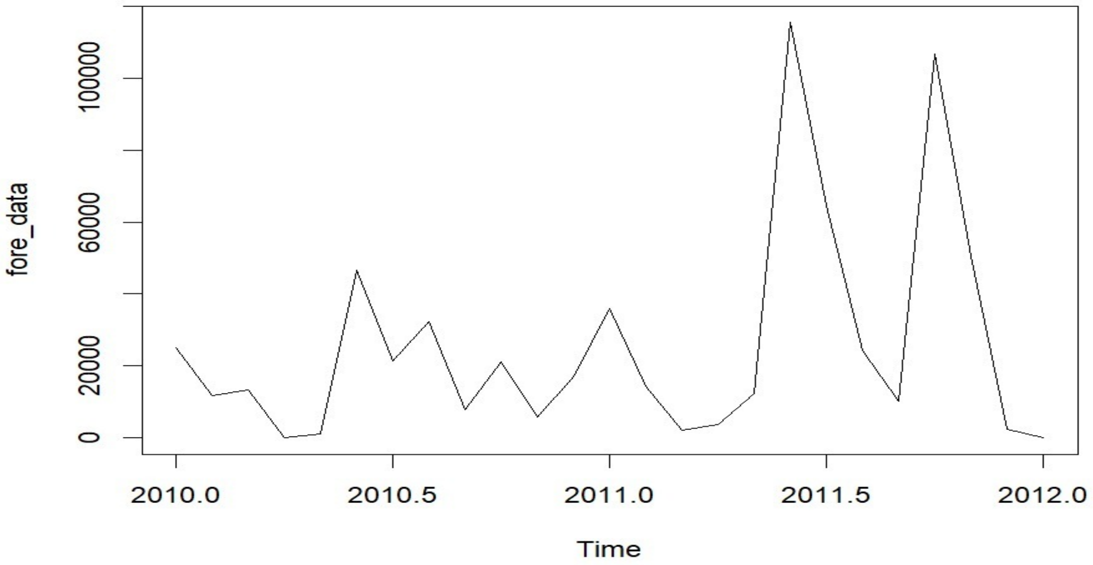
```
mean_markdown5 <- mean(final_df$MarkDown5)
final_markdown <- mean_markdown1 + mean_markdown2 +
mean_markdown3 + mean_markdown4 + mean_markdown5
final_markdown

average_final_markdown <-final_markdown/5
average_final_markdown

install.packages("classInt")
library(classInt)
bin_data <- final_df$Weekly_Sales
bin_data
classIntervals(bin_data,5,style = "equal")

classIntervals(bin_data,5,style="quantile")
fore_data <- ts(final_df$Weekly_Sales, start=2010, end=2012,frequency=12)
plot(fore_data)

install.packages("forecast")
library(forecast)
fore_data <- ts(final_df$Weekly_Sales, start=2010, end=2012,frequency=12)
plot(fore_data)
hw <- HoltWinters(fore_data)
plot(hw)

install.packages("ggplot2")
library(ggplot2)
install.packages("reshape")
library(reshape)
install.packages("dplyr")
library(dplyr)
```

```
subset2 <- subset(final_df,select=
c("Size","Weekly_Sales","Temperature","Fuel_price","MarkDown1","MarkDo
wn2","MarkDown3","MarkDown4","MarkDown5","CPI","Unemployment"))

res <-cor(subset2)
install.packages("corrplot")
library(corrplot)
corrplot(res,type = "upper",order = "hclust",tl.col = "black",tl.srt = 45)

col <-colorRampPalette(c("blue","white","red"))(20)
heatmap(x = res, col = col, symm = TRUE )



input<-
final_df[c("Weekly_Sales","Temperature","Fuel_price","MarkDown1","Mark
Down2","MarkDown3","MarkDown4","MarkDown5","CPI","Unemployment"
),]
head(input)
model <-
lm(Weekly_Sales~Temperature+Fuel_Price+MarkDown1+MarkDown2+Mark
Down3+MarkDown4+MarkDown5+CPI+Unemployment,data= final_df)
print(model)

cat("# # # # The Coefficeint value # # #","\n")
a <- coef(model)[1]
print(a)

XTemperature <- coef(model)[2]
XFuel_Price <- coef(model)[3]
XMarkDown1 <- coef(model)[4]
XMarkDown2 <- coef(model)[5]
XMarkDown3 <- coef(model)[6]
XMarkDown4 <- coef(model)[7]
```

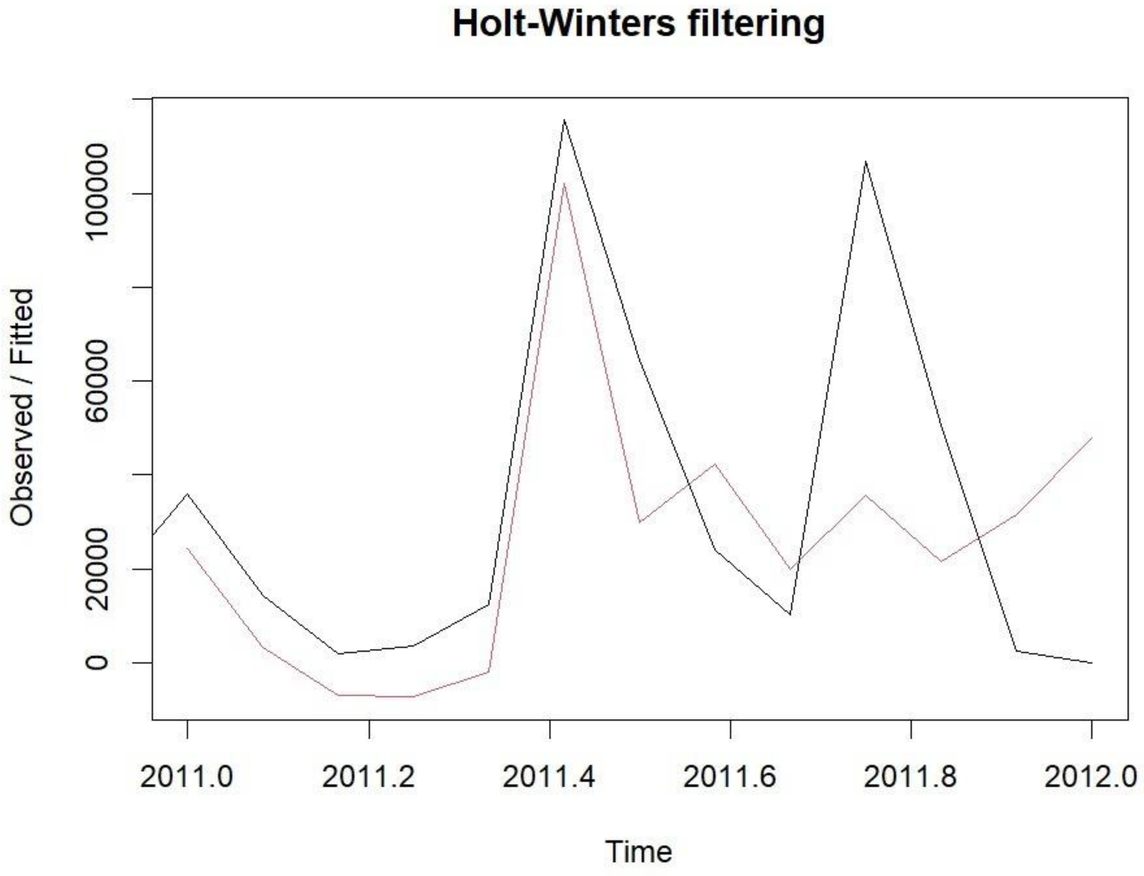
```
XMarkDown5 <- coef(model)[8]
XCPI <- coef(model)[9]
XUnemployment <- coef(model)[10]

print(XTemperature)
print(XFuel_Price)
print(XMarkDown1)
print(XMarkDown2)
print(XMarkDown3)
print(XMarkDown4)
print(XMarkDown5)
print(XCPI)
print(XUnemployment)

x1= 41.17
x2 = 2.562
x3 = 16305.11
x4 = 3551.41
x5=16.16
x6 = 3611.60
x7 = 1240.2
x8 = 220.806
x9 = 7.931

y=a+XTemperature*x1+XFuel_Price*x2+XMarkDown1*x3+XMarkDown2*x4+XMarkDown3*x5+XMarkDown4*x6+XMarkDown5*x7+XCPI*x8+XUnemployment*x9
print(y)
```

5.RESULTS

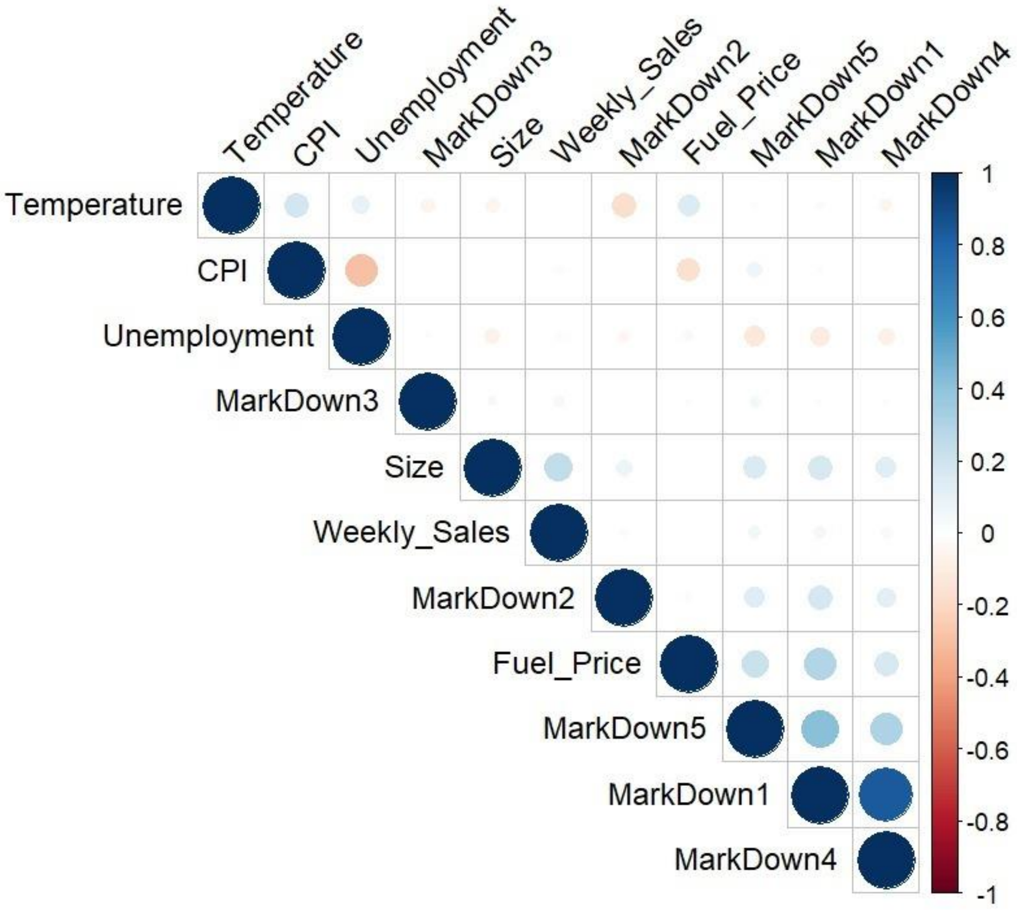
1)Result for Time series:



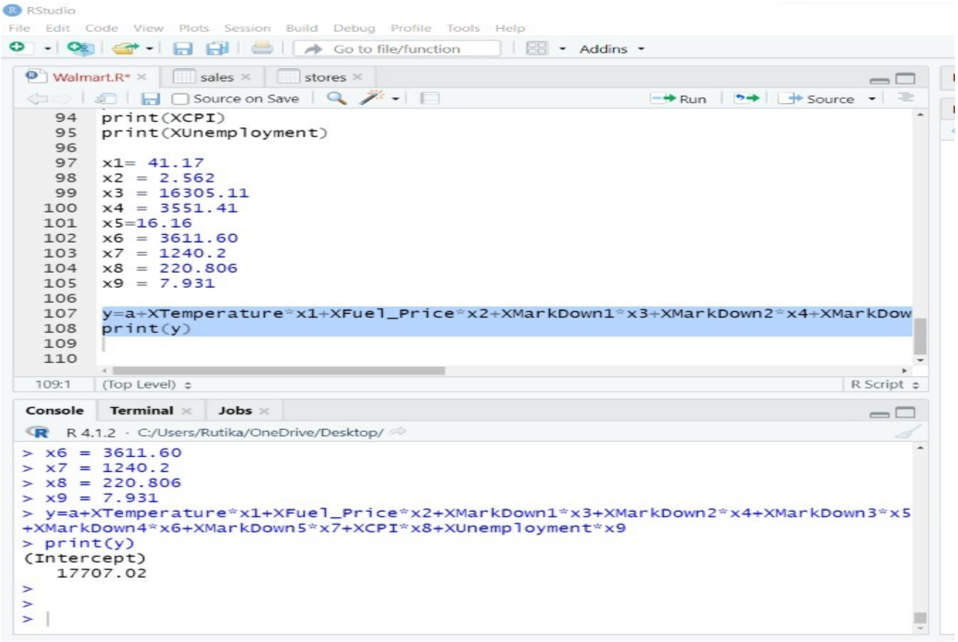
2)Result for Holt-winters Filtering.



3)Corelation Plot:



5)Result for Y-intercept-(Future week’s sales prediction):



```r
print(XCPI)
print(XUnemployment)

x1= 41.17
x2 = 2.562
x3 = 16305.11
x4 = 3551.41
x5=16.16
x6 = 3611.60
x7 = 1240.2
x8 = 220.806
x9 = 7.931

y=a+XTemperature*x1+XFuel_Price*x2+XMarkDown1*x3+XMarkDown2*x4+XMarkDow
print(y)
```

```
> x6 = 3611.60
> x7 = 1240.2
> x8 = 220.806
> x9 = 7.931
> y=a+XTemperature*x1+XFuel_Price*x2+XMarkDown1*x3+XMarkDown2*x4+XMarkDown3*x5
+XMarkDown4*x6+XMarkDown5*x7+XCPI*x8+XUnemployment*x9
> print(y)
(Intercept)
   17707.02
>
>
> |
```

6.CONCLUSION

In conclusion, Wal-Mart is the number one retailer in the USA and it also operates in many other countries all around the world and is moving into new countries as years pass by. There, are other companies who are constantly rising as well and would give Walmart a tough competition in the future if Walmart does not stay to the top of their game. In order to do so, they will need to understand their business trends, the customer needs and manage the resources wisely. In this era when the technologies are reaching out to new levels, Big Data is taking over the traditional method of managing and analyzing data. These technologies are constantly used to understand complex datasets in a matter of time with beautiful visual representations. Through observing the history of the company's datasets, clearer ideas on the sales for the previous years was realized which will be very helpful to the company on its own. Additionally, seasonality trend and randomness and future forecasts will help to analyse sale drops which the companies can avoid by using a more focused and efficient tactics to minimize the sale drop and maximize the profit and remain in competition.

7. REFERENCES

1. https://medium.datadriveninvestor.com/walmart-sales-data-analysis-sales-prediction-using-multiple-linear-regression-in-r-programming-adb14afd56fb

2.M. Franco-Santos and M. Bourne, “The impact of performance targets on behaviour: a close look at sales force contexts,” Research executive summaries series, vol. 5, 2009.

[2] D. Silverman, Interpreting Qualitative Data: Methods for Analyzing Talk, Text and Interaction 3rd Ed. Text and Interaction, Sage Publications Ltd: Methods for Analyzing Talk, 2006.

[3] UBM. (2003) Big Data analytics: Descriptive vs. predictive vs. prescriptive. [Accessed 17 September 2017]. [Online].