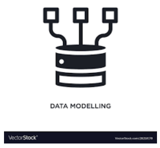
# PROJECT REPORT

## RETAIL ANALYSIS OF WALMART SALES

## &

## SALES FORECASTING

Walmart


By- Ruben Easo Thomas

# Contents



# Introduction

Wal-mart Stores, Inc is an American multinational retail corporation that operates a chain of discount department stores and wholesale warehouse stores. Headquartered in Bentonville, Arkansas , USA, the company was founded by Sam Walton in 1962 and incorporated in 1968. It has 11,000 stores in 28 countries under 65 banners. It operates under the name of Walmart in the USA and Canada. It has bases of operations in Central American region, Brazil, Argentina and Chile. Walmart is the world's largest company by revenue, with US$548.743 billion, according to the Fortune Global 500 list in 2020. It is also the largest private employer in the world with 2.2 million employees. It is a publicly traded family-owned business, as the company is controlled by the Walton family. Sam Walton's heirs own over 50 percent of Walmart through both their holding company Walton Enterprises and their individual holdings.Walmart was the largest United States grocery retailer in 2019, and 65 percent of Walmart's US$510.329 billion sales came from U.S. operations.

Walmart was listed on the New York Stock Exchange in 1972. By 1988, it was the most profitable retailer in the U.S., and it had become the largest in terms of revenue by October 1989. The company was originally geographically limited to the South and lower Midwest, but it had stores from coast to coast by the early 1990s. Sam's Club opened in New Jersey in November 1989, and the first California outlet opened in Lancaster, in July 1990. A Walmart in York, Pennsylvania, opened in October 1990, the first main store in the Northeast.

# Business Scenario

Walmart stores decided that they would like to predict the sales and demand accurately. There are certain events and holidays which impact sales on each day. There are sales data available for 45 stores of Walmart. The business is facing a challenge due to unforeseen demands and runs out of stock some times, due to the inappropriate machine learning algorithm.
An ideal ML algorithm will predict demand accurately and ingest factors like economic conditions including CPI, Unemployment Index, etc.

Walmart runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of all, which are the Super Bowl, Labour Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks. Part of the challenge presented by this competition is modeling the effects of markdowns on these holiday weeks in the absence of complete/ideal historical data. Historical sales data for 45 Walmart stores located in different regions are available.

# Data Overview

## Dataset Description

This is the historical data which covers sales from 2010-02-05 to 2012-11-01, in the file Walmart_Store_sales. Within this file you will find the following fields:

- Store - the store number
- Date - the week of sales
- Weekly_Sales - sales for the given store
- Holiday_Flag - whether the week is a special holiday week 1 – Holiday week 0 – non-holiday week
- Temperature - Temperature on the day of sale
- Fuel_Price - Cost of fuel in the region
- CPI – Prevailing consumer price index
- Unemployment - Prevailing unemployment rate

## Holiday Events

Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13
Labour Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13
Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13
Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

# Procedural Analysis

In the Walmart_sales_csv dataset, we have the various parameters such as Store No., Holiday_Flag, Temperature, Fuel_price, CPI and Unemployment. We try to find out the impact and influence these factors have on the Weekly_sales. The objective is to develop a statistical model based on the dataset available. We are using the historical sales data of 45 Walmart stores located in different regions to predicting the Weekly sales of each store.

We load the dataset and the libraries:

```
#Load the libraries
library(dplyr)
library(ggplot2)
library(caTools)
library(MLmetrics)
library(corrplot)

#Load the dataset
dataset <- read.csv("Walmart_store_sales.csv")
dataset1 <- dataset #Copying into another dataframe for analysis
```

Let us visualize the sales with respect to each variable and understand their influence.

```
#Visualization of all independent variables with respect to the target variable(Weekly_Sales)
par(mfrow=c(3,2))  #Arranges the plots in 3 rows and 2 columns
for(i in 4:8)
{
  plot(dataset[,i],
     dataset$Weekly_Sales,
```

```
    main=names(dataset[i]),
     ylab=names(dataset$Weekly_Sales),
     xlab="", col='indianred4')
}
```

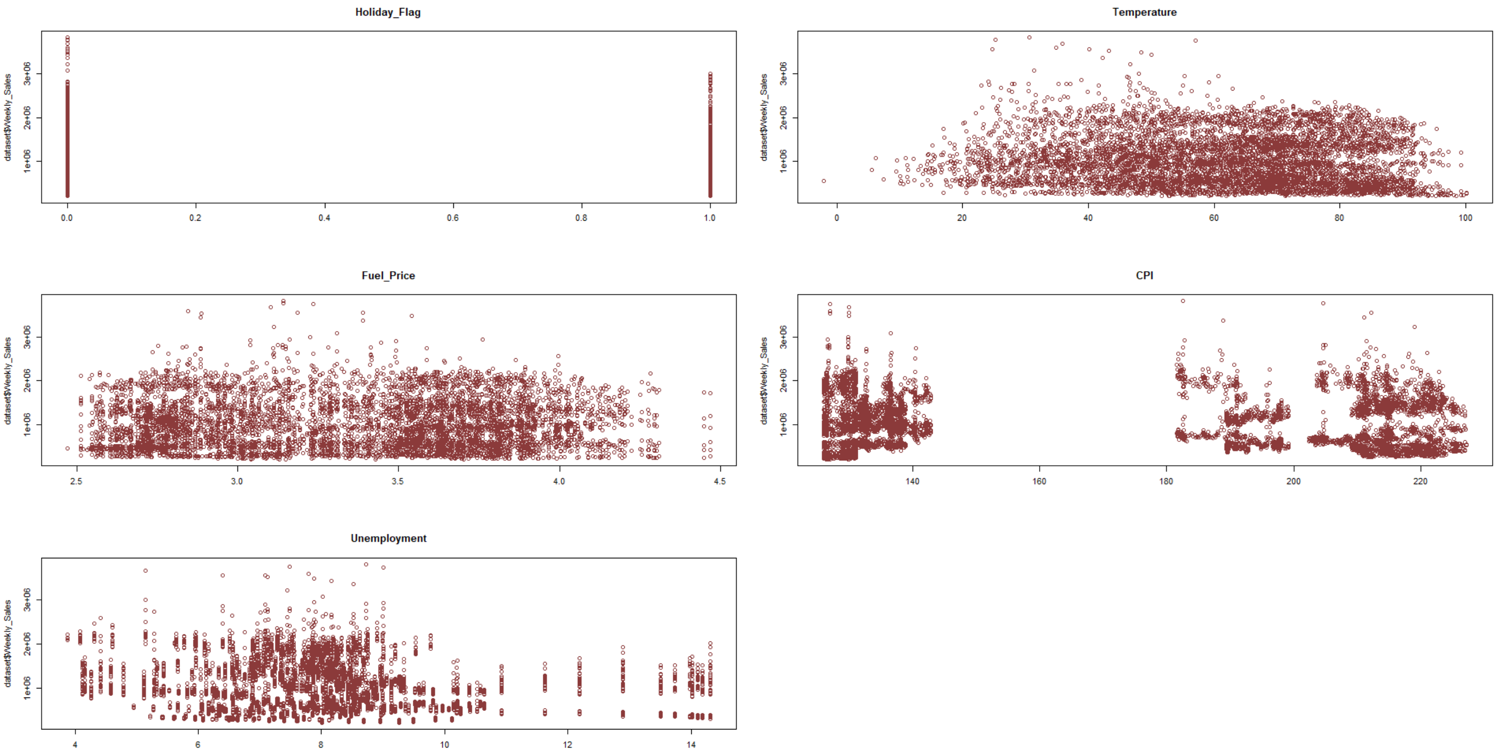



Fig 1: Variation of sales with respect to each variable in the dataset

As we notice from the plots (Fig 1), holiday Flag is a binary variable with only 0 and 1 values, while the remaining variables are continuous. Since we aren't able to get a clear idea of the impact of the variables on the weekly sales, let us do some analytical tasks to deepen our understanding of the data.

## Basic Statistical Tasks

1. **Which store has maximum sales?**
   To find the store with maximum sales, the total sales was found for each store and the dataset is sorted according to maximum sales and the store number is obtained.

```
#Task 1: Determination of which store has the maximum sales
names(dataset1)
```

stat <- summarize (group_by(dataset1, Store),sales_sum = sum(Weekly_Sales)) *#Getting the sum for each store in a seperate dataframe*
max_sum <- stat[which.max(stat$sales_sum),] *#returns the store with max sales*
ggplot(data=stat, aes(x=Store, y=sales_sum)) +
  geom_bar(stat="identity", fill="darkblue")+
  ggtitle("Weekly sales for each Store") +
  geom_text(aes(label=sales_sum), vjust= -1, size=3) +
  theme_minimal()  *#Displays a bar chart of the weekly sales of each store*

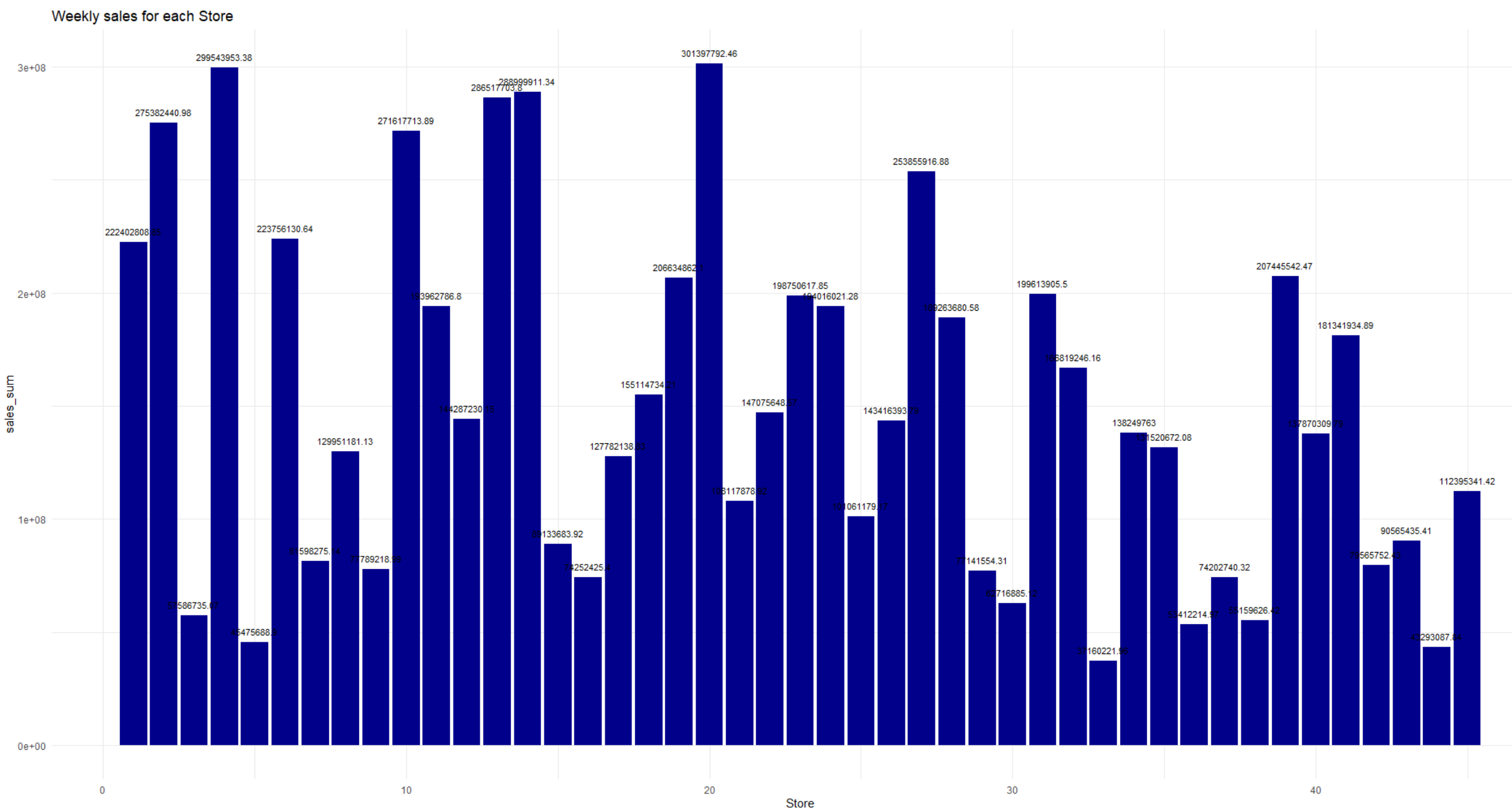



Fig. 2 Weekly sales for each store


Fig.3  Maximum value of sales

From the above results, we can see that the **store no. 20** has got the maximum sales of **$ 301397792**.

2. **Which store has maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of mean to standard deviation?**
   To find the store with the max std deviation of sales or the maximum variation in sales with each week, the std deviation for the sales is calculated for each store and the dataset is arranged in descending order . The store with the highest standard deviation is extracted. The coefficient of variance is also calculated by the std deviation of sales in the store by the mean sales of that store.

```
#Task 2: Determination of which store has the maximum std deviation
between the sales of each week and finding the coeff of variance
names(dataset1)
salessd <- summarise(group_by(dataset1,Store),sales_sd =
sd(Weekly_Sales), sales_mean = mean(Weekly_Sales)) #Getting the sd
for each store in a seperate dataframe
stat <- merge(stat,salessd,by ='Store',all.x = TRUE)
max_sd <- stat[which.max(stat$sales_sd),] #returns the store with max
variation in sales
ggplot(data=stat, aes(x=Store, y=sales_sd)) +
  geom_bar(stat="identity", fill="orange")+
  geom_text(aes(label=sales_sd), vjust= -1, size=3) +
  ggtitle("Variation of Weekly sales for each Store") +
  theme_minimal()       #Displays a bar chart of the weekly sales of each
store
stat$coeff_var <- stat$sales_sd/stat$sales_mean #finding the coeff of
variance
```

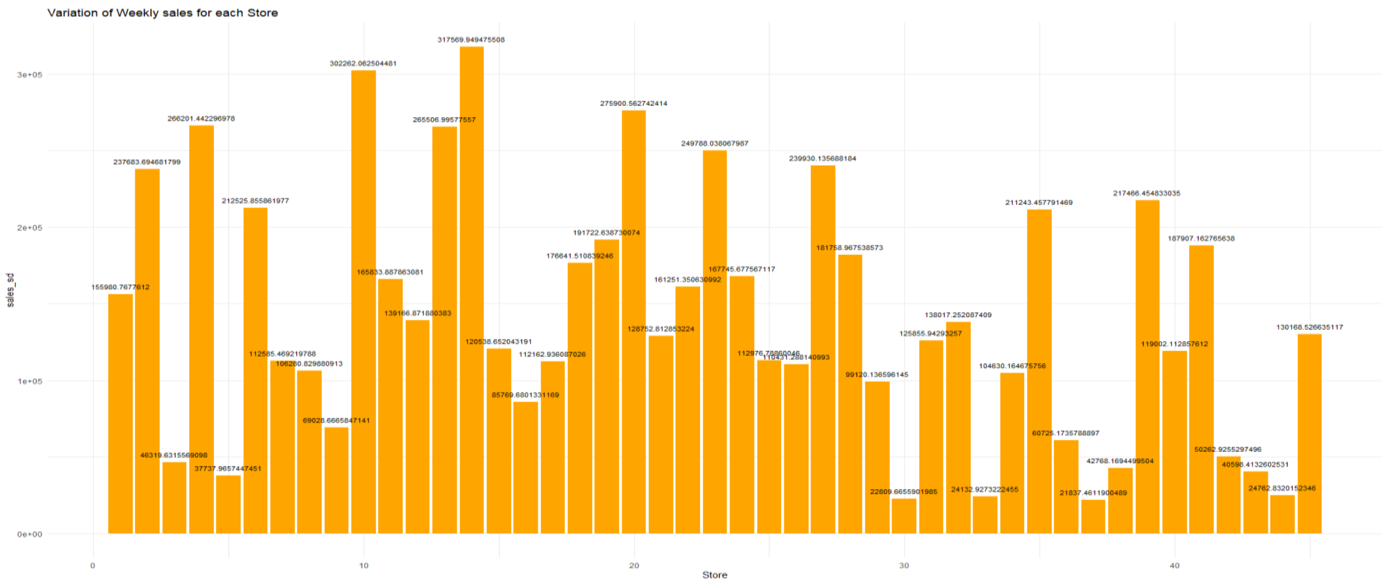



Fig 4.  Variation in Weekly Sales for each store

```
> max_sd
   Store sales_sum sales_sd sales_mean
14    14 288999911 317569.9    2020978
> 
```

Fig 5. Store with Maximum standard deviation of sales

From the above results, we can see that the **store no. 14** has got the maximum variation in sales of **$ 317569**. The coefficient of variance are present in the newly created stat dataframe.



| | Store | sales_sum | sales_sd | sales_mean | coeff_var |
|---|---|---|---|---|---|
| 1 | 1 | 222402809 | 155980.77 | 1555264.4 | 0.10029212 |
| 2 | 2 | 275382441 | 237683.69 | 1925751.3 | 0.12342388 |
| 3 | 3 | 57586735 | 46319.63 | 402704.4 | 0.11502141 |
| 4 | 4 | 299543953 | 266201.44 | 2094713.0 | 0.12708254 |
| 5 | 5 | 45475689 | 37737.97 | 318011.8 | 0.11866844 |
| 6 | 6 | 223756131 | 212525.86 | 1564728.2 | 0.13582286 |
| 7 | 7 | 81598275 | 112585.47 | 570617.3 | 0.19730469 |
| 8 | 8 | 129951181 | 106280.83 | 908749.5 | 0.11695283 |
| 9 | 9 | 77789219 | 69028.67 | 543980.6 | 0.12689547 |
| 10 | 10 | 271617714 | 302262.06 | 1899424.6 | 0.15913349 |
| 11 | 11 | 193962787 | 165833.89 | 1356383.1 | 0.12226183 |
| 12 | 12 | 144287230 | 139166.87 | 1009001.6 | 0.13792532 |
| 13 | 13 | 286517704 | 265507.00 | 2003620.3 | 0.13251363 |
| 14 | 14 | 288999911 | 317569.95 | 2020978.4 | 0.15713674 |
| 15 | 15 | 89133684 | 120538.65 | 623312.5 | 0.19338399 |
| 16 | 16 | 74252425 | 85769.68 | 519247.7 | 0.16518065 |
| 17 | 17 | 127782139 | 112162.94 | 893581.4 | 0.12552067 |
| 18 | 18 | 155114734 | 176641.51 | 1084718.4 | 0.16284550 |
| 19 | 19 | 206634862 | 191722.64 | 1444999.0 | 0.13268012 |
| 20 | 20 | 301397792 | 275900.56 | 2107676.9 | 0.13090269 |
| 21 | 21 | 108117879 | 128752.81 | 756069.1 | 0.17029239 |
| 22 | 22 | 147075649 | 161251.35 | 1028501.0 | 0.15678288 |
| 23 | 23 | 198750618 | 249788.04 | 1389864.5 | 0.17972115 |
| 24 | 24 | 194016021 | 167745.68 | 1356755.4 | 0.12363738 |
| 25 | 25 | 101061179 | 112976.79 | 706721.5 | 0.15986040 |
| 26 | 26 | 143416394 | 110431.29 | 1002911.8 | 0.11011066 |
| 27 | 27 | 253855917 | 239930.14 | 1775216.2 | 0.13515544 |
| 28 | 28 | 189263681 | 181758.97 | 1323522.2 | 0.13732974 |
| 29 | 29 | 77141554 | 99120.14 | 539451.4 | 0.18374247 |
| 30 | 30 | 62716885 | 22809.67 | 438579.6 | 0.05200804 |
| 31 | 31 | 199613906 | 125855.94 | 1395901.4 | 0.09016105 |
| 32 | 32 | 166819246 | 138017.25 | 1166568.2 | 0.11831049 |
| 33 | 33 | 37160222 | 24132.93 | 259861.7 | 0.09286835 |

Fig .6 : Coefficient of variation calculation for each store

| | | | | | |
|---|---|---|---|---|---|
| 34 | 34 | 138249763 | 104630.16 | 966781.6 | 0.10822524 |
| 35 | 35 | 131520672 | 211243.46 | 919725.0 | 0.22968111 |
| 36 | 36 | 53412215 | 60725.17 | 373512.0 | 0.16257891 |
| 37 | 37 | 74202740 | 21837.46 | 518900.3 | 0.04208412 |
| 38 | 38 | 55159626 | 42768.17 | 385731.7 | 0.11087545 |
| 39 | 39 | 207445542 | 217466.45 | 1450668.1 | 0.14990779 |
| 40 | 40 | 137870310 | 119002.11 | 964128.0 | 0.12342978 |
| 41 | 41 | 181341935 | 187907.16 | 1268125.4 | 0.14817711 |
| 42 | 42 | 79565752 | 50262.93 | 556403.9 | 0.09033533 |
| 43 | 43 | 90565435 | 40598.41 | 633324.7 | 0.06410363 |
| 44 | 44 | 43293088 | 24762.83 | 302748.9 | 0.08179331 |
| 45 | 45 | 112395341 | 130168.53 | 785981.4 | 0.16561273 |

Here from the table we can see that store no. 35 has got the maximum level of variability of sales across the average sales of that store.

3. **Which store/s has good quarterly growth rate in Q3'2012?**
   In order to achieve this, the month and year column are extracted from the date. The dataset is subsetted into two data frames which has the 2nd quarter (April, May, June) and the 3rd quarter(July, August and September) of 2012 respectively. The total sales of both the quarters are obtained for each store and merged. Growth rate column is calculated by:

$$\text{Growth rate} = \left(\frac{\text{Q3 sales 2012} - \text{Q2 sales 2012}}{\text{Q3 sales 2012}}\right) \times 100$$

*#Task 3: Determination of which store has good quaterly growth for the quarter Q3-2012*

```
dataset1$Month <- as.integer(substr(dataset1$Date,4,5))
dataset1$Year <- as.integer(substr(dataset1$Date,7,10))
q3 <- subset(dataset1,Year == 2012 & (Month == 7 | Month == 8 | Month ==
9))
q2 <- subset(dataset1,Year == 2012 & (Month == 4 | Month == 5| Month ==
6))
q3_sales <- summarise(group_by(q3,Store),Q3_sales =
sum(Weekly_Sales)) #Getting the sum for each store for third quarter
```

q2_sales <- summarise(group_by(q2,Store),Q2_sales = sum(Weekly_Sales)) *#Getting the sum for each store for second quarter*

q3_sales <- merge(q3_sales,q2_sales,by = "Store",all.x = TRUE)

q3_sales$netgrowth <- ((q3_sales$Q3_sales - q3_sales$Q2_sales)/q3_sales$Q3_sales)*100 *#Obtaining the net growth of each store from second to third quarter*

View(subset(q3_sales,netgrowth > 0)) *#Filter the data with growth rate greater than 0 (positive)*

| | Store | Q3_sales | Q2_sales | netgrowth |
|---|---|---|---|---|
| 7 | 7 | 8262787 | 7290859 | 11.7627149 |
| 16 | 16 | 7121542 | 6564336 | 7.8242281 |
| 23 | 23 | 18641489 | 18488883 | 0.8186381 |
| 24 | 24 | 17976378 | 17684219 | 1.6252374 |
| 26 | 26 | 13675692 | 13155336 | 3.8049727 |
| 35 | 35 | 11322421 | 10838313 | 4.2756590 |
| 39 | 39 | 20715116 | 20214128 | 2.4184647 |
| 40 | 40 | 12873195 | 12727738 | 1.1299280 |
| 41 | 41 | 18093844 | 17659943 | 2.3980602 |
| 44 | 44 | 4411251 | 4306406 | 2.3767719 |

Fig .7 Stores which have a positive growth rate in Q3-2012

From the results taken, it is seen that **Stores 7,16,23,24,26,35,39,40,41** and **44** have had a positive growth rate with Store 7 having the highest growth rate between third quarter and second quarter.

4. **Some holidays have a negative impact on sales. Find out holidays which have higher sales than the mean sales in non-holiday season for all stores together.** To find out which holidays have a higher sales than non holiday seasons, a dataframe is created with the specified holidays and merged with the original dataset. The dataset is subsetted on the bases of each holiday and the mean sales was calculated. Simultaneously the non- holiday season mean sales is calculated for all stores. A comparison is done between the mean sales of each holiday and the non- holiday.

*#Task 4: Find out which holiday period has a positive impact and has higher sales than the mean sales in non holiday season*

```
holiday_df <- data.frame(Date = c("12-02-2010", "11-02-2011", "10-02-2012", "8-02-2013", "10-09-2010", "9-09-2011", "7-09-2012", "6-09-2013","26-11-2010", "25-11-2011", "23-11-2012", "29-11-2013", "31-12-2010", "30-12-2011", "28-12-2012", "27-12-2013"),
                         Name_Holiday = c("Super Bowl","Super Bowl","Super Bowl","Super Bowl","Labour Day","Labour Day","Labour Day","Labour Day","Thanksgiving","Thanksgiving","Thanksgiving","Thanksgiving","Christmas","Christmas","Christmas","Christmas")) #Creating a specific holiday dataframe
dataset1 <- merge(dataset1,holiday_df,by = 'Date',all.x = TRUE) #Left merge
non_holiday <- subset(dataset1,Holiday_Flag == 0)
noholiday_mean <- mean(non_holiday$Weekly_Sales)  #Mean sales for non holiday season
super_bowl <- subset(dataset1,Name_Holiday == "Super Bowl" )
super_bowl_mean <- mean(super_bowl$Weekly_Sales)  #Mean sales on super bowl days
labour <- subset(dataset1,Name_Holiday == "Labour Day" )
labour_mean <- mean(labour$Weekly_Sales)   #Mean sales on Labour days
thanksgiving <- subset(dataset1,Name_Holiday == "Thanksgiving" )
thanksgiving_mean <- mean(thanksgiving$Weekly_Sales)  #Mean sales on thanksgiving days
christmas <- subset(dataset1,Name_Holiday == "Christmas" )
christmas_mean <- mean(christmas$Weekly_Sales)  #Mean sales on Christmas days
holiday_df <- data.frame(Name = c("Super Bowl","Labour Day","Thanksgiving","Christmas"), Mean_sales = c(super_bowl_mean,labour_mean,thanksgiving_mean,christmas_mean))
#Creating a new data frame with mean sales of each holiday
holiday_df$Positive_Impact <- holiday_df$Mean_sales > noholiday_mean
#Checking which Holiday has a positive or negative impact
```

| | Name | Mean_sales | Positive_Impact |
|---|---|---|---|
| 1 | Super Bowl | 1079128.0 | TRUE |
| 2 | Labour Day | 1014097.7 | FALSE |
| 3 | Thanksgiving | 1471273.4 | TRUE |
| 4 | Christmas | 960833.1 | FALSE |

From the above observation table, it is noticed that Super Bowl and Thanksgiving Days have a higher sales output than non holiday seasonal working days.

```
#Visualizing the holiday sales plot which have higher impact than non-holiday seasonal sales
ggplot()+
  geom_bar(aes(x=holiday_df$Name, y = holiday_df$Mean_sales, fill = holiday_df$Positive_Impact),stat = "identity",position=position_dodge())+
  xlab("Holidays")+
  ylab("Mean sales")+
  ggtitle("Graphical analysis of holiday sales with non- holiday sales")+
  theme_minimal()
```

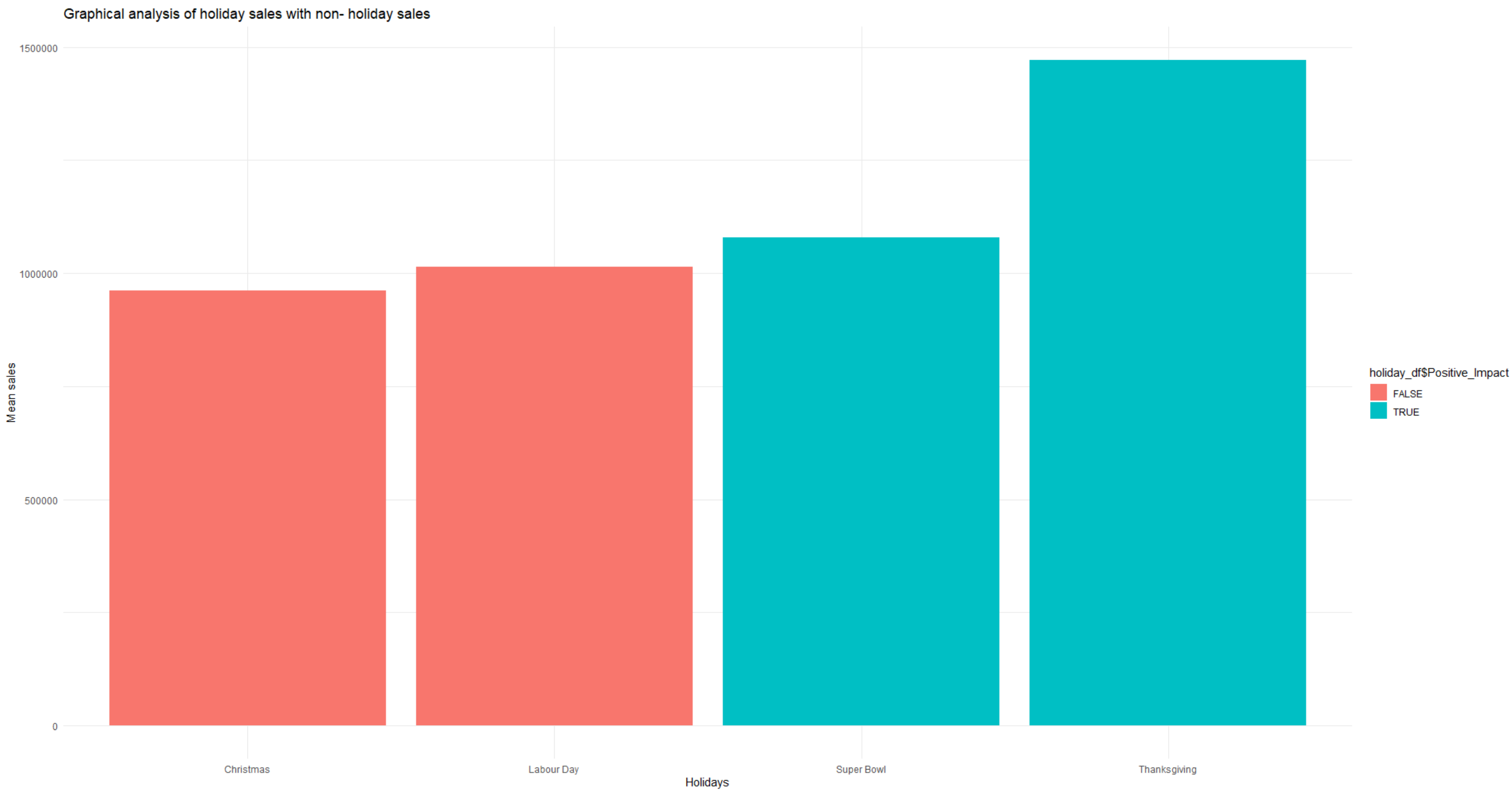



Fig 8 . Visualization of holiday sales with non holiday sales

5. **Provide a monthly and semester view of sales in units and give insights**
   Two bar plots are implemented for the analysis of monthly and semester view of sales. Monthly sales is the sales for every month while semesterly sales are the sales for every 6 months.

```
#Task 5:Monthly and semester view of sales in units
ggplot(data = dataset1, aes(x=Month,y= Weekly_Sales))+
  geom_bar(stat = "identity",fill = "red")+
  xlab("Month")+
  ylab("Sales")+
  ggtitle("Graphical analysis of Monthly sales")
  + theme_minimal()#Monthly Sales Visualization
dataset1$semester <- ifelse(dataset1$Month %in% c(1,2,3,4,5,6),1,2)
#Creating semester column
ggplot()+
  geom_bar(aes(x=dataset1$semester,y=dataset1$Weekly_Sales),stat =
"identity",fill = "green",width = 0.5)+
  xlab("Semesterwise") +
  ylab("Sales")+
  ggtitle("Graphical analysis of Semester Sales")+
   theme_minimal()     #Semesterly Sales Visualization
ggplot(data = dataset1, aes(x=Month,y= Temperature))+
  geom_bar(stat = "identity",fill = "steelblue")+
  xlab("Month")+
  ylab("Temeprature")+
  ggtitle("Average Monthly Temperature ")+
  theme_minimal()     #Monthly Temperature
```

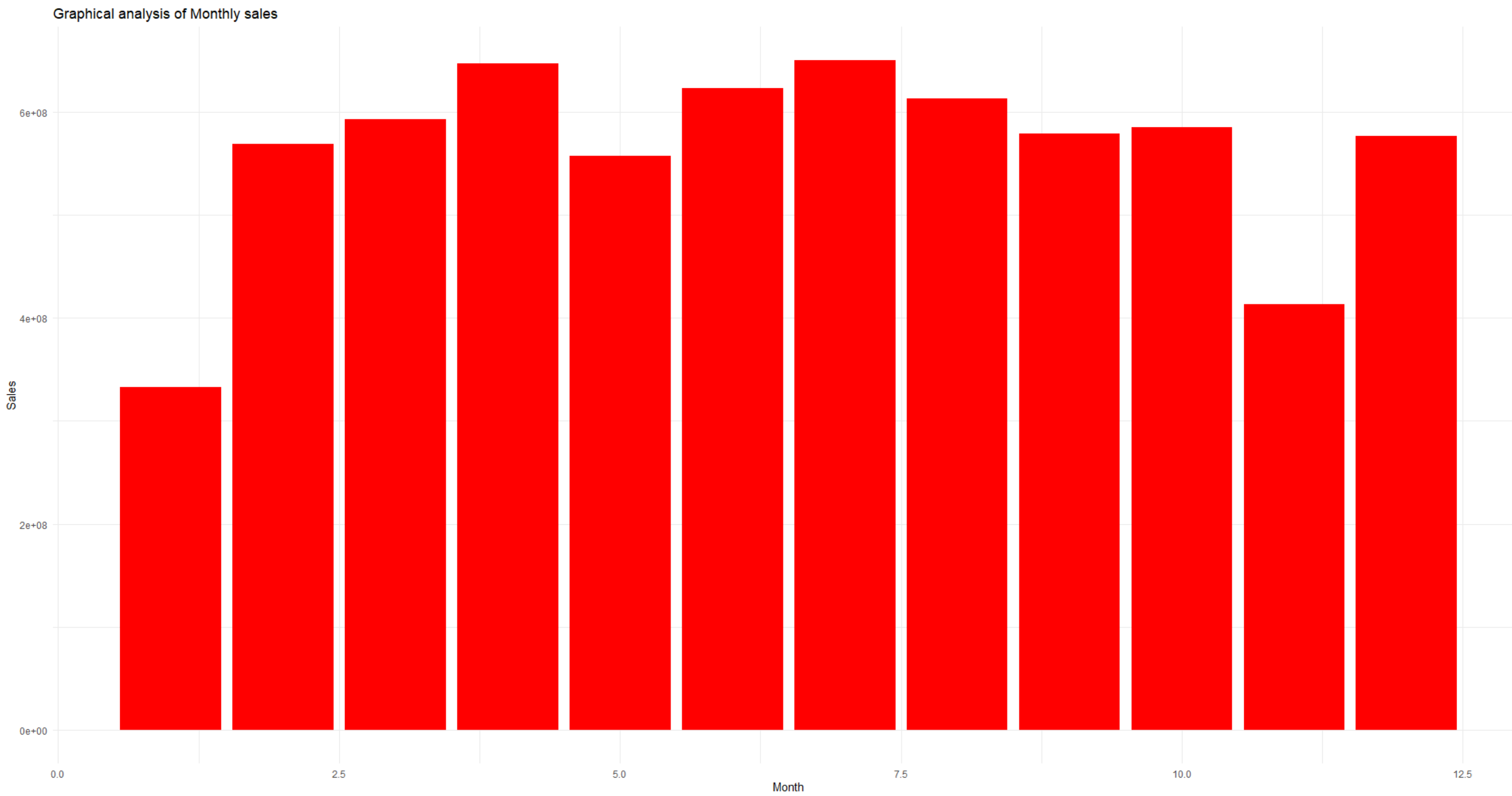



Fig 9. Monthly Sales Analysis

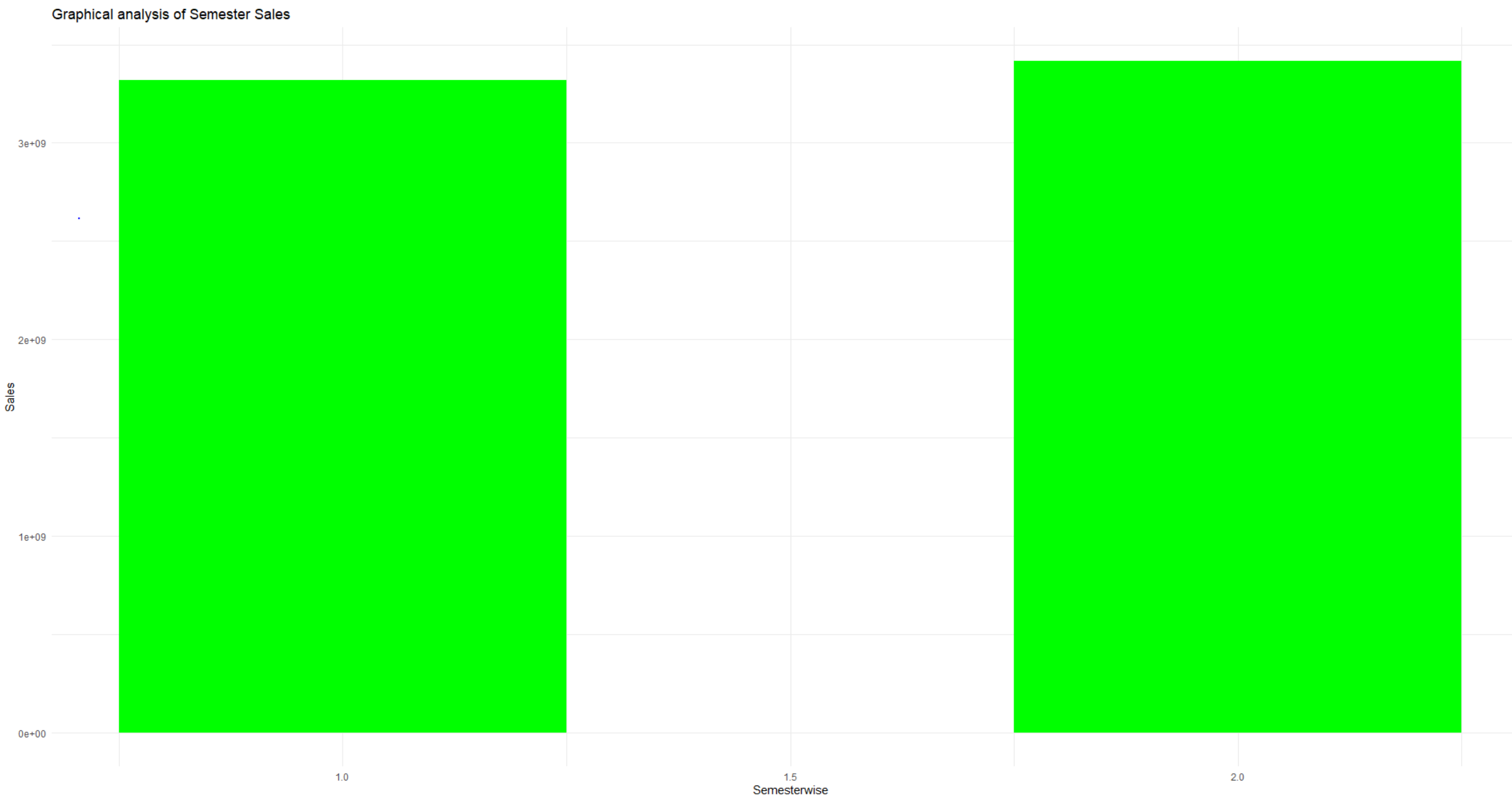



Fig 10. Semesterly Sales Analysis

From the above analysis we notice that the sales in both the the semesterly sales are almost relative with the second semester higher than the first semester. The reason most likely for the drop of sales in the first semester is because low sales generated in the month of January. This might be due to very low temperature in January which makes it difficult for customers to commute to the stores for shopping, hence reducing the sales.

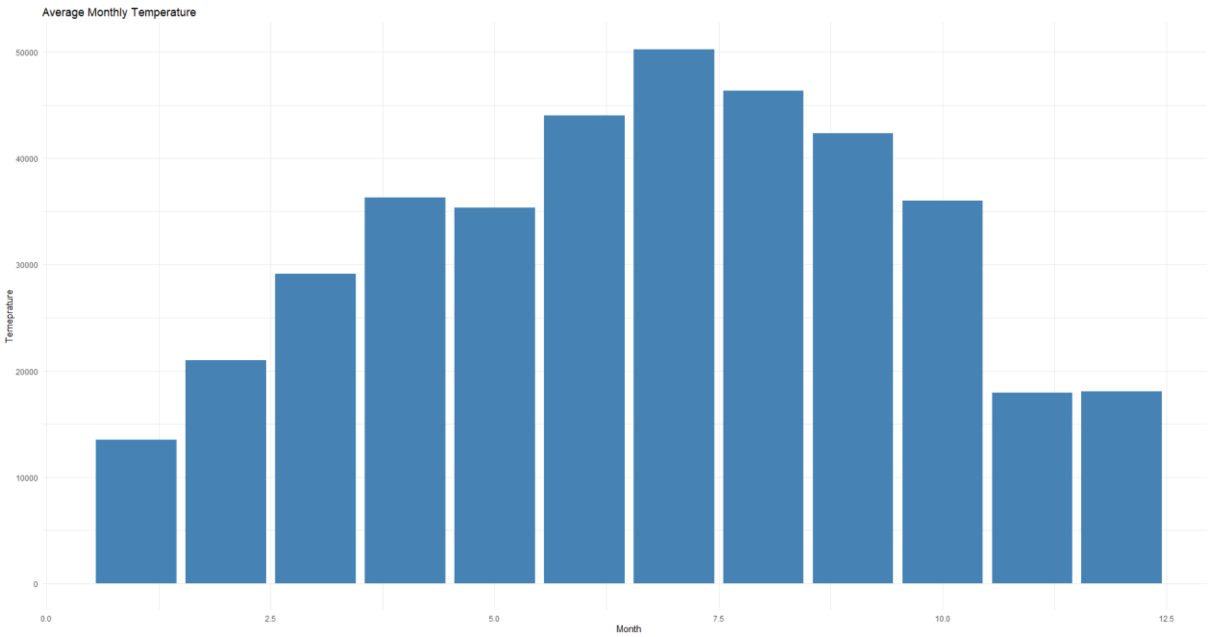



Fig. 11 Variation of temperature with each month

## Data Modelling

**Predict the weekly sales of the stores using a linear Regression model – Utilize variables like date and restructure dates as 1 for 5 Feb 2010 (starting from the earliest date in order). Hypothesize if CPI, unemployment, and fuel price have any impact on sales.**

First we need to restructure the dates starting as 1 for the $5^{th}$ Feb 2010 in a sequential order till the last date for all stores. To do that we extract the unique value of all dates in a new vector and we create another vector with the sequence from 1 till the length of the unique date vector and we create a dataframe and merge with it the original dataframe under "Week".We hence drop the date column as it will be of no further use to our model.

- We create a Null Hypothesis that the CPI, Unemployment and fuel price do not have any impact on the weekly sales

*##DATA MODELLING: LINEAR REGRESSION##*

*#Create the week column and drop the date column*

dataset <- dataset1 *#Copying the new analysed dataframe*

arrange(dataset,Store)

Date <- unique(dataset$Date)

Week <- seq(1:length(Date))

week_df <- data.frame(Date,Week)

dataset <- merge(dataset,week_df,by = "Date",all.x = TRUE)

dataset$Date <- NULL

arrange(dataset,Week)

We replace the categorical holiday NA values with 0 to help in the formation of dummy variables later.

*#Replace NA values in holiday with 0s*

dataset$Name_Holiday[is.na(dataset$Name_Holiday)] <- 0

View(dataset)

| | Store | Weekly_Sales | Holiday_Flag | Temperature | Fuel_Price | CPI | Unemployment | Month | Year | Name_Holiday | semester | Week |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1495064.8 | 0 | 59.17 | 3.524 | 214.8372 | 7.682 | 4 | 2011 | 0 | 1 | 1 |
| 2 | 2 | 1800171.4 | 0 | 55.43 | 3.524 | 214.4887 | 7.931 | 4 | 2011 | 0 | 1 | 1 |
| 3 | 3 | 374556.1 | 0 | 68.76 | 3.524 | 218.2114 | 7.574 | 4 | 2011 | 0 | 1 | 1 |
| 4 | 4 | 1900246.5 | 0 | 56.99 | 3.521 | 128.7199 | 5.946 | 4 | 2011 | 0 | 1 | 1 |
| 5 | 5 | 314316.5 | 0 | 61.50 | 3.524 | 215.4024 | 6.489 | 4 | 2011 | 0 | 1 | 1 |
| 6 | 6 | 1459276.8 | 0 | 62.25 | 3.524 | 216.3841 | 6.855 | 4 | 2011 | 0 | 1 | 1 |
| 7 | 7 | 513409.7 | 0 | 24.83 | 3.461 | 192.2692 | 8.595 | 4 | 2011 | 0 | 1 | 1 |
| 8 | 8 | 878762.3 | 0 | 49.86 | 3.524 | 218.2586 | 6.297 | 4 | 2011 | 0 | 1 | 1 |
| 9 | 9 | 520962.1 | 0 | 56.12 | 3.524 | 218.4452 | 6.380 | 4 | 2011 | 0 | 1 | 1 |
| 10 | 10 | 1827733.2 | 0 | 67.64 | 3.772 | 128.7199 | 8.494 | 4 | 2011 | 0 | 1 | 1 |
| 11 | 11 | 1258674.1 | 0 | 69.10 | 3.524 | 218.2114 | 7.574 | 4 | 2011 | 0 | 1 | 1 |
| 12 | 12 | 1005463.5 | 0 | 63.63 | 3.772 | 128.7199 | 13.736 | 4 | 2011 | 0 | 1 | 1 |
| 13 | 13 | 1864238.6 | 0 | 42.49 | 3.487 | 128.7199 | 7.193 | 4 | 2011 | 0 | 1 | 1 |
| 14 | 14 | 1869110.6 | 0 | 37.27 | 3.638 | 185.1790 | 8.521 | 4 | 2011 | 0 | 1 | 1 |
| 15 | 15 | 542556.1 | 0 | 30.34 | 3.811 | 134.0683 | 7.658 | 4 | 2011 | 0 | 1 | 1 |
| 16 | 16 | 459756.1 | 0 | 35.75 | 3.461 | 192.2692 | 6.339 | 4 | 2011 | 0 | 1 | 1 |
| 17 | 17 | 795859.2 | 0 | 39.38 | 3.487 | 128.7199 | 6.774 | 4 | 2011 | 0 | 1 | 1 |
| 18 | 18 | 938083.2 | 0 | 35.06 | 3.638 | 134.0683 | 8.975 | 4 | 2011 | 0 | 1 | 1 |

Fig 12. Replacement of NA values of holiday column with 0s

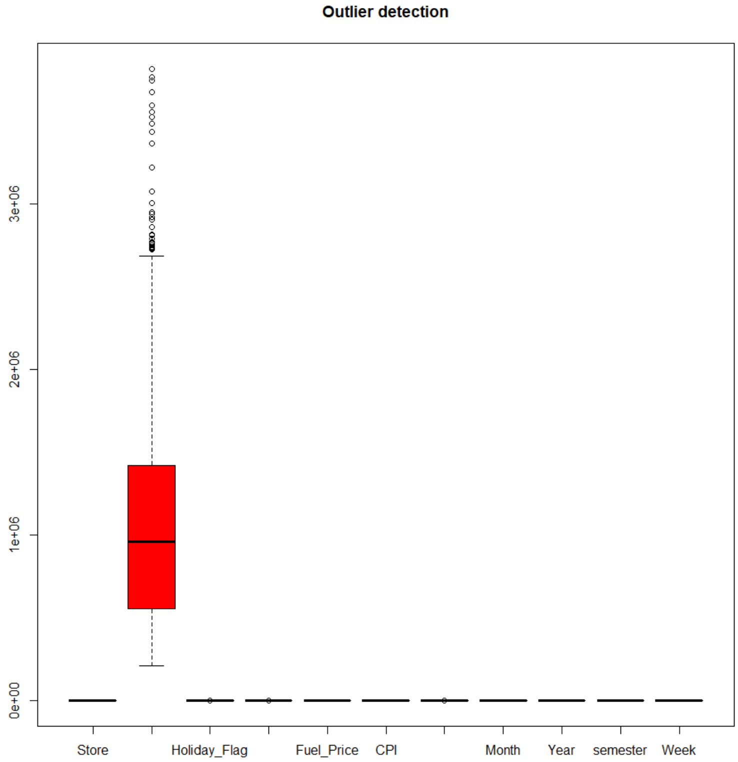
We checked if there is any missing values in the dataset. If there is any, we drop those records. As of now in the current data, there is no missing values present.

```
#Check for missing values

Noofna <- dim(dataset[is.na(dataset),])[1]

if(Noofna > 0 )

{

  cat("No.of missing values:",Noofna)

  cat("\n Removing missing values....")

  dataset <- dataset[complete.cases(datset),]

  cat("\n Removed succcessfully!")

}
```

```
> Noofna
[1] 0
>
```

Outliers cause a huge disturbance to the entire dataset and instigates a lot of errors contributed by variable leading the model to be less efficient. It disturbs the mean of the population sample a lot. Hence we need to deal with the outliers. Since the no. of outliers of weekly sales are low, we would drop them.

```
#Check for outliers

boxplot(dataset[,-10],main = "Outlier detection", col=c("blue","red")) #We remove the non numeric Holiday column from our boxplot analysis

#Getting rid of the outliers of weekly sales by some means

iqr <-IQR(dataset$Weekly_Sales)

quant <- quantile(dataset$Weekly_Sales)

ll <- round(quant[2] - iqr*1.5)

ul <- round(quant[4] + iqr*1.5)

#Extracting the outliers beyond the upper and lower limits

View(subset(dataset,Weekly_Sales >ul | dataset$Weekly_Sales < ll)) #only a few outliers are present, hence we drop them

dataset <- dataset[!(dataset$Weekly_Sales > ul | dataset$Weekly_Sales < ll),] #Outliers deleted!!

boxplot(dataset[,-10],main = "Outlier detection", col=c("blue","red"))
```



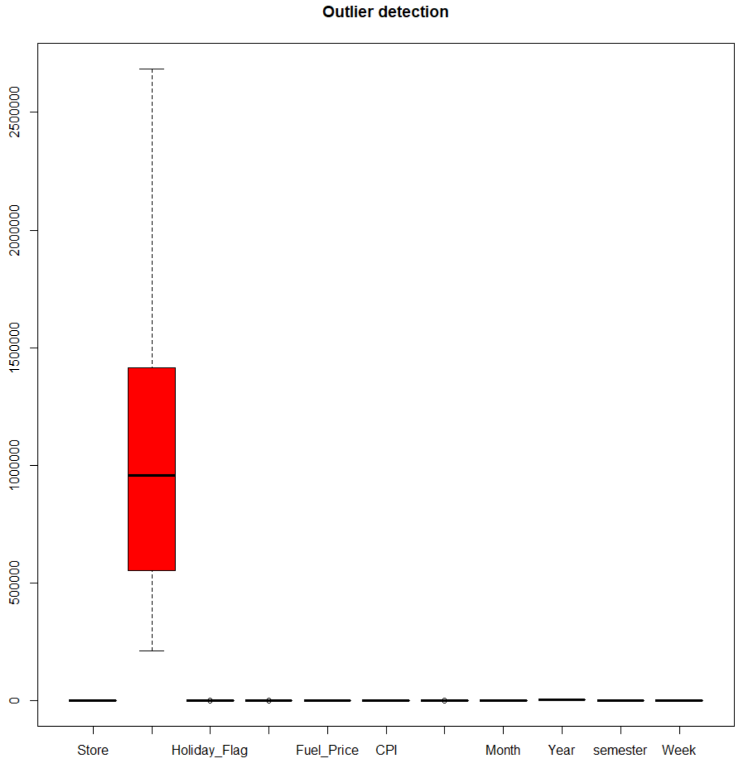




Fig 13. Before outlier removal                                                                After outlier removal

| | Store | Weekly_Sales | Holiday_Flag | Temperature | Fuel_Price | CPI | Unemployment | Month | Year | Name_Holiday | semester | Week |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2090 | 20 | 2752122 | 0 | 24.27 | 3.109 | 204.6877 | 7.484 | 12 | 2010 | 0 | 2 | 47 |
| 3334 | 4 | 2771397 | 0 | 36.44 | 3.149 | 129.8981 | 5.143 | 12 | 2011 | 0 | 2 | 75 |
| 3343 | 13 | 2760347 | 0 | 27.85 | 3.282 | 129.8981 | 6.392 | 12 | 2011 | 0 | 2 | 75 |
| 3350 | 20 | 2762817 | 0 | 37.16 | 3.413 | 212.0685 | 7.082 | 12 | 2011 | 0 | 2 | 75 |
| 3559 | 4 | 2740057 | 0 | 46.57 | 2.884 | 126.8795 | 7.127 | 12 | 2010 | 0 | 2 | 80 |
| 3565 | 10 | 2811647 | 0 | 59.15 | 3.125 | 126.8795 | 9.003 | 12 | 2010 | 0 | 2 | 80 |
| 3568 | 13 | 2771647 | 0 | 35.21 | 2.842 | 126.8795 | 7.795 | 12 | 2010 | 0 | 2 | 80 |
| 3569 | 14 | 2762861 | 0 | 30.51 | 3.140 | 182.5177 | 8.724 | 12 | 2010 | 0 | 2 | 80 |
| 3575 | 20 | 2819193 | 0 | 24.07 | 3.140 | 204.6321 | 7.484 | 12 | 2010 | 0 | 2 | 80 |
| 4817 | 2 | 3224370 | 0 | 46.66 | 3.112 | 218.9995 | 7.441 | 12 | 2011 | 0 | 2 | 108 |
| 4819 | 4 | 3676389 | 0 | 35.92 | 3.103 | 129.9845 | 5.143 | 12 | 2011 | 0 | 2 | 108 |
| 4825 | 10 | 3487987 | 0 | 48.36 | 3.541 | 129.9845 | 7.874 | 12 | 2011 | 0 | 2 | 108 |
| 4828 | 13 | 3556766 | 0 | 24.76 | 3.186 | 129.9845 | 6.392 | 12 | 2011 | 0 | 2 | 108 |
| 4829 | 14 | 3369069 | 0 | 42.27 | 3.389 | 188.9300 | 8.523 | 12 | 2011 | 0 | 2 | 108 |
| 4835 | 20 | 3555371 | 0 | 40.19 | 3.389 | 212.2360 | 7.082 | 12 | 2011 | 0 | 2 | 108 |
| 4842 | 27 | 2739020 | 0 | 41.59 | 3.587 | 140.5288 | 7.906 | 12 | 2011 | 0 | 2 | 108 |
| 5042 | 2 | 3436008 | 0 | 49.97 | 2.886 | 211.0647 | 8.163 | 12 | 2010 | 0 | 2 | 113 |
| 5044 | 4 | 3526713 | 0 | 43.21 | 2.887 | 126.9836 | 7.127 | 12 | 2010 | 0 | 2 | 113 |
| 5046 | 6 | 2727575 | 0 | 55.07 | 2.886 | 212.9165 | 7.007 | 12 | 2010 | 0 | 2 | 113 |
| 5050 | 10 | 3749058 | 0 | 57.06 | 3.236 | 126.9836 | 9.003 | 12 | 2010 | 0 | 2 | 113 |
| 5053 | 13 | 3595903 | 0 | 34.90 | 2.846 | 126.9836 | 7.795 | 12 | 2010 | 0 | 2 | 113 |
| 5054 | 14 | 3818686 | 0 | 30.59 | 3.141 | 182.5446 | 8.724 | 12 | 2010 | 0 | 2 | 113 |
| 5060 | 20 | 3766687 | 0 | 25.17 | 3.141 | 204.6377 | 7.484 | 12 | 2010 | 0 | 2 | 113 |
| 5063 | 23 | 2734277 | 0 | 22.96 | 3.150 | 132.7477 | 5.287 | 12 | 2010 | 0 | 2 | 113 |
| 5067 | 27 | 3078162 | 0 | 31.34 | 3.309 | 136.5973 | 8.021 | 12 | 2010 | 0 | 2 | 113 |
| 5266 | 10 | 2950199 | 1 | 60.68 | 3.760 | 129.8364 | 7.874 | 11 | 2011 | Thanksgiving | 2 | 118 |
| 5273 | 13 | 2864171 | 1 | 38.89 | 3.445 | 129.8364 | 6.392 | 11 | 2011 | Thanksgiving | 2 | 118 |
| 5282 | 4 | 3004702 | 1 | 47.96 | 3.225 | 129.8364 | 5.143 | 11 | 2011 | Thanksgiving | 2 | 118 |
| 5307 | 20 | 2906233 | 1 | 46.38 | 3.492 | 211.4121 | 7.082 | 11 | 2011 | Thanksgiving | 2 | 118 |
| 5500 | 10 | 2939946 | 1 | 55.33 | 3.162 | 126.6693 | 9.003 | 11 | 2010 | Thanksgiving | 2 | 123 |
| 5508 | 13 | 2766400 | 1 | 28.22 | 2.830 | 126.6693 | 7.795 | 11 | 2010 | Thanksgiving | 2 | 123 |
| 5517 | 14 | 2921710 | 1 | 46.15 | 3.039 | 182.7833 | 8.724 | 11 | 2010 | Thanksgiving | 2 | 123 |
| 5527 | 20 | 2811634 | 1 | 46.66 | 3.039 | 204.9621 | 7.484 | 11 | 2010 | Thanksgiving | 2 | 123 |
| 5532 | 4 | 2789469 | 1 | 48.08 | 2.752 | 126.6693 | 7.127 | 11 | 2010 | Thanksgiving | 2 | 123 |

Fig14 Outlier subset of the dataset of sales higher than upper and lower limit

We now check the correlation of all the factors with the target variable (Weekly_sales) and see if how much impact they contribute. We use a heat map to display our findings more easier to understand. Note: we haven't use the holiday name column as its non numeric.

```
#check for the corelation between the variables

corr = cor(dataset[, -10])

View(corr)

corrplot(corr = corr, method = "color", outline = T, cl.pos = 'n', rect.col = "black",  tl.col = "indianred4", addCoef.col = "black", number.digits = 2, number.cex = 0.60, tl.cex = 0.7, cl.cex = 1, col = colorRampPalette(c("green4","white","red"))(100))

names(dataset)
```

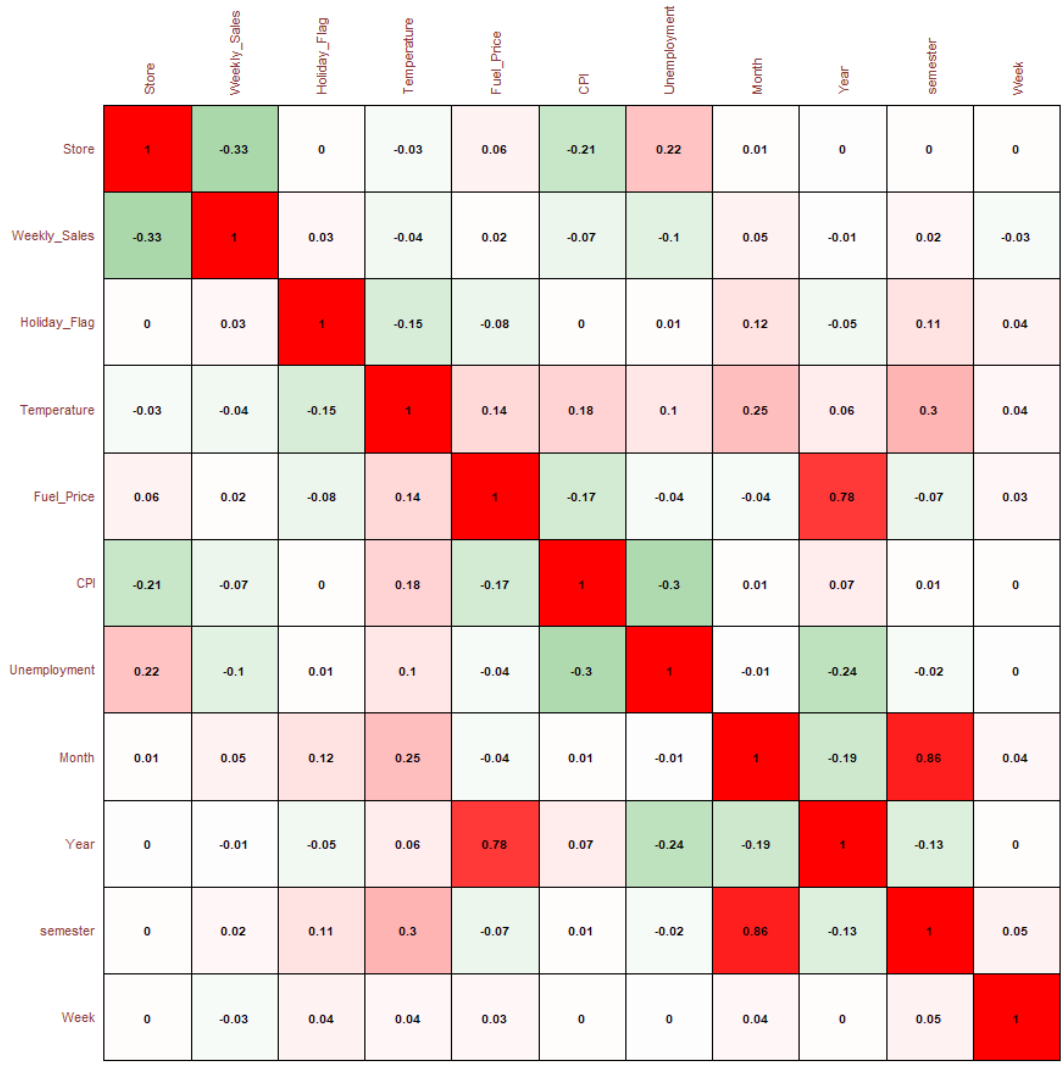



Fig 15. Heat Map of the correlation of variables with each other

From the plot, we can see that correlation of different variables with the weekly sales is very low including the CPI, unemployment and fuel prices. Hence we accept the NULL hypothesis that the CPI, unemployment and fuel prices have very low to no impact on the sales value.

We create dummy variables for the categorical column Name_holiday as it wont be possible to add this to the model. We then delete the original categorical variable and merge the dummy variable dataframe to the original dataset.

```
#Create dummy variables for the holiday categorical column
holiday_fact <- as.factor(dataset$Name_Holiday)
dummy_holiday <- data.frame(model.matrix(~holiday_fact))[,-1]
#Merging the dummy variables to the final dataset
dataset <- cbind(dataset,dummy_holiday)
dataset <- subset(dataset,select = -Name_Holiday) #Dropping the categorical column
```

| | holiday_factChristmas | holiday_factLabour.Day | holiday_factSuper.Bowl | holiday_factThanksgiving |
|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 |
| 12 | 0 | 0 | 0 | 0 |
| 13 | 0 | 0 | 0 | 0 |
| 14 | 0 | 0 | 0 | 0 |
| 15 | 0 | 0 | 0 | 0 |
| 16 | 0 | 0 | 0 | 0 |
| 17 | 0 | 0 | 0 | 0 |
| 18 | 0 | 0 | 0 | 0 |
| 19 | 0 | 0 | 0 | 0 |
| 20 | 0 | 0 | 0 | 0 |
| 21 | 0 | 0 | 0 | 0 |
| 22 | 0 | 0 | 0 | 0 |
| 23 | 0 | 0 | 0 | 0 |
| 24 | 0 | 0 | 0 | 0 |
| 25 | 0 | 0 | 0 | 0 |
| 26 | 0 | 0 | 0 | 0 |
| 27 | 0 | 0 | 0 | 0 |
| 28 | 0 | 0 | 0 | 0 |
| 29 | 0 | 0 | 0 | 0 |
| 30 | 0 | 0 | 0 | 0 |
| 31 | 0 | 0 | 0 | 0 |
| 32 | 0 | 0 | 0 | 0 |
| 33 | 0 | 0 | 0 | 0 |

Fig 16. Dummy variable of holiday column

Now that we have taken care of the key aspects of the data, the next step is to proceed in building the model. We set a seed 123 and split the data into train and test sets. We continue to build 7 models each having different factors contributing to make it an efficient model.

*#Splitting the data into train and test sets*

set.seed(123)

sample <- sample.split(dataset, SplitRatio = 0.7)

trainSet <- subset(dataset, sample ==T)

testSet <- subset(dataset, sample == F)

*#Model-1*

*#Create the model*

model1 = lm(formula = Weekly_Sales ~.,data = trainSet)

summary(model1)

```
Call:
lm(formula = Weekly_Sales ~ ., data = trainSet)

Residuals:
     Min       1Q   Median       3Q      Max 
-1130278  -387165   -28889   374521  1779216 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)              -4.916e+06  3.594e+07  -0.137 0.891205    
Store                    -1.561e+04  5.916e+02 -26.388  < 2e-16 ***
Holiday_Flag             -2.430e+04  6.528e+04  -0.372 0.709731    
Temperature               1.365e+03  4.761e+02   2.867 0.004163 ** 
Fuel_Price               -1.203e+04  3.092e+04  -0.389 0.697182    
CPI                      -2.840e+03  2.235e+02 -12.706  < 2e-16 ***
Unemployment             -2.016e+04  4.535e+03  -4.446 8.99e-06 ***
Month                     1.954e+04  5.096e+03   3.834 0.000128 ***
Year                      3.442e+03  1.791e+04   0.192 0.847625    
semester                 -9.963e+04  3.161e+04  -3.152 0.001632 ** 
Week                      7.837e+01  1.912e+02   0.410 0.681873    
holiday_factChristmas    -4.967e+04  9.570e+04  -0.519 0.603758    
holiday_factLabour.Day   -6.182e+04  1.113e+05  -0.556 0.578534    
holiday_factSuper.Bowl    1.575e+05  8.545e+04   1.843 0.065343 .  
holiday_factThanksgiving  3.745e+05  9.554e+04   3.919 9.02e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 490700 on 4251 degrees of freedom
Multiple R-squared:  0.1728,	Adjusted R-squared:  0.1701 
F-statistic: 63.44 on 14 and 4251 DF,  p-value: < 2.2e-16
```

Fig 17-24 : Model Summaries

*#Model-2*

names(dataset)

*#Create the model*

model2 = lm(formula = Weekly_Sales ~ Store+CPI+Unemployment+Week+Temperature+Fuel_Price+holiday_factChristmas+holiday_factLabour.Day+holiday_factSuper.Bowl+holiday_factThanksgiving,data = trainSet) *#removing the semester, month and year factors*

summary(model2)

```
Call:
lm(formula = Weekly_Sales ~ Store + CPI + Unemployment + Week +
    Temperature + Fuel_Price + holiday_factChristmas + holiday_factLabour.Day +
    holiday_factSuper.Bowl + holiday_factThanksgiving, data = trainSet)

Residuals:
     Min       1Q   Median       3Q      Max
-1129819  -385796   -29621   370744  1840501

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               1954532.16   88876.31  21.992  < 2e-16 ***
Store                      -15598.99     591.99 -26.350  < 2e-16 ***
CPI                         -2815.41     212.35 -13.258  < 2e-16 ***
Unemployment               -20184.93    4353.43  -4.637 3.65e-06 ***
Week                           20.04     189.49   0.106  0.91577
Temperature                  1394.14     446.51   3.122  0.00181 **
Fuel_Price                  -4391.97   17240.09  -0.255  0.79893
holiday_factChristmas       -9520.42   66319.01  -0.144  0.88586
holiday_factLabour.Day     -85167.10   91125.43  -0.935  0.35004
holiday_factSuper.Bowl      98466.57   53850.46   1.829  0.06754 .
holiday_factThanksgiving   393301.44   68312.41   5.757 9.14e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 491400 on 4255 degrees of freedom
Multiple R-squared:  0.1696,	Adjusted R-squared:  0.1677
F-statistic: 86.93 on 10 and 4255 DF,  p-value: < 2.2e-16
```

*#Model-3*

names(dataset)

#Create the model

model3 = lm(formula = Weekly_Sales ~ Store+CPI+Unemployment+Week+Temperature+holiday_factChristmas+holiday_factLabour.Day+holiday_factSuper.Bowl+holiday_factThanksgiving,data = trainSet) *#removing fuel price factor* summary(model3)

```
Call:
lm(formula = Weekly_Sales ~ Store + CPI + Unemployment + Week +
    Temperature + holiday_factChristmas + holiday_factLabour.Day +
    holiday_factSuper.Bowl + holiday_factThanksgiving, data = trainSet)

Residuals:
     Min       1Q   Median       3Q      Max
-1127850  -385568   -29623   369654  1842137

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1937887.7    60244.9  32.167  < 2e-16 ***
Store                     -15603.7      591.6 -26.373  < 2e-16 ***
CPI                        -2803.3      206.9 -13.547  < 2e-16 ***
Unemployment              -20043.1     4317.2  -4.643 3.54e-06 ***
Week                          20.0      189.5   0.106  0.91593
Temperature                 1373.2      438.9   3.129  0.00177 **
holiday_factChristmas      -9081.9    66289.4  -0.137  0.89103
holiday_factLabour.Day    -82082.4    90307.4  -0.909  0.36344
holiday_factSuper.Bowl     99009.3    53802.4   1.840  0.06580 .
holiday_factThanksgiving  393841.7    68272.0   5.769 8.55e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 491400 on 4256 degrees of freedom
Multiple R-squared:  0.1696,	Adjusted R-squared:  0.1679
F-statistic:  96.6 on 9 and 4256 DF,  p-value: < 2.2e-16
```

*#Model-4*

names(dataset)

*#Create the model*

model4 = lm(formula = Weekly_Sales ~Store+CPI+Unemployment+Temperature+Fuel_Price+holiday_factChristmas+holiday_factLabour.Day+holiday_factSuper.Bowl+holiday_factThanksgiving,data = trainSet) *#removing the week factor*

summary(model4)

```
Call:
lm(formula = Weekly_Sales ~ Store + CPI + Unemployment + Temperature +
    Fuel_Price + holiday_factChristmas + holiday_factLabour.Day +
    holiday_factSuper.Bowl + holiday_factThanksgiving, data = trainSet)

Residuals:
     Min       1Q   Median       3Q      Max
-1129850  -385442   -28911   370256  1841441

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1956223.1    87416.5  22.378  < 2e-16 ***
Store                     -15601.4      591.5 -26.376  < 2e-16 ***
CPI                        -2816.4      212.1 -13.276  < 2e-16 ***
Unemployment              -20212.3     4345.2  -4.652 3.39e-06 ***
Temperature                 1396.8      445.8   3.134  0.00174 **
Fuel_Price                 -4390.4    17238.1  -0.255  0.79897
holiday_factChristmas      -8029.3    64795.9  -0.124  0.90139
holiday_factLabour.Day    -85690.1    90980.6  -0.942  0.34632
holiday_factSuper.Bowl     98083.9    53722.6   1.826  0.06796 .
holiday_factThanksgiving  394328.6    67610.8   5.832 5.87e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 491400 on 4256 degrees of freedom
Multiple R-squared:  0.1696,	Adjusted R-squared:  0.1679
F-statistic: 96.61 on 9 and 4256 DF,  p-value: < 2.2e-16

>
```

*#Model-5*

names(dataset)

*#Create the model*

model5 = lm(formula = Weekly_Sales ~Store+CPI+Unemployment+Week+Temperature+Fuel_Price+holiday_factLabour.Day+holiday_factSuper.Bowl+holiday_factThanksgiving,data = trainSet) *#removing the christmas holiday factor*

summary(model5)

```
Call:
lm(formula = Weekly_Sales ~ Store + CPI + Unemployment + Week + 
    Temperature + Fuel_Price + holiday_factLabour.Day + holiday_factSuper.Bowl + 
    holiday_factThanksgiving, data = trainSet)

Residuals:
     Min       1Q   Median       3Q      Max 
-1129956  -385844   -29439   370182  1840976 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)              1954381.38   88859.87  21.994  < 2e-16 ***
Store                     -15598.40     591.91 -26.353  < 2e-16 ***
CPI                        -2816.34     212.23 -13.270  < 2e-16 ***
Unemployment              -20219.84    4346.13  -4.652 3.38e-06 ***
Week                          14.26     185.14   0.077  0.93861    
Temperature                 1404.52     440.56   3.188  0.00144 ** 
Fuel_Price                 -4327.74   17232.30  -0.251  0.80172    
holiday_factLabour.Day    -85236.82   91113.65  -0.936  0.34958    
holiday_factSuper.Bowl     98762.94   53804.68   1.836  0.06649 .  
holiday_factThanksgiving  393867.87   68190.51   5.776 8.20e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 491400 on 4256 degrees of freedom
Multiple R-squared:  0.1696,	Adjusted R-squared:  0.1679 
F-statistic:  96.6 on 9 and 4256 DF,  p-value: < 2.2e-16
```

*#Model-6*

names(dataset)

*#Create the model*

model6 = lm(formula = Weekly_Sales ~ Store+CPI+Unemployment+Week+Temperature+Fuel_Price+holiday_factChristmas+holiday_factLabour.Day+holiday_factSuper.Bowl+holiday_factThanksgiving +semester,data = trainSet) *#adding semester factor*

summary(model6)

```
Call:
lm(formula = Weekly_Sales ~ Store + CPI + Unemployment + Week +
    Temperature + Fuel_Price + holiday_factChristmas + holiday_factLabour.Day +
    holiday_factSuper.Bowl + holiday_factThanksgiving + semester,
    data = trainSet)

Residuals:
     Min       1Q   Median       3Q      Max
-1128971  -384995   -31155   371328  1838376

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              1947361.72   93699.12  20.783  < 2e-16 ***
Store                     -15600.95     592.11 -26.348  < 2e-16 ***
CPI                        -2810.33     213.41 -13.169  < 2e-16 ***
Unemployment              -20082.32    4374.52  -4.591 4.54e-06 ***
Week                          20.93     189.55   0.110  0.91206
Temperature                 1355.18     474.71   2.855  0.00433 **
Fuel_Price                 -3822.23   17402.14  -0.220  0.82616
holiday_factChristmas     -12449.17   67422.48  -0.185  0.85352
holiday_factLabour.Day    -86438.94   91287.07  -0.947  0.34375
holiday_factSuper.Bowl     99524.05   54033.57   1.842  0.06556 .
holiday_factThanksgiving  390787.98   69105.68   5.655 1.66e-08 ***
semester                    3991.38   16501.27   0.242  0.80888
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 491500 on 4254 degrees of freedom
Multiple R-squared:  0.1696,	Adjusted R-squared:  0.1675
F-statistic: 79.01 on 11 and 4254 DF,  p-value: < 2.2e-16
```

*#Model-7*

names(dataset)

*#Create the model*

model7 = lm(formula = Weekly_Sales ~ Store+CPI+Unemployment+Week+Temperature+holiday_factChristmas+holiday_factLabour.Day+holiday_factSuper.Bowl+holiday_factThanksgiving +semester+Year,data = trainSet) *#adding year factor and removing fuel price*

summary(model7)

```
Call:
lm(formula = Weekly_Sales ~ Store + CPI + Unemployment + Week +
    Temperature + holiday_factChristmas + holiday_factLabour.Day +
    holiday_factSuper.Bowl + holiday_factThanksgiving + semester +
    Year, data = trainSet)

Residuals:
     Min       1Q   Median       3Q      Max
-1133423  -384703   -29598   371798  1832378

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.942e+07  2.013e+07   0.965  0.33468
Store                   -1.558e+04  5.924e+02 -26.302  < 2e-16 ***
CPI                     -2.802e+03  2.075e+02 -13.507  < 2e-16 ***
Unemployment            -2.098e+04  4.490e+03  -4.672 3.08e-06 ***
Week                     2.139e+01  1.895e+02   0.113  0.91017
Temperature              1.386e+03  4.674e+02   2.965  0.00304 **
holiday_factChristmas   -1.399e+04  6.744e+04  -0.207  0.83569
holiday_factLabour.Day  -9.159e+04  9.100e+04  -1.006  0.31426
holiday_factSuper.Bowl   1.008e+05  5.397e+04   1.869  0.06176 .
holiday_factThanksgiving 3.887e+05  6.914e+04   5.622 2.01e-08 ***
semester                 2.303e+03  1.654e+04   0.139  0.88925
Year                    -8.691e+03  1.000e+04  -0.869  0.38499
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 491400 on 4254 degrees of freedom
Multiple R-squared:  0.1698,	Adjusted R-squared:  0.1676
F-statistic: 79.09 on 11 and 4254 DF,  p-value: < 2.2e-16
```

We have successfully created 7 models of varying R sqr and adjusted R sqr. We choose the 7th model as it has a relatively higher R sqr and it has a low difference between the R sqr and adjusted R sqr is relatively low. We move forward to test our model with the test set and get the predicted values.

*#Test and find the predictions with the test set*

testSet$pred_price <- predict(model7,newdata = testSet) #we select model 7 due to best value of rsqr and adjusted rsqr

View(subset(testSet, select = -c(holiday_factChristmas,holiday_factLabour.Day,holiday_factSuper.Bowl,holiday_factThanksgiving))) *#for easier view*

*#Visualization of actual vs predicted price*

```
ggplot()+
  geom_point(aes(x = testSet$Weekly_Sales,y = testSet$pred_price))+
  xlab("Actual price")+
  ylab("Predicted price")+
  ggtitle("Graphical Analysis of actual vs predicted prices")
```

| | Store | Weekly_Sales | Holiday_Flag | Temperature | Fuel_Price | CPI | Unemployment | Month | Year | semester | Week | pred_price |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 992 | 1 | 1670786.0 | 0 | 68.55 | 3.617 | 223.1815 | 6.573 | 10 | 2012 | 2 | 23 | 1252806 |
| 2048 | 1 | 1507460.7 | 1 | 78.69 | 2.565 | 211.4952 | 7.787 | 9 | 2010 | 2 | 46 | 1200429 |
| 3782 | 1 | 1539483.7 | 0 | 62.25 | 3.308 | 218.2205 | 7.866 | 11 | 2011 | 2 | 85 | 1240870 |
| 5279 | 1 | 2033320.7 | 1 | 60.14 | 3.236 | 218.4676 | 7.866 | 11 | 2011 | 2 | 118 | 1626675 |
| 5311 | 1 | 1409727.6 | 0 | 46.63 | 2.561 | 211.3196 | 8.106 | 2 | 2010 | 1 | 119 | 1240639 |
| 5356 | 1 | 1404429.9 | 0 | 51.45 | 2.732 | 211.0180 | 8.106 | 3 | 2010 | 1 | 120 | 1248186 |
| 5401 | 1 | 1464693.5 | 0 | 87.96 | 3.523 | 215.7332 | 7.962 | 8 | 2011 | 2 | 121 | 1282233 |
| 5446 | 1 | 1493659.7 | 0 | 69.16 | 3.506 | 223.4443 | 6.573 | 10 | 2012 | 2 | 122 | 1255033 |
| 2 | 2 | 1800171.4 | 0 | 55.43 | 3.524 | 214.4887 | 7.931 | 4 | 2011 | 1 | 1 | 1220831 |
| 47 | 2 | 1910092.4 | 0 | 78.38 | 3.501 | 221.3853 | 6.891 | 6 | 2012 | 1 | 2 | 1246461 |
| 92 | 2 | 1866243.0 | 0 | 85.69 | 3.524 | 214.8369 | 7.852 | 7 | 2011 | 2 | 3 | 1265801 |
| 137 | 2 | 1827440.4 | 0 | 69.24 | 2.603 | 211.3299 | 8.163 | 10 | 2010 | 2 | 4 | 1255017 |
| 182 | 2 | 1952555.7 | 0 | 58.79 | 3.630 | 220.4867 | 7.057 | 3 | 2012 | 1 | 5 | 1218408 |
| 227 | 2 | 2066187.7 | 0 | 63.27 | 2.719 | 210.4799 | 8.200 | 4 | 2010 | 1 | 6 | 1246087 |
| 272 | 2 | 2003940.6 | 0 | 82.74 | 2.669 | 210.8804 | 8.099 | 7 | 2010 | 2 | 7 | 1276395 |
| 317 | 2 | 1809119.7 | 0 | 89.64 | 3.533 | 215.4509 | 7.852 | 9 | 2011 | 2 | 8 | 1269663 |
| 362 | 2 | 1954952.0 | 0 | 48.74 | 3.172 | 218.3590 | 7.441 | 12 | 2011 | 2 | 9 | 1213465 |
| 407 | 2 | 1935299.9 | 0 | 55.21 | 3.360 | 219.8119 | 7.057 | 2 | 2012 | 1 | 10 | 1215443 |
| 452 | 2 | 1933756.2 | 0 | 83.07 | 3.699 | 214.9255 | 7.931 | 6 | 2011 | 1 | 11 | 1258132 |
| 497 | 2 | 1946104.6 | 0 | 90.22 | 3.417 | 221.5870 | 6.565 | 8 | 2012 | 2 | 12 | 1271663 |
| 542 | 2 | 1904608.1 | 0 | 81.83 | 2.577 | 211.1887 | 8.099 | 9 | 2010 | 2 | 13 | 1274398 |
| 632 | 2 | 1929346.2 | 0 | 38.25 | 2.989 | 212.2241 | 8.028 | 2 | 2011 | 1 | 15 | 1201628 |
| 677 | 2 | 1981607.8 | 0 | 57.77 | 3.288 | 213.4775 | 8.028 | 3 | 2011 | 1 | 16 | 1225194 |
| 722 | 2 | 1923957.1 | 0 | 76.73 | 3.749 | 221.3095 | 6.891 | 5 | 2012 | 1 | 17 | 1244707 |
| 767 | 2 | 2102539.9 | 0 | 81.81 | 2.705 | 210.8336 | 8.200 | 6 | 2010 | 1 | 18 | 1271050 |
| 812 | 2 | 1959707.9 | 0 | 55.53 | 3.332 | 217.4854 | 7.441 | 11 | 2011 | 2 | 19 | 1225538 |
| 857 | 2 | 2136989.5 | 0 | 40.19 | 2.572 | 210.7526 | 8.324 | 2 | 2010 | 1 | 20 | 1211030 |
| 902 | 2 | 1991013.1 | 0 | 47.17 | 2.625 | 211.0068 | 8.324 | 3 | 2010 | 1 | 21 | 1220014 |
| 947 | 2 | 1876704.3 | 0 | 93.34 | 3.684 | 215.1979 | 7.852 | 8 | 2011 | 2 | 22 | 1275800 |
| 994 | 2 | 1998321.0 | 0 | 70.27 | 3.617 | 222.8159 | 6.170 | 10 | 2012 | 2 | 23 | 1249087 |
| 1037 | 2 | 1939061.4 | 0 | 57.85 | 2.689 | 211.6135 | 8.163 | 11 | 2010 | 2 | 24 | 1238862 |
| 1082 | 2 | 1799520.1 | 0 | 46.75 | 3.157 | 219.3551 | 7.057 | 1 | 2012 | 1 | 25 | 1205318 |
| 1127 | 2 | 2129035.9 | 0 | 68.43 | 3.891 | 221.0738 | 6.891 | 4 | 2012 | 1 | 26 | 1234056 |
| 1172 | 2 | 1837743.6 | 0 | 61.48 | 3.906 | 215.4449 | 7.931 | 5 | 2011 | 1 | 27 | 1227093 |

Fig 25. Predicted values of the Weekly prices of the test set

From the snippet of the predicted sales, we can see the forecasted sales for store 1. Visualization of the results is given below.

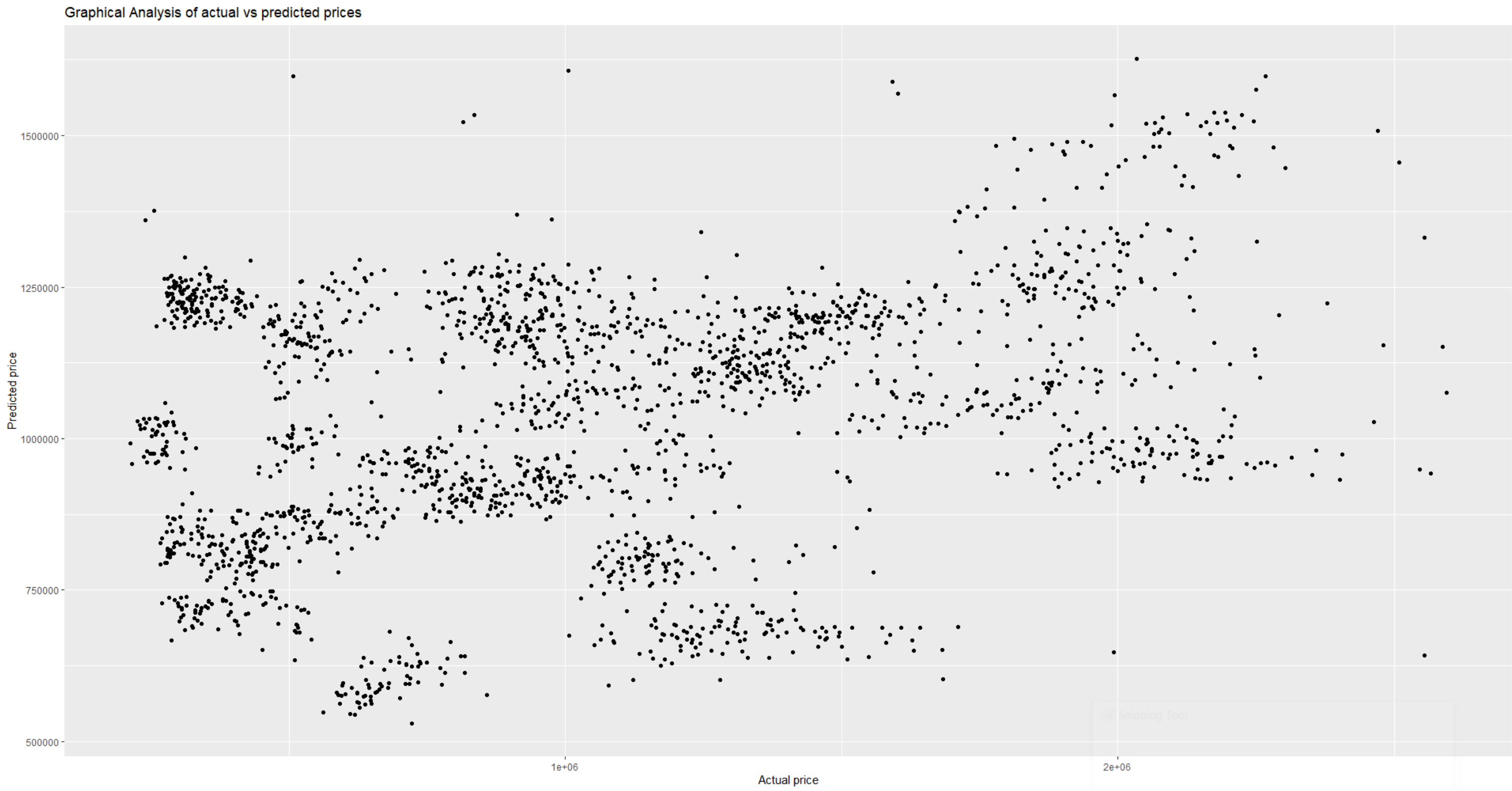


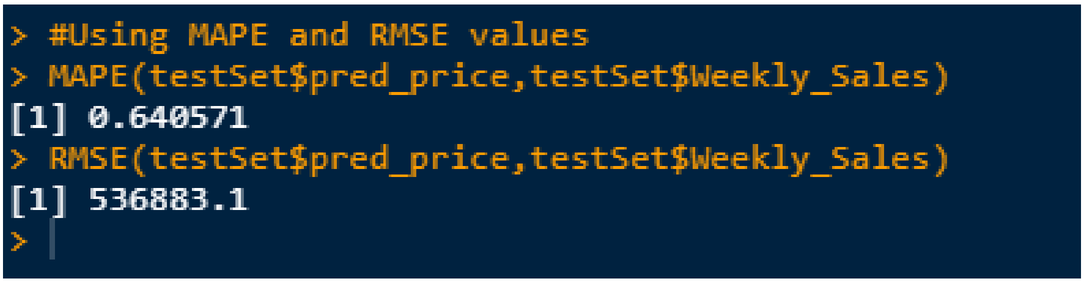

Fig 26. Visualization of actual sales vs predicted sales

Evaluation of the model can be performed by MAPE and RMSE calculations.

*#Using MAPE and RMSE values*

MAPE(testSet$pred_price,testSet$Weekly_Sales)

RMSE(testSet$pred_price,testSet$Weekly_Sales)



```
> #Using MAPE and RMSE values
> MAPE(testSet$pred_price,testSet$Weekly_Sales)
[1] 0.640571
> RMSE(testSet$pred_price,testSet$Weekly_Sales)
[1] 536883.1
>
```

Fig 27. RMSE and MAPE values

# Conclusions and Inferences

We concluded from our deeper statistical analysis that **Store No. 20** had the maximum total sales while **store No. 14** had the maximum variation in sales from the years 2010-2012. We also found out from the calculated coefficient of variations that **Store No. 35** had the maximum variation of sales about the average sales. It is seen that **Stores 7,16,23,24,26,35,39,40,41** and **44** have had a positive growth rate in Q3-2012 from Q2-2012. It is observed that the mean sales on **Super Bowl** and **Thanksgiving days** are higher than the mean sales on non-holiday season days. Finally, visualizations on monthly and semesterly sales are plotted and it is found out that the sales in both the semesters are almost relatively equal with the first semester lower than the second semester due the low sales in the month of January. We proceeded to build a statistical model to predict and forecast the weekly prices and found out that CPI, Unemployment and Fuel Prices have very weak correlation with the weekly sales. Due to this issue, the R sqr value was affected and the model was very accurate in the predictions. Hence to improve the model, more new variables can be included to the dataset to improve the model quality.

# Next Steps

In the next step, to obtain better prediction of values and accuracy, we should test it with time series models and random forest algorithms.

*********************************