**ANALYSIS & PREDICTION OF WALMART SALES**[¶](https://www.kaggle.com/code/sarvaninandipati/analysis-prediction-of-walmart-sales-using-r/notebook#ANALYSIS-&-PREDICTION-OF-WALMART-SALES)

**PROBLEM STATEMENT**

**DESCRIPTION**

One of the leading retail stores in the US, Walmart, 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 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

**ANALYSIS TASKS**

1. Which store has maximum sales
2. Which store has maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of mean to standard deviation
3. Which store/s has good quarterly growth rate in Q3’2012
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
5. Provide a monthly and semester view of sales in units and give insights

**STATISTICAL MODEL**

For Store 1 – Build prediction models to forecast demand

Linear Regression – 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.

Change dates into days by creating new variable.

Select the model which gives best accuracy

**PROJECT**

*#getting work directory*

getwd()

'/kaggle/working'

*#uploading data*

data1 <- read.csv("../input/retail-analysis-with-walmart-sales-data/WALMART\_SALES\_DATA.csv")

*#Calling libraries*

library("dplyr") *#Calling dplyr function for data manipulation*

library("ggplot2") *# for data visualisation*

library("scales") *#for change of scales in data visualisation*

library("zoo")

library("tidyverse")

library("tidyr")

library("lubridate")

library(car) *#Companion to Applied Regression for Regression Visualisations*

require(stats)

library(corrplot)

library(caTools)

library(MLmetrics)

library("repr")

Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

filter, lag

The following objects are masked from ‘package:base’:

intersect, setdiff, setequal, union

Attaching package: ‘zoo’

The following objects are masked from ‘package:base’:

as.Date, as.Date.numeric

── **Attaching packages** ─────────────────────────────────────── tidyverse 1.3.1 ──

✔ tibble 3.1.2 ✔ purrr 0.3.4

✔ tidyr 1.1.3 ✔ stringr 1.4.0

✔ readr 1.4.0 ✔ forcats 0.5.1

── **Conflicts** ────────────────────────────────────────── tidyverse\_conflicts() ──

✖ readr::col\_factor() masks scales::col\_factor()

✖ purrr::discard() masks scales::discard()

✖ dplyr::filter() masks stats::filter()

✖ dplyr::lag() masks stats::lag()

Attaching package: ‘lubridate’

The following objects are masked from ‘package:base’:

date, intersect, setdiff, union

Loading required package: carData

Attaching package: ‘car’

The following object is masked from ‘package:purrr’:

some

The following object is masked from ‘package:dplyr’:

recode

corrplot 0.88 loaded

Attaching package: ‘MLmetrics’

The following object is masked from ‘package:base’:

Recall

Understanding Data

head(data1)

| A data.frame: 6 × 8 | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Store | Date | Weekly\_Sales | Holiday\_Flag | Temperature | Fuel\_Price | CPI | Unemployment |
|  | <int> | <chr> | <dbl> | <int> | <dbl> | <dbl> | <dbl> | <dbl> |
| 1 | 1 | 05-02-2010 | 1643691 | 0 | 42.31 | 2.572 | 211.0964 | 8.106 |
| 2 | 1 | 12-02-2010 | 1641957 | 1 | 38.51 | 2.548 | 211.2422 | 8.106 |
| 3 | 1 | 19-02-2010 | 1611968 | 0 | 39.93 | 2.514 | 211.2891 | 8.106 |
| 4 | 1 | 26-02-2010 | 1409728 | 0 | 46.63 | 2.561 | 211.3196 | 8.106 |
| 5 | 1 | 05-03-2010 | 1554807 | 0 | 46.50 | 2.625 | 211.3501 | 8.106 |
| 6 | 1 | 12-03-2010 | 1439542 | 0 | 57.79 | 2.667 | 211.3806 | 8.106 |

*#Data Exploration - dimensions*

dim(data1)

1. 6435
2. 8

*#Data Exploration - class*

class(data1)

'data.frame'

*#Data Exploration -structure*

str(data1)

'data.frame': 6435 obs. of 8 variables:

$ Store : int 1 1 1 1 1 1 1 1 1 1 ...

$ Date : chr "05-02-2010" "12-02-2010" "19-02-2010" "26-02-2010" ...

$ Weekly\_Sales: num 1643691 1641957 1611968 1409728 1554807 ...

$ Holiday\_Flag: int 0 1 0 0 0 0 0 0 0 0 ...

$ Temperature : num 42.3 38.5 39.9 46.6 46.5 ...

$ Fuel\_Price : num 2.57 2.55 2.51 2.56 2.62 ...

$ CPI : num 211 211 211 211 211 ...

$ Unemployment: num 8.11 8.11 8.11 8.11 8.11 ...

*#Data Exploration - summary*

summary(data1)

Store Date Weekly\_Sales Holiday\_Flag

Min. : 1 Length:6435 Min. : 209986 Min. :0.00000

1st Qu.:12 Class :character 1st Qu.: 553350 1st Qu.:0.00000

Median :23 Mode :character Median : 960746 Median :0.00000

Mean :23 Mean :1046965 Mean :0.06993

3rd Qu.:34 3rd Qu.:1420159 3rd Qu.:0.00000

Max. :45 Max. :3818686 Max. :1.00000

Temperature Fuel\_Price CPI Unemployment

Min. : -2.06 Min. :2.472 Min. :126.1 Min. : 3.879

1st Qu.: 47.46 1st Qu.:2.933 1st Qu.:131.7 1st Qu.: 6.891

Median : 62.67 Median :3.445 Median :182.6 Median : 7.874

Mean : 60.66 Mean :3.359 Mean :171.6 Mean : 7.999

3rd Qu.: 74.94 3rd Qu.:3.735 3rd Qu.:212.7 3rd Qu.: 8.622

Max. :100.14 Max. :4.468 Max. :227.2 Max. :14.313

*#Data Exploration - tables*

table(data1$Store)

table(data1$Holiday\_Flag)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

143 143 143 143 143 143 143 143 143 143 143 143 143 143 143 143 143 143 143 143

21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40

143 143 143 143 143 143 143 143 143 143 143 143 143 143 143 143 143 143 143 143

41 42 43 44 45

143 143 143 143 143

0 1

5985 450

*#Checking NA values*

colSums(is.na(data1)) *#Observed no NA values*

**Store**

0

**Date**

0

**Weekly\_Sales**

0

**Holiday\_Flag**

0

**Temperature**

0

**Fuel\_Price**

0

**CPI**

0

**Unemployment**

0

*#Checking Duplicate Values*

all(duplicated(data1) == TRUE)

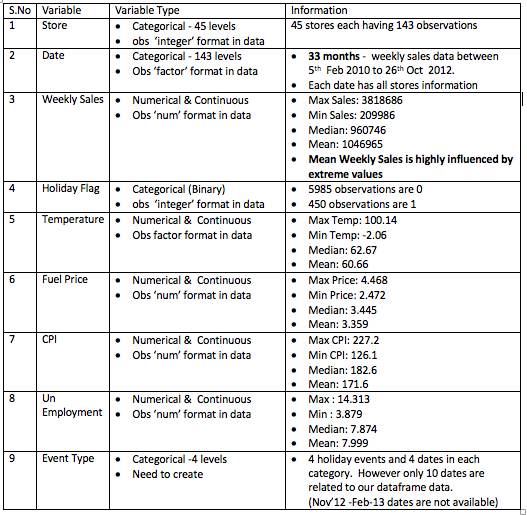
*#observed no duplicate values*

FALSE

**Summarisation of Data**

Exploratory Data Analysis of Retail Analysis with Walmart Data

* Industry: Retail Industry
* This data is about different stores weekly sales and consists possible influence factors CPI, Un Employment Rate Fuel Price, Holiday Type that may affect sales
* Dimensions: 6435 observations, 8 variables.
* Missing Data / Null Values: No
* Duplicated Values: No
* Structure: Data frame



**ANALYSIS TASKS**

Which Store has maximum sales?

*#Aggregating data by 'Store' and Finding sum of 'Weekly\_Sales'*

Store\_Sales<- aggregate(Weekly\_Sales ~ Store, data = data1, sum)

*#Changing column name of sales*

colnames(Store\_Sales)[2] <- "Total\_Sales\_by\_Store"

*#Finding out Store with highest Sales*

Store\_Sales <-arrange(Store\_Sales, desc(Total\_Sales\_by\_Store)) *#Arranged Stores based on Sales in descending order*

Store\_Sales[1,] *#Choosing the first store that comes in this order*

|  |  |  |
| --- | --- | --- |
| A data.frame: 1 × 2 | | |
|  | Store | Total\_Sales\_by\_Store |
|  | <int> | <dbl> |
| 1 | 20 | 301397792 |

*#Printing the output*

print(paste('Store no.', Store\_Sales[1,]$Store,

'has the maximum sales and the value is = ', Store\_Sales[1,]$Total\_Sales\_by\_Store))

[1] "Store no. 20 has the maximum sales and the value is = 301397792.46"

*# Converting Store column into factor so that order won't change for graph*

Store\_Sales$Store <- as.character(Store\_Sales$Store)

Store\_Sales$Store <- factor(Store\_Sales$Store, levels=unique(Store\_Sales$Store))

*#Plotting Store vs TotalSales*

options(repr.plot.width = 14, repr.plot.height = 8)

a<-ggplot(data=Store\_Sales, aes(x=Store, y=Total\_Sales\_by\_Store)) + geom\_bar(stat="identity",fill="steelblue") +

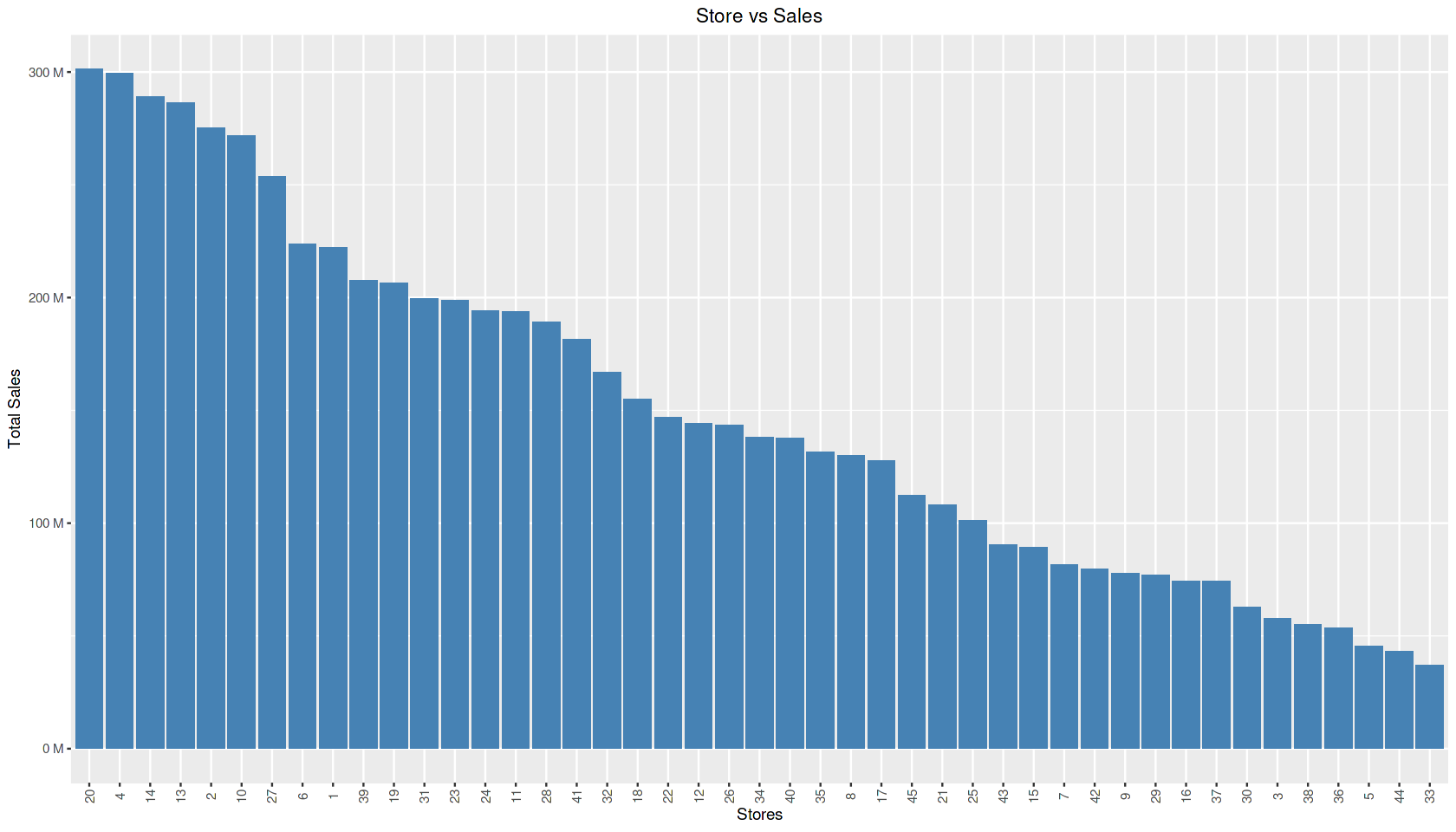
theme(axis.text.x = element\_text(angle = 90,vjust = 0.5, hjust=0.5))+ scale\_x\_discrete(breaks = data1$Store)+

scale\_y\_continuous(labels = label\_number(suffix = " M", scale = 1e-6))+ ggtitle('Store vs Sales')+

theme(plot.title = element\_text(hjust = 0.5))+

xlab("Stores") + ylab("Total Sales")

a



**Insights:**

* Store 20 has the maximum sales and the value is 301397792.46 (i.e., 301.39M)
* Store 4 is the second largest store in terms of sales and the value is 299543953 (i.e., 299.54M)
* Store 33 has the least sales 37160222 (i.e., 37.16M)

Which store has maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of mean to standard deviation

In [15]:

*#Aggregating data by 'Store' and Finding Standard Deviation of 'Weekly\_Sales'*

Store\_Sales\_Variation<-summarise(group\_by(data1,Store),sd(Weekly\_Sales), mean(Weekly\_Sales))

*#Changing column names*

colnames(Store\_Sales\_Variation)[2] <- "StandardDeviation\_Sales\_by\_Store"

colnames(Store\_Sales\_Variation)[3] <- "Mean\_Sales\_by\_Store"

*#Creating Coefficient of Variation for Sales by Store in Store\_Sales\_Variation dataframe*

Store\_Sales\_Variation<- mutate(Store\_Sales\_Variation,CV\_Sales\_by\_Store = (StandardDeviation\_Sales\_by\_Store/Mean\_Sales\_by\_Store)\*100)

*#------Finding Store with highest Standard deviation-------#*

*#Finding out the row with highest standard deviation*

Store\_Sales\_Variation[which.max(Store\_Sales\_Variation$StandardDeviation\_Sales\_by\_Store), ]

*#Storing store number with max std deviation value*

store\_sales\_max\_std <- Store\_Sales\_Variation[which.max(Store\_Sales\_Variation$StandardDeviation\_Sales\_by\_Store), ]$Store

*#Storing max std deviation value*

max\_sd <- Store\_Sales\_Variation[which.max(Store\_Sales\_Variation$StandardDeviation\_Sales\_by\_Store), ]$StandardDeviation\_Sales\_by\_Store

*#Storing CV value for max std deviation*

CV\_max\_sd <- Store\_Sales\_Variation[which.max(Store\_Sales\_Variation$StandardDeviation\_Sales\_by\_Store), ]$CV\_Sales\_by\_Store

*#Store with highest variation in Sales - Store 14 & Standard Deviation - 317570, C.V - 5.7137*

*#printing the output*

print(paste('Store no. ', store\_sales\_max\_std,

'has the maximum standard deviation of ', max\_sd, 'Coefficient of Variation = ',CV\_max\_sd ))

|  |  |  |  |
| --- | --- | --- | --- |
| A tibble: 1 × 4 | | | |
| Store | StandardDeviation\_Sales\_by\_Store | Mean\_Sales\_by\_Store | CV\_Sales\_by\_Store |
| <int> | <dbl> | <dbl> | <dbl> |
| 14 | 317569.9 | 2020978 | 15.71367 |

[1] "Store no. 14 has the maximum standard deviation of 317569.949475508 Coefficient of Variation = 15.7136736009483"

options(repr.plot.width = 14, repr.plot.height = 8)

*#Density Plot for Store 14*

Store\_14 <- data1[data1$Store == 14, ]

p <- ggplot(Store\_14, aes(x=Weekly\_Sales)) + geom\_density(color="darkblue", fill="lightblue",alpha=0.2)+

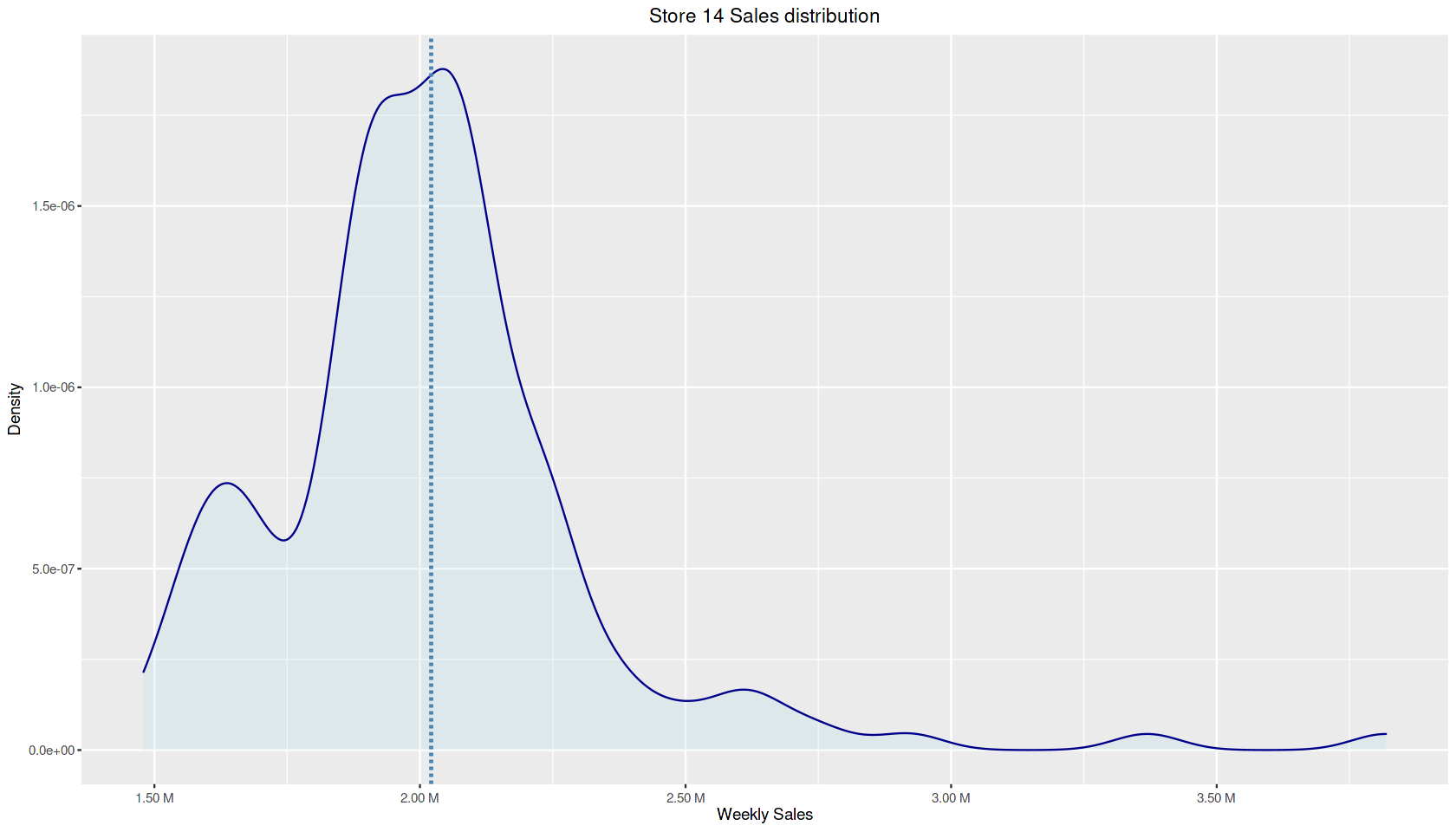
geom\_vline(aes(xintercept= mean(Weekly\_Sales)),color="steelblue", linetype="dashed", size=1)+

theme(axis.text.x = element\_text(vjust = 0.5, hjust=0.5))+ scale\_x\_continuous(labels = label\_number(suffix = " M", scale = 1e-6))+ ggtitle('Store 14 Sales distribution')+

theme(plot.title = element\_text(hjust = 0.5))+

xlab("Weekly Sales") + ylab("Density")

p



**Insights:**

* Store 14 has highest SD 317569.95 (317.5K) & CV = 15.714
* Store 14 sales are right skewed i.e, It had very high sales in few weeks which resulted in increasing of Standard deviation
* Inspite of having highest Standard deviation Store 14 CV is relatively low

Which store/s has good quarterly growth rate in Q3’2012

*#Creating new dataframe to do alterations*

data2<-data1

*#Creating a month- year column in data2*

data2$month\_Year = substr(data2$Date, 4, 10)

*#Subsetting Q3-2012 data (i.e, 07-2012,08-2012,09-2012), Q2-2012 data (i.e, 04-2012,05- 2012,06-2012)*

Q3\_2012 <- filter(data2,month\_Year == "07-2012" | month\_Year== "08-2012" | month\_Year== "09-2012")

Q2\_2012 <- filter(data2,month\_Year == "04-2012" | month\_Year== "05-2012" | month\_Year== "06-2012")

*#Aggregating sales by store for Q3-2012*

Q3\_2012\_Sales<-summarise(group\_by(Q3\_2012,Store),sum(Weekly\_Sales))

*#Changing column names*

colnames(Q3\_2012\_Sales)[2] <- "Q3\_2012\_Sales\_by\_Store"

*#Aggregating sales by store each Q2-2012*

Q2\_2012\_Sales<-summarise(group\_by(Q2\_2012,Store),sum(Weekly\_Sales))

*#Changing column names*

colnames(Q2\_2012\_Sales)[2] <- "Q2\_2012\_Sales\_by\_Store"

*#merging two quarters data by store*

Q3\_2012\_Growthrate <- merge ( Q2\_2012\_Sales , Q3\_2012\_Sales , by = 'Store')

*#Creating Growth rate column for Sales by Store in the above dataframe*

Q3\_2012\_Growthrate <- mutate(Q3\_2012\_Growthrate, Growth\_Rate = ((Q3\_2012\_Sales\_by\_Store - Q2\_2012\_Sales\_by\_Store)\*100) / Q2\_2012\_Sales\_by\_Store)

*#Creating only positive growth rates*

positive\_growthrate <- filter(Q3\_2012\_Growthrate, Growth\_Rate > 0 )

positive\_growthrate<-arrange(positive\_growthrate, desc(Growth\_Rate))

View(positive\_growthrate)

a<- positive\_growthrate$Store

*#printing the output*

print(paste(c('The positive growth rate Stores are', a),collapse=" " ))

print(paste('Store',positive\_growthrate[1,1], 'has highest growth rate & it is',positive\_growthrate[1,4]))

*# Store 7 -13.33% , Store 16 - 8.49% , Store 35 - 4.47% and 7 more stores with positive growth rates.*

|  |  |  |  |
| --- | --- | --- | --- |
| A data.frame: 10 × 4 | | | |
| Store | Q2\_2012\_Sales\_by\_Store | Q3\_2012\_Sales\_by\_Store | Growth\_Rate |
| <int> | <dbl> | <dbl> | <dbl> |
| 7 | 7290859 | 8262787 | 13.3307760 |
| 16 | 6564336 | 7121542 | 8.4883781 |
| 35 | 10838313 | 11322421 | 4.4666372 |
| 26 | 13155336 | 13675692 | 3.9554775 |
| 39 | 20214128 | 20715116 | 2.4784040 |
| 41 | 17659943 | 18093844 | 2.4569801 |
| 44 | 4306406 | 4411251 | 2.4346377 |
| 24 | 17684219 | 17976378 | 1.6520877 |
| 40 | 12727738 | 12873195 | 1.1428413 |
| 23 | 18488883 | 18641489 | 0.8253951 |

[1] "The positive growth rate Stores are 7 16 35 26 39 41 44 24 40 23"

[1] "Store 7 has highest growth rate & it is 13.330776030738"

options(repr.plot.width = 14, repr.plot.height = 8)

*# Visual representation of growth rates*

c<-ggplot(data=Q3\_2012\_Growthrate, aes(x=Store, y=Growth\_Rate)) +geom\_bar(stat ="identity",fill="steelblue")+

ggtitle('Growth rates of Q3- 2012')+

theme(plot.title = element\_text(hjust = 0.5))+

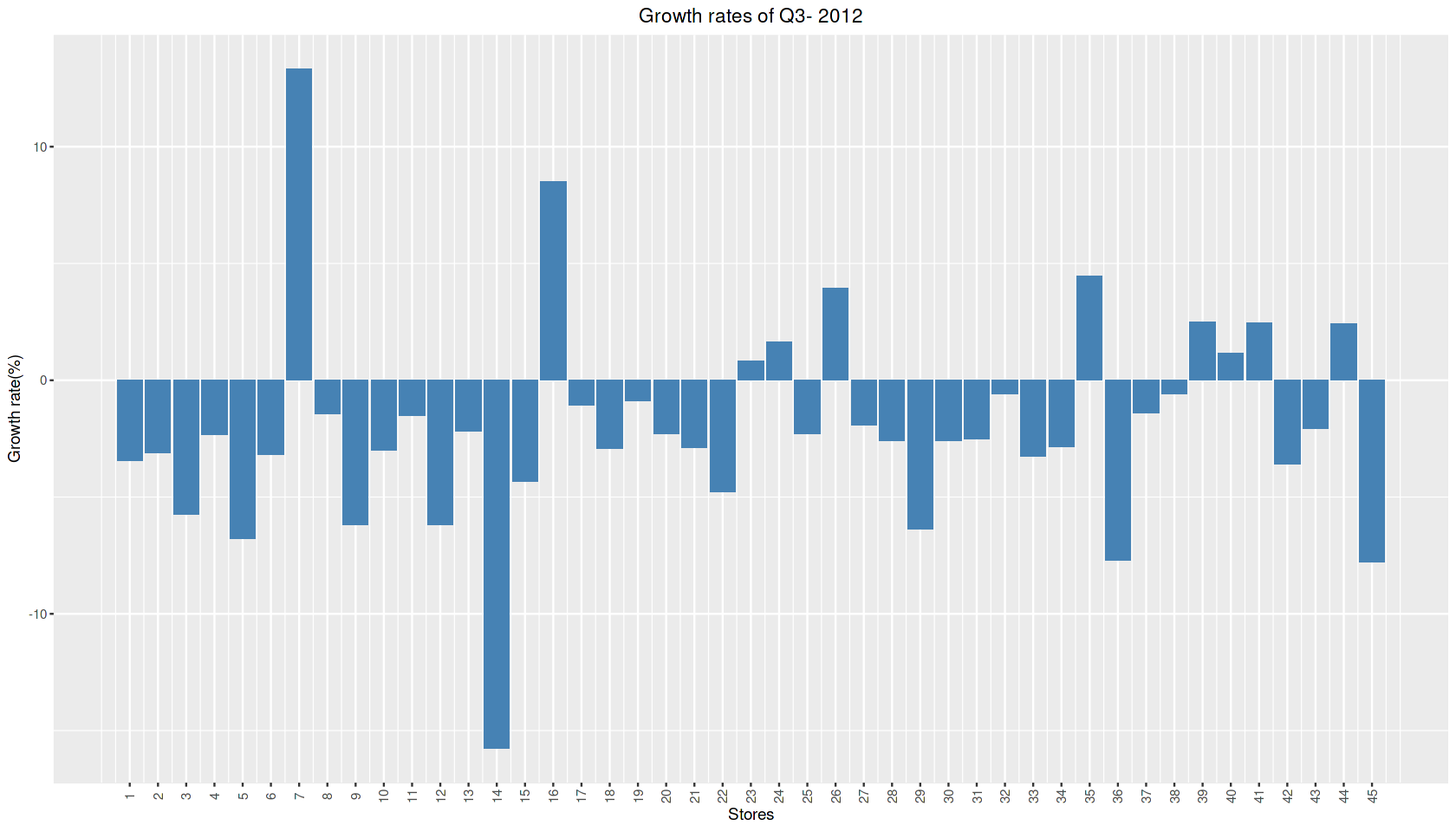
xlab("Stores") + ylab("Growth rate(%)") +

scale\_x\_continuous("Stores", labels = as.character(Q3\_2012\_Growthrate$Store), breaks =

Q3\_2012\_Growthrate$Store)+

theme(axis.text.x = element\_text(angle = 90,vjust = 0.5, hjust=0.5))

c



**Insights:**

* Store 7 has highest growth rate 13.33%, Store 16 second highest growth rate 8.49%, Store 35 third highest growth rate 4.47%
* It is observed that seven other stores has positive growth rates and they are 26,39,41,44,24,40,23
* Store 14 has highest negative growth rate.

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

*#Creating Holidays Data dataframe*

Holiday\_date <- c("12-02-2010", "11-02-2011", "10-02-2012", "08-02-2013","10-09-2010", "09-09-2011", "07-09-2012", "06-09-2013","26-11-2010", "25-11-2011", "23-11-2012", "29- 11-2013","31-12-2010", "30-12-2011", "28-12-2012", "27-12-2013")

Events <-c(rep("Super Bowl", 4), rep("Labour Day", 4),rep("Thanksgiving", 4), rep("Christmas", 4))

Holidays\_Data <- data.frame(Events,Holiday\_date)

*#merging both dataframes*

data3<-merge(data1,Holidays\_Data, by.x= "Date", by.y="Holiday\_date", all.x = TRUE)

*#Replacing null values in Event with No\_Holiday*

data3$Events = as.character(data3$Events)

data3$Events[is.na(data3$Events)]= "No\_Holiday"

head(data3)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A data.frame: 6 × 9 | | | | | | | | | | |
|  | Date | Store | Weekly\_Sales | Holiday\_Flag | Temperature | Fuel\_Price | CPI | Unemployment | Events |
|  | <chr> | <int> | <dbl> | <int> | <dbl> | <dbl> | <dbl> | <dbl> | <chr> |
| 1 | 01-04-2011 | 1 | 1495064.8 | 0 | 59.17 | 3.524 | 214.8372 | 7.682 | No\_Holiday |
| 2 | 01-04-2011 | 2 | 1800171.4 | 0 | 55.43 | 3.524 | 214.4887 | 7.931 | No\_Holiday |
| 3 | 01-04-2011 | 3 | 374556.1 | 0 | 68.76 | 3.524 | 218.2114 | 7.574 | No\_Holiday |
| 4 | 01-04-2011 | 4 | 1900246.5 | 0 | 56.99 | 3.521 | 128.7199 | 5.946 | No\_Holiday |
| 5 | 01-04-2011 | 5 | 314316.5 | 0 | 61.50 | 3.524 | 215.4024 | 6.489 | No\_Holiday |
| 6 | 01-04-2011 | 6 | 1459276.8 | 0 | 62.25 | 3.524 | 216.3841 | 6.855 | No\_Holiday |

*#Creating dataframe the mean of sales for No\_Holiday and also mean of sales for different events*

Holiday\_Sales<-aggregate(Weekly\_Sales ~ Events, data = data3, mean)

*#Changing column names*

colnames(Holiday\_Sales)[2] <- "Mean\_Sales\_by\_Event\_Type"

View(Holiday\_Sales)

*# Christmas and Labour Day has negative impact on sales where as Thanks giving and Super Bowl has positive impact on sales*

|  |  |
| --- | --- |
| A data.frame: 5 × 2 | |
| Events | Mean\_Sales\_by\_Event\_Type |
| <chr> | <dbl> |
| Christmas | 960833.1 |
| Labour Day | 1042427.3 |
| No\_Holiday | 1041256.4 |
| Super Bowl | 1079128.0 |
| Thanksgiving | 1471273.4 |

*# checking negative impact based on holiday date and non- holiday date*

*#Filtering holiday dates and finding mean of Weekly Sales*

Holiday\_date <- filter(data3,Holiday\_Flag ==1)

Holiday\_Date\_Sales<-summarise(group\_by(Holiday\_date,Date),mean(Weekly\_Sales))

*#Caluclating mean of Weekly Sales for non holidays*

mean\_non\_holiday\_sales <- mean(filter(data3,Holiday\_Flag ==0)$Weekly\_Sales)

Holiday\_Date\_Sales$higher\_than\_non\_holiday <- Holiday\_Date\_Sales[,2] > mean\_non\_holiday\_sales

View(Holiday\_Date\_Sales)

|  |  |  |
| --- | --- | --- |
| A tibble: 10 × 3 | | |
| Date | mean(Weekly\_Sales) | higher\_than\_non\_holiday |
| <chr> | <dbl> | <lgl[,1]> |
| 07-09-2012 | 1074001.3 | TRUE |
| 09-09-2011 | 1039182.8 | FALSE |
| 10-02-2012 | 1111320.2 | TRUE |
| 10-09-2010 | 1014097.7 | FALSE |
| 11-02-2011 | 1051915.4 | TRUE |
| 12-02-2010 | 1074148.4 | TRUE |
| 25-11-2011 | 1479857.9 | TRUE |
| 26-11-2010 | 1462689.0 | TRUE |
| 30-12-2011 | 1023165.8 | FALSE |
| 31-12-2010 | 898500.4 | FALSE |

**Insights:**

* Super Bowl, Thanks giving, Labour day has higher sales than mean sales of a Non Holiday and creating positive impact on sales.
* 9th Sept 2011, 10th Sept 2010 ,30th Dec 2011, 31st Dec 2010 were the dates which created negative impact on sales
* All the dates related to Christmas have low sales than mean, whereas all the dates related to Super Bowl, Thanks giving have high sales than mean.
* It is interesting to note that Labour Day has overall positive impact on sales inspite of having two days days below mean value.

weekly\_sales <- aggregate(Weekly\_Sales~Date, data=data1,mean)

weekly\_sales$Date <-as.Date(weekly\_sales$Date, "%d-%m-%Y")

weekly\_sales <-arrange(weekly\_sales,Date)

weekly\_sales$Date <-factor(weekly\_sales$Date)

options(repr.plot.width = 14, repr.plot.height = 8)

*# plotting weekly mean sales*

d <- ggplot(data=weekly\_sales, aes(x=Date, y=Weekly\_Sales, group=1)) +

geom\_line(color="steelblue")+

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))+

scale\_x\_discrete(breaks = levels(weekly\_sales$Date)[c(T, rep(F, 9))])+

scale\_y\_continuous(labels = label\_number(suffix = " M", scale = 1e-6))+

theme(plot.title = element\_text(hjust = 0.5))+

xlab("Week") + ylab("Mean Sales of Week")

*#Plotting Christmas*

d +ggtitle('CHRISTMAS')+

geom\_point(aes(x = factor("2010-12-31"), y = 898500.4), color = "red", size = 2) +

geom\_point(aes(x = factor("2011-12-30"), y = 1023165.8), color = "red", size = 2) +

geom\_hline(aes(yintercept = mean\_non\_holiday\_sales), linetype="dashed")

*#Plotting Labourday*

d + ggtitle('LABOUR DAY')+

geom\_point(aes(x = factor("2010-09-10"), y = 1014097.7), color = "deeppink", size = 2) +

geom\_point(aes(x = factor("2011-09-09"), y = 1039182.8), color = "deeppink", size = 2) +

geom\_point(aes(x = factor("2012-09-07"), y = 1074001.3), color = "deeppink", size = 2) +

geom\_hline(aes(yintercept = mean\_non\_holiday\_sales), linetype="dashed")

*#Plotting Thanks Giving*

d + ggtitle('THANKS GIVING')+

geom\_point(aes(x = factor("2010-11-26"), y = 1462689.0), color = "indianred4", size = 2) +

geom\_point(aes(x = factor("2011-11-25"), y = 1479857.9), color = "indianred4", size = 2) +

geom\_hline(aes(yintercept = mean\_non\_holiday\_sales), linetype="dashed")

*#Plotting Superbowl*

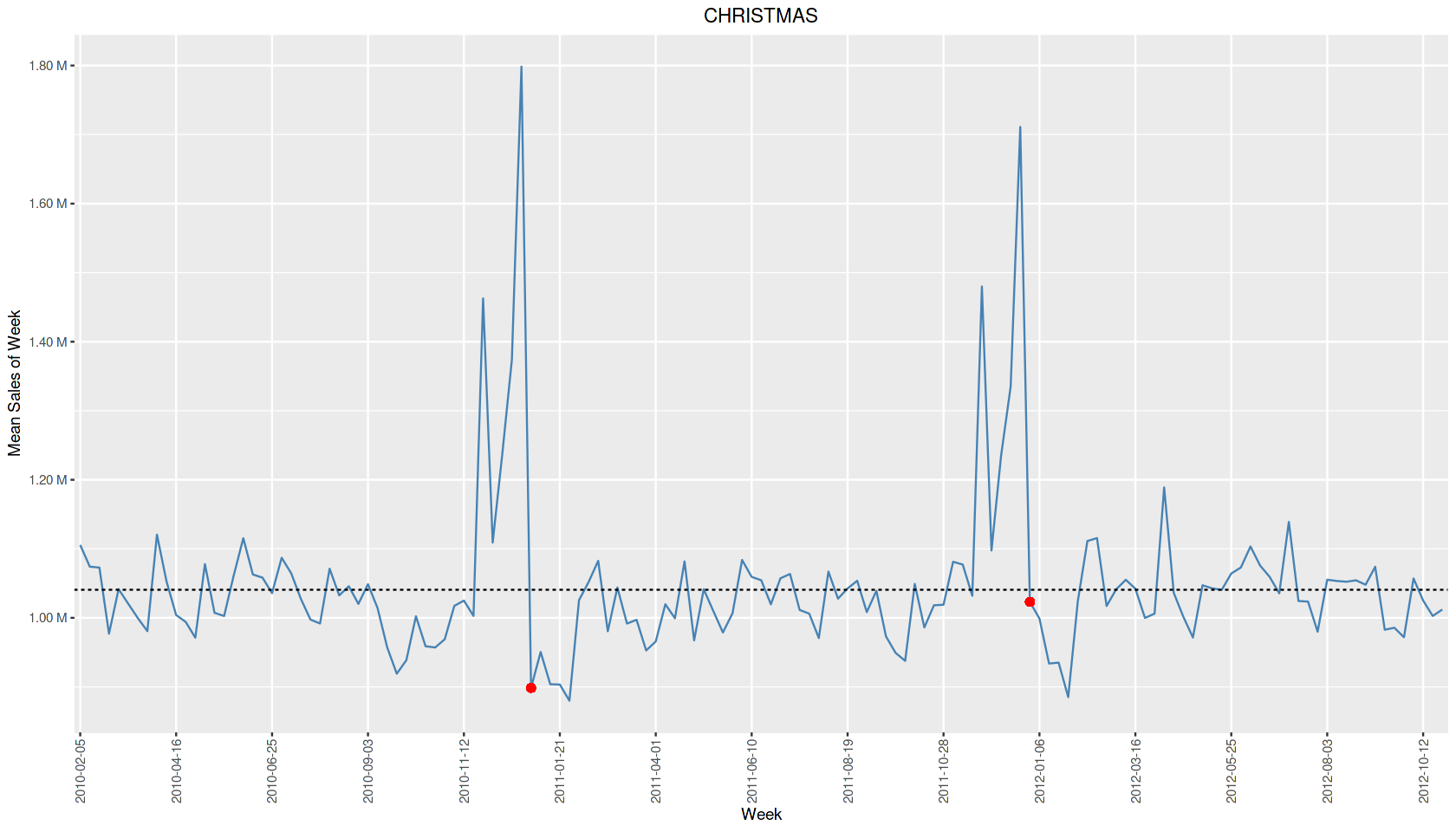
d + ggtitle('SUPER BOWL')+

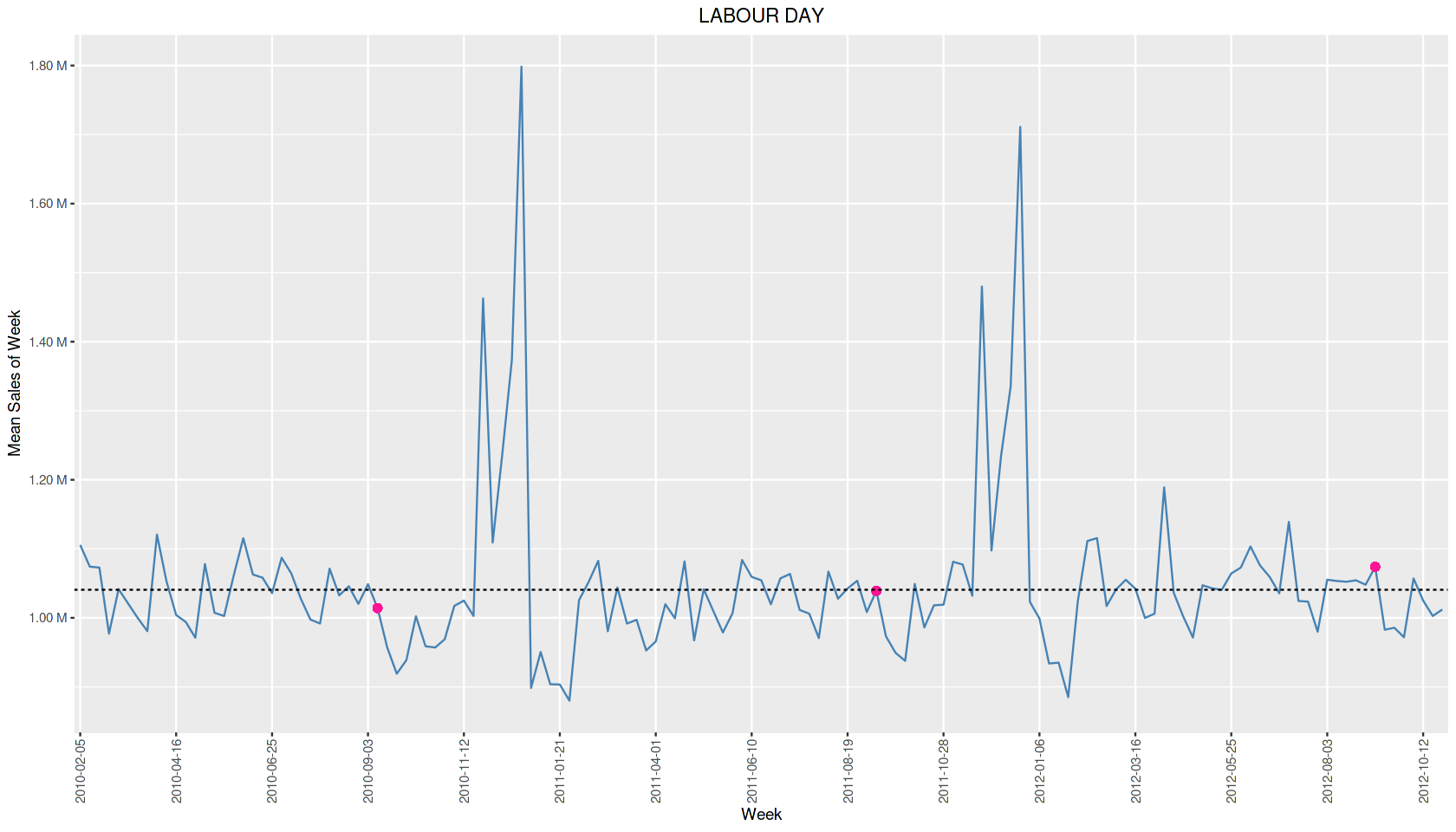
geom\_point(aes(x = factor("2010-02-12"), y = 1074148.4), color = "goldenrod4", size = 2) +

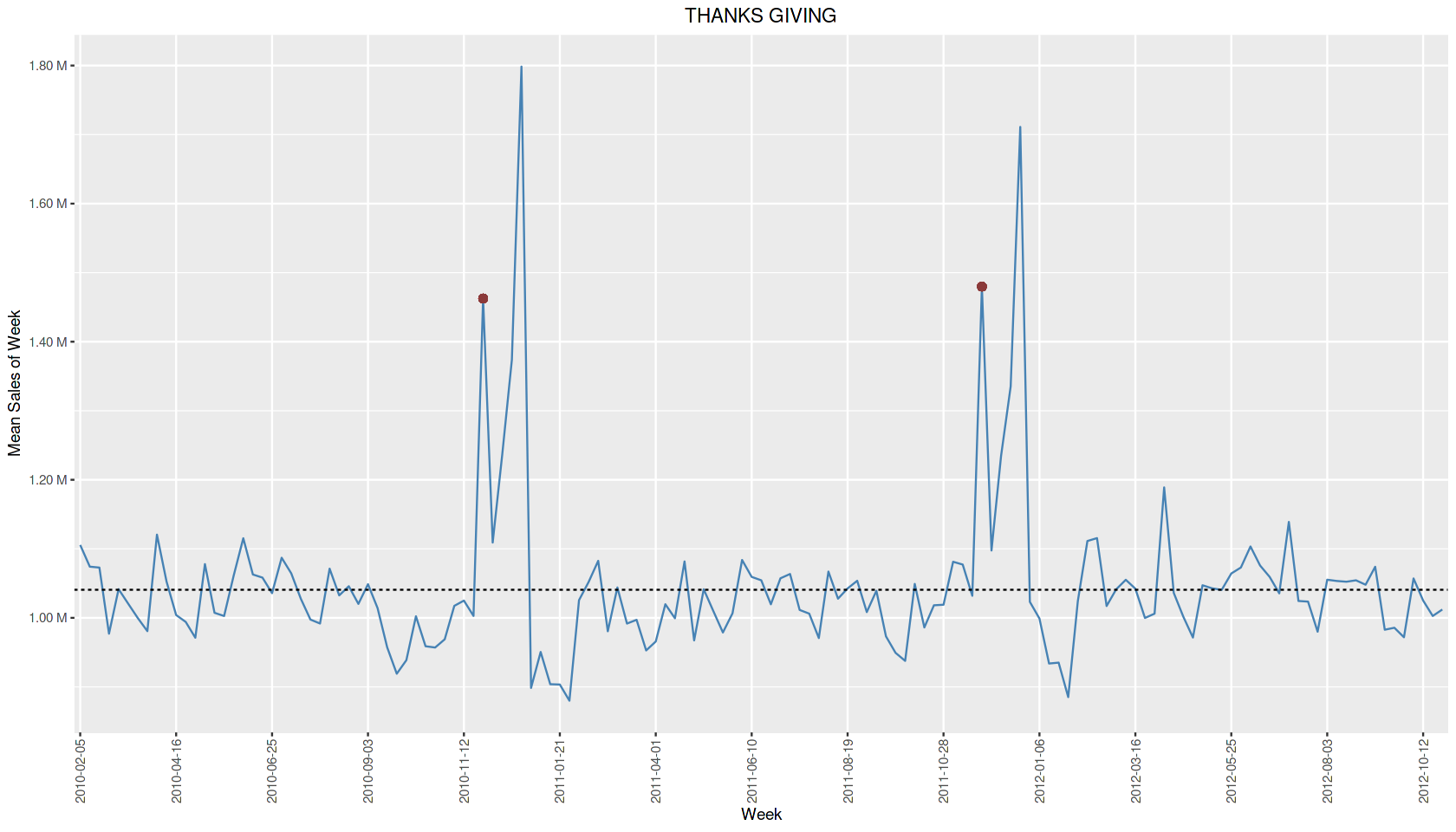
geom\_point(aes(x = factor("2011-02-11"), y = 1051915.4), color = "goldenrod4", size = 2) +

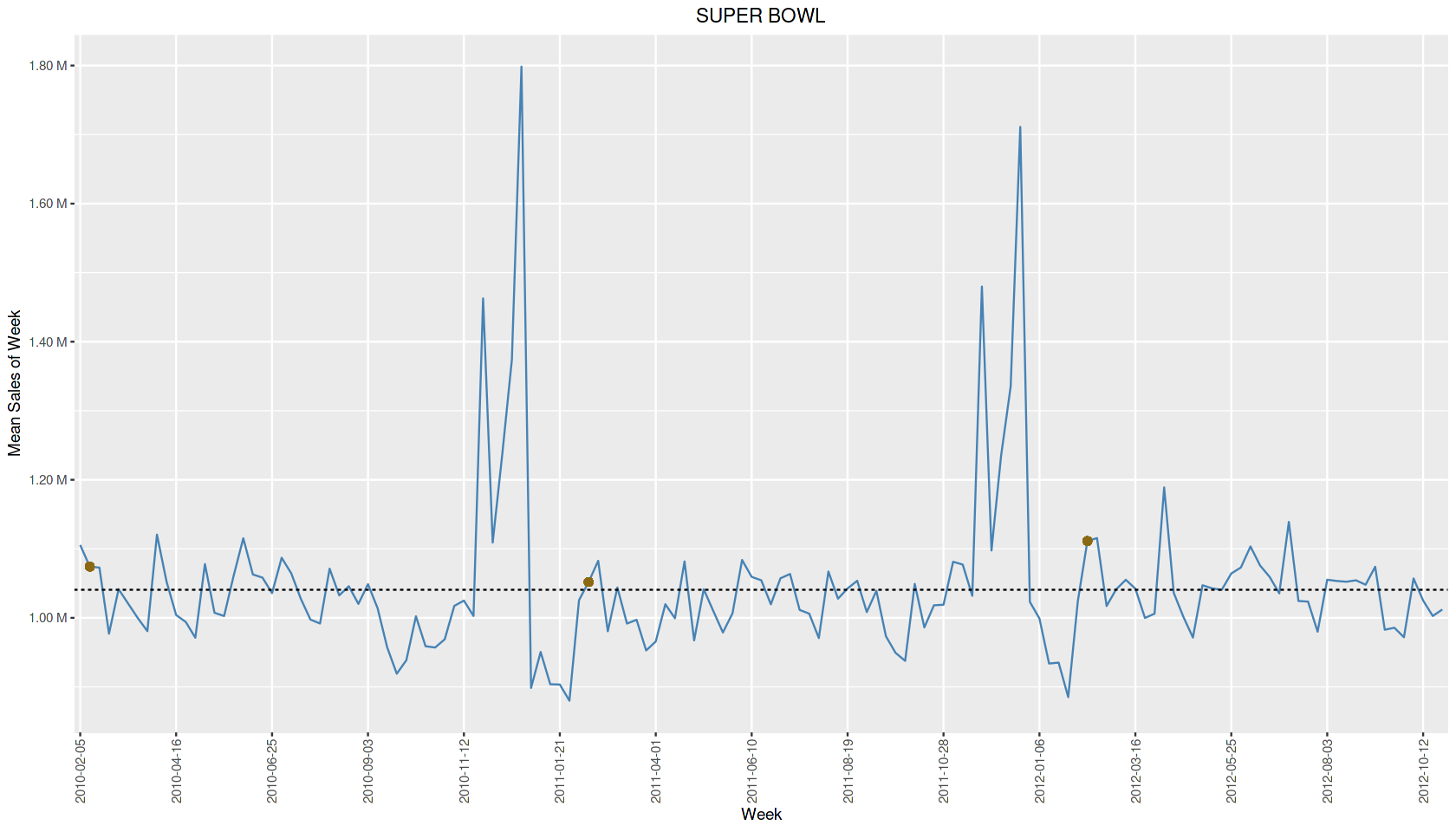
geom\_point(aes(x = factor("2012-02-10"), y = 1111320.2), color = "goldenrod4", size = 2) +

geom\_hline(aes(yintercept = mean\_non\_holiday\_sales), linetype="dashed")









**Insights:**

* Though Super Bowl, Labour day have higher sales they are very close to mean
* Thanks giving does create a high positive impact on sales than others
* Christmas Holiday Flag has lower sales than mean sales of a Non Holiday. Both the dates (30- Dec-2011, 31-Dec-2010 ) related to Christmas has negative impact on sales.
* **However, from the graph it is clear that the week just before the Christmas bagged highest sales. It may be because customers did shopping beforehand for preparation/ Advent is popularly celebrated there**

Provide a monthly and semester view of sales in units and give insights

*#Converting date into factor*

x<-as.factor(data2$Date)

*#defining what is the original format of date*

abis<-strptime(x,format="%d-%m-%Y")

*#defining what is the desired format of your date*

data2$Mon\_Year<-as.Date(abis,format="%Y-%m-%d")

data2$Mon\_Year = as.yearmon(data2$Mon\_Year)

*#Aggregating data by 'Month -Year' and Finding sum of 'Weekly\_Sales' and convrting it into dataframe*

Month\_Year\_Sales<-summarise(group\_by(data2,Mon\_Year),sum(Weekly\_Sales))

colnames(Month\_Year\_Sales)[2] <- "Sales\_by\_Month"

Month\_Year\_Sales<- as.data.frame(Month\_Year\_Sales)

*#Converting year-mon to factor for plotting so that order wont change*

Month\_Year\_Sales$Mon\_Year<- as.character(Month\_Year\_Sales$Mon\_Year)

Month\_Year\_Sales$Mon\_Year<- factor(Month\_Year\_Sales$Mon\_Year, levels=Month\_Year\_Sales$Mon\_Year)

*#plotting line graph as it is time series data*

p <- ggplot(data=Month\_Year\_Sales, aes(x=Mon\_Year, y=Sales\_by\_Month, group=1)) +

geom\_line(color="steelblue")+

geom\_point()+

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))+

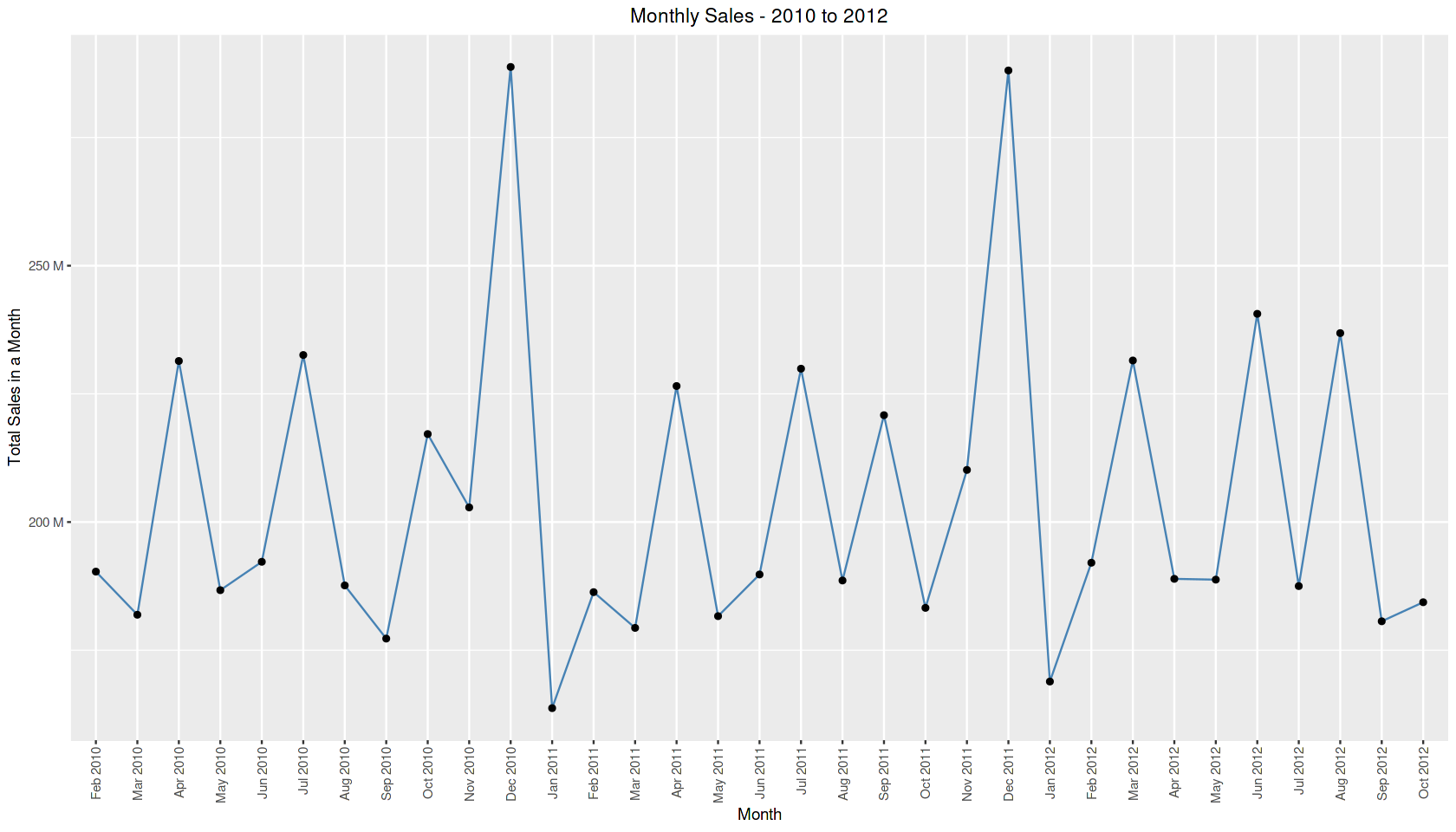
scale\_y\_continuous(labels = label\_number(suffix = " M", scale = 1e-6))+

ggtitle('Monthly Sales - 2010 to 2012')+

theme(plot.title = element\_text(hjust = 0.5))+

xlab("Month") + ylab("Total Sales in a Month")

p



\

*#sales vs semester - using lubridate*

*#converting to date format*

data2$Date <- dmy(data2$Date)

*#creating semester column with year*

data2$sem <- semester(data2$Date, with\_year=TRUE)

*#creating a dataframe 's' which has total sales for every sem*

s <- aggregate(Weekly\_Sales~sem,data=data2, sum)

*# Addding a new column by Rewriting semester and yr to different format*

s$sem\_year <- paste(substr(s$sem,1,4),'-S',substr(s$sem,6,6),sep = '')

*#Plotting the graph semester vs Sales*

q <- ggplot(data=s, aes(x=sem\_year, y=Weekly\_Sales, group=1)) +

geom\_line(color="steelblue")+

geom\_point()+

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))+

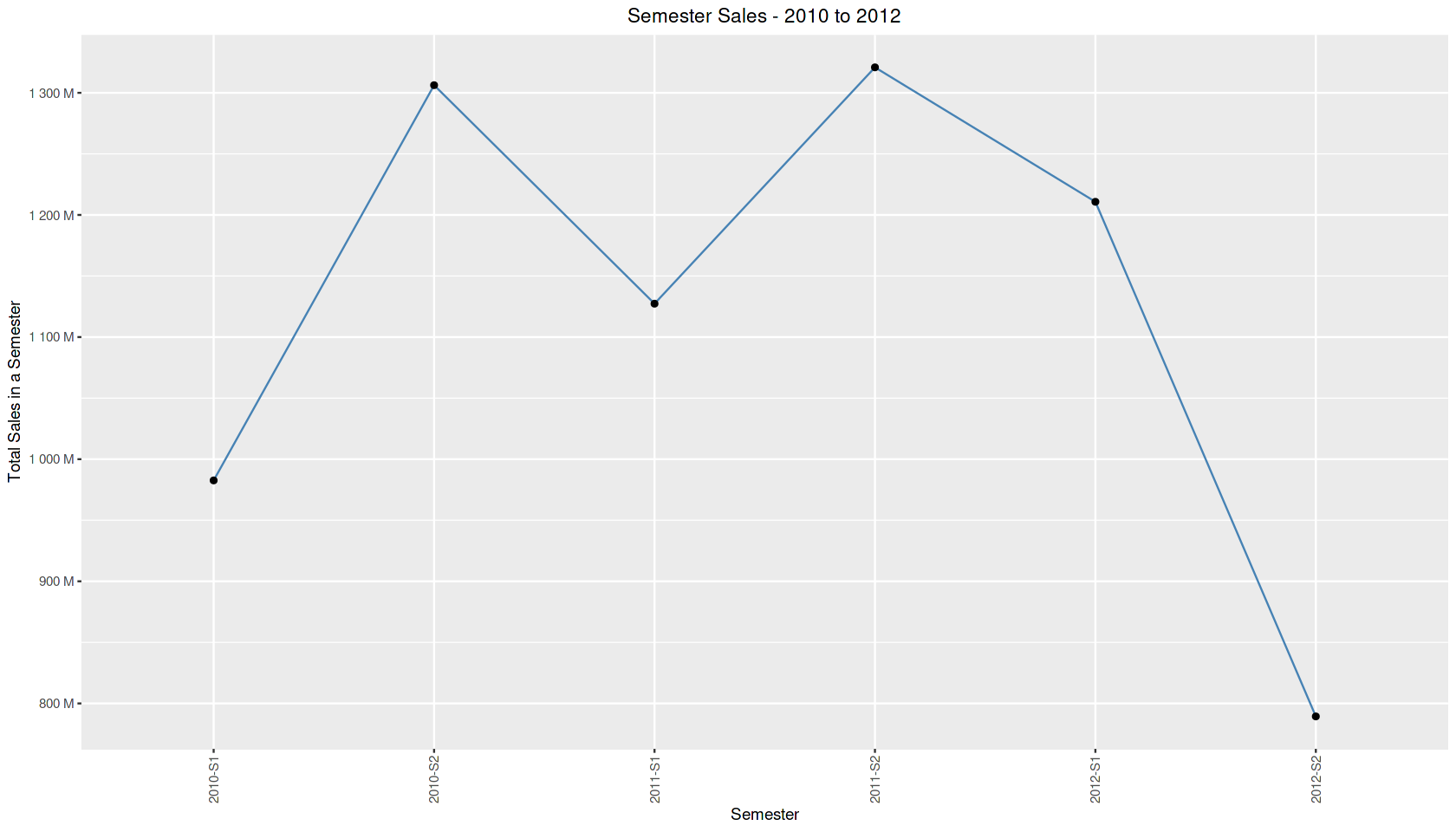
scale\_y\_continuous(labels = label\_number(suffix = " M", scale = 1e-6))+

ggtitle('Semester Sales - 2010 to 2012')+

theme(plot.title = element\_text(hjust = 0.5))+

xlab("Semester") + ylab("Total Sales in a Semester")

q



**Insights:**

* The sales are highest in December and Lowest in January
* The sales are higher in second semester of every year
* The plot shows a drop in S2-2012 & S1-2010. It is due to absence of Jan data in S1-2010 & Nov-Dec 2012 data in S2-2012.

**STASTICAL MODEL USING MULTI LINEAR REGRESSION**

**Creating a dataframe with required columns**

*#creating same data for alterations*

data4 <- data1

*#selecting only first store as prediction Required only for first Store*

data4<- dplyr::filter(data4, Store ==1)

*#changing date column in dataframe to date format & arranging in ascending order as per dates*

data4$Date <- lubridate::dmy(data4$Date)

data4 <- dplyr::arrange(data4,Date)

*#Creating a week number,month,quarter column in dataframe*

data4$Week\_Number <- seq(1:length(unique(data4$Date)))

*#adding quarter & month columns*

data4$month <- lubridate::month(data4$Date)

data4$quarter <- lubridate::quarter(data4$Date)

*##Creating a event type dataframe##*

*# creating Holiday\_date vector*

Holiday\_date <- c("12-02-2010", "11-02-2011", "10-02-2012", "08-02-2013","10-09-2010", "09-09-2011", "07-09-2012", "06-09-2013","26-11-2010", "25-11-2011", "23-11-2012", "29-11-2013","31-12-2010", "30-12-2011", "28-12-2012", "27-12-2013")

*#assigning date format to Holiday\_date vector*

Holiday\_date <- lubridate::dmy(Holiday\_date)

*#Creating Events vector*

Events <-c(rep("Super Bowl", 4), rep("Labour Day", 4),rep("Thanksgiving", 4), rep("Christmas", 4))

*#Creating dataframe with Events and date*

Holidays\_Data <- data.frame(Events,Holiday\_date)

*#merging both dataframes*

data4<-merge(data4,Holidays\_Data, by.x= "Date", by.y="Holiday\_date", all.x = TRUE)

*#Replacing null values in Event with No\_Holiday*

data4$Events = as.character(data4$Events)

data4$Events[is.na(data4$Events)]= "No\_Holiday"

**Scatter Plot for data visualisation**

*#linear regression graph*

par(mfrow=c(3,3))

for(i in 3:11){

plot(data4[,i],

data4$Weekly\_Sales,

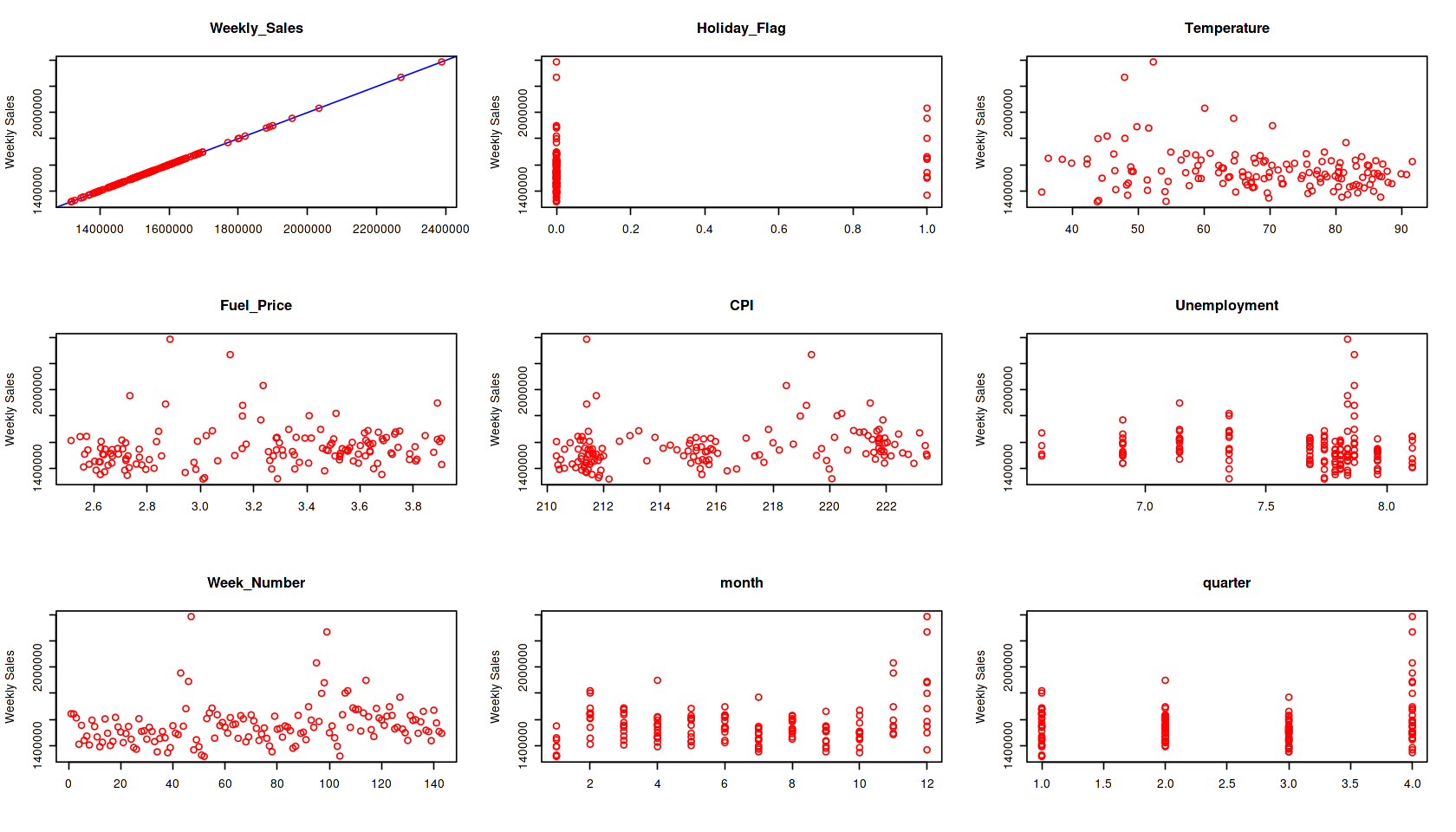
main=names(data4[i]),

ylab="Weekly Sales", xlab =" ",

col='red',

abline(lm(data4[,i] ~ data4$Weekly\_Sales, data = data4), col = "blue"))

}



**Detection and removal of outliers using Bi Variate Box Plots**

*#Boxplot for checking outliers & removing them*

par(mfrow=c(1,1))

*#Creating a dataframe for outlier treatment*

data5 <- data4

*#As we are predicting sales, Thought of removing outliers in Sales based on Various parameters*

*#Temperature Outlier treatment -- found 5 outlier and removed them*

boxplot(data5$Weekly\_Sales ~ cut(data5$Temperature, pretty(data5$Temperature)), main="Temperature vs Weekly Sales", xlab ="Temperature", ylab="Weekly Sales", cex.axis=0.5, col="Steel Blue")

outliers\_temp <- boxplot(data5$Weekly\_Sales ~ cut(data5$Temperature, pretty(data5$Temperature)), main="Temperature vs Weekly Sales", cex.axis=0.5,plot=FALSE)$out

data5<- data5[-which(data5$Weekly\_Sales %in% outliers\_temp),]

*#CPI Outlier treatment-found 1 outlier and removed them*

boxplot(data5$Weekly\_Sales ~ cut(data5$CPI, pretty(data5$CPI)), main="CPI vs Weekly Sales",xlab ="CPI", ylab="Weekly Sales", cex.axis=0.5,col="Steel Blue")

outliers\_CPI <- boxplot(data5$Weekly\_Sales ~ cut(data5$CPI, pretty(data5$CPI)), main="CPI vs Weekly Sales", cex.axis=0.5,plot=FALSE)$out

data5<- data5[-which(data5$Weekly\_Sales %in% outliers\_CPI),]

*#Unemployment outlier treatment--found 3 outlier and removed them*

boxplot(data5$Weekly\_Sales ~ cut(data5$Unemployment, pretty(data5$Unemployment)), main="Unemployment vs Weekly Sales",xlab ="Unemployment", ylab="Weekly Sales", cex.axis=0.5,col="Steel Blue")

outliers\_Unemployment <- boxplot(data5$Weekly\_Sales ~ cut(data5$Unemployment, pretty(data5$Unemployment)), main="Unemployment vs Weekly Sales", cex.axis=0.5,plot=FALSE)$out

data5<- data5[-which(data5$Weekly\_Sales %in% outliers\_Unemployment),]

*#fuel price outlier treatment -- found 2 outliers and removed*

boxplot(data5$Weekly\_Sales ~ cut(data5$Fuel\_Price, pretty(data5$Fuel\_Price)), main="Fuel\_Price vs Weekly Sales", xlab ="Fuel Price", ylab="Weekly Sales", cex.axis=0.5,col="Steel Blue")

outliers\_fuel\_price <- boxplot(data5$Weekly\_Sales ~ cut(data5$Fuel\_Price, pretty(data5$Fuel\_Price)), main="Fuel\_Price vs Weekly Sales", cex.axis=0.5,plot=FALSE)$out

data5<- data5[-which(data5$Weekly\_Sales %in% outliers\_fuel\_price),]

*#Outlier treatment for Holiday Flag - No outliers found*

boxplot(data5$Weekly\_Sales ~ data5$Holiday\_Flag, main = 'Weekly Sales - Holiday\_Flag',xlab ="Holiday Flag", ylab="Weekly Sales",col="Steel Blue" )

*#outlier treatment for month - 4 outliers found and removed*

boxplot(data5$Weekly\_Sales ~ data5$month, main = 'Weekly Sales - month', xlab ="Month", ylab="Weekly Sales", col="Steel Blue")

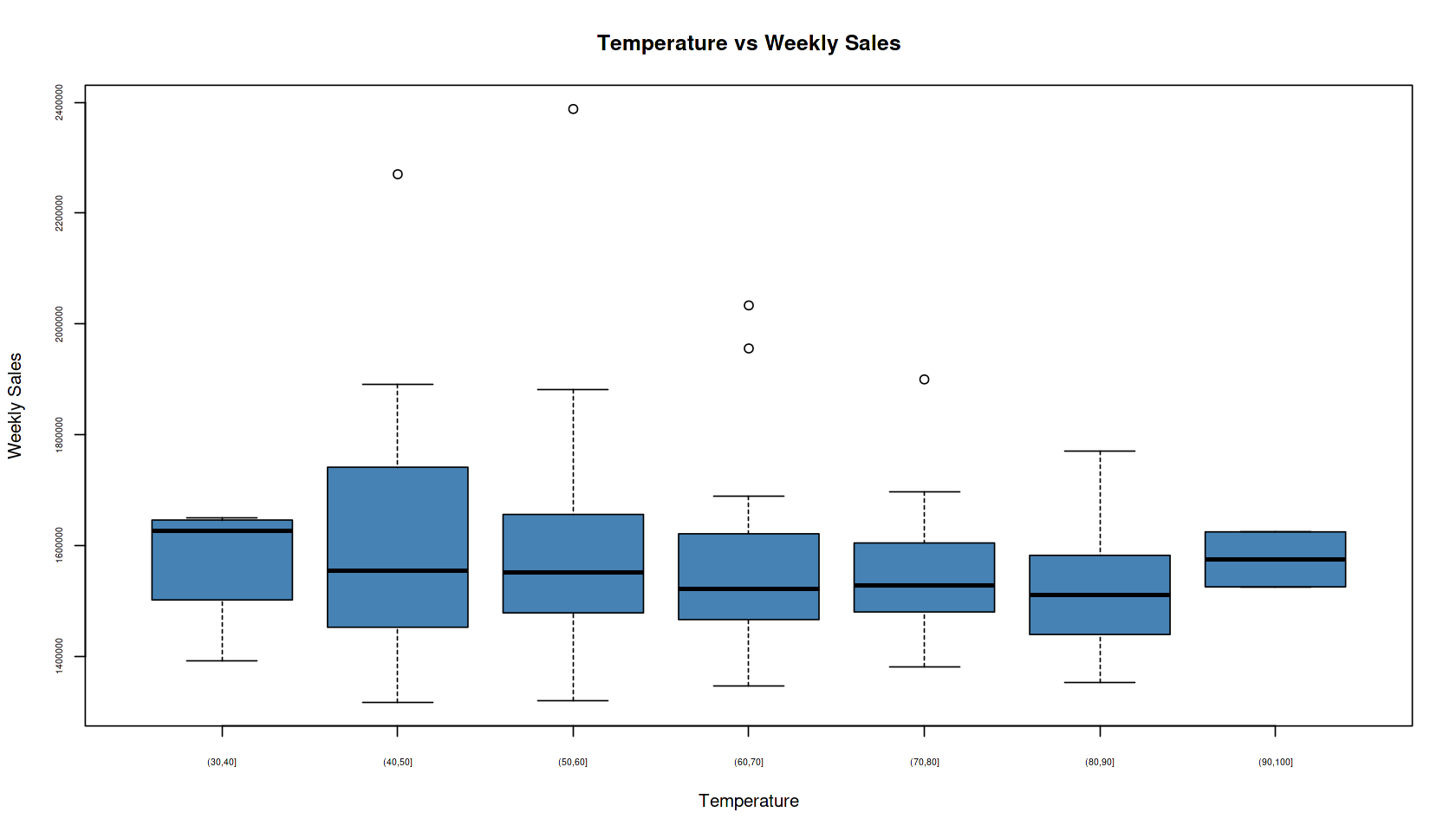
outliers\_month <- boxplot(data5$Weekly\_Sales ~ data5$month, main = 'Weekly Sales - month',plot=FALSE)$out

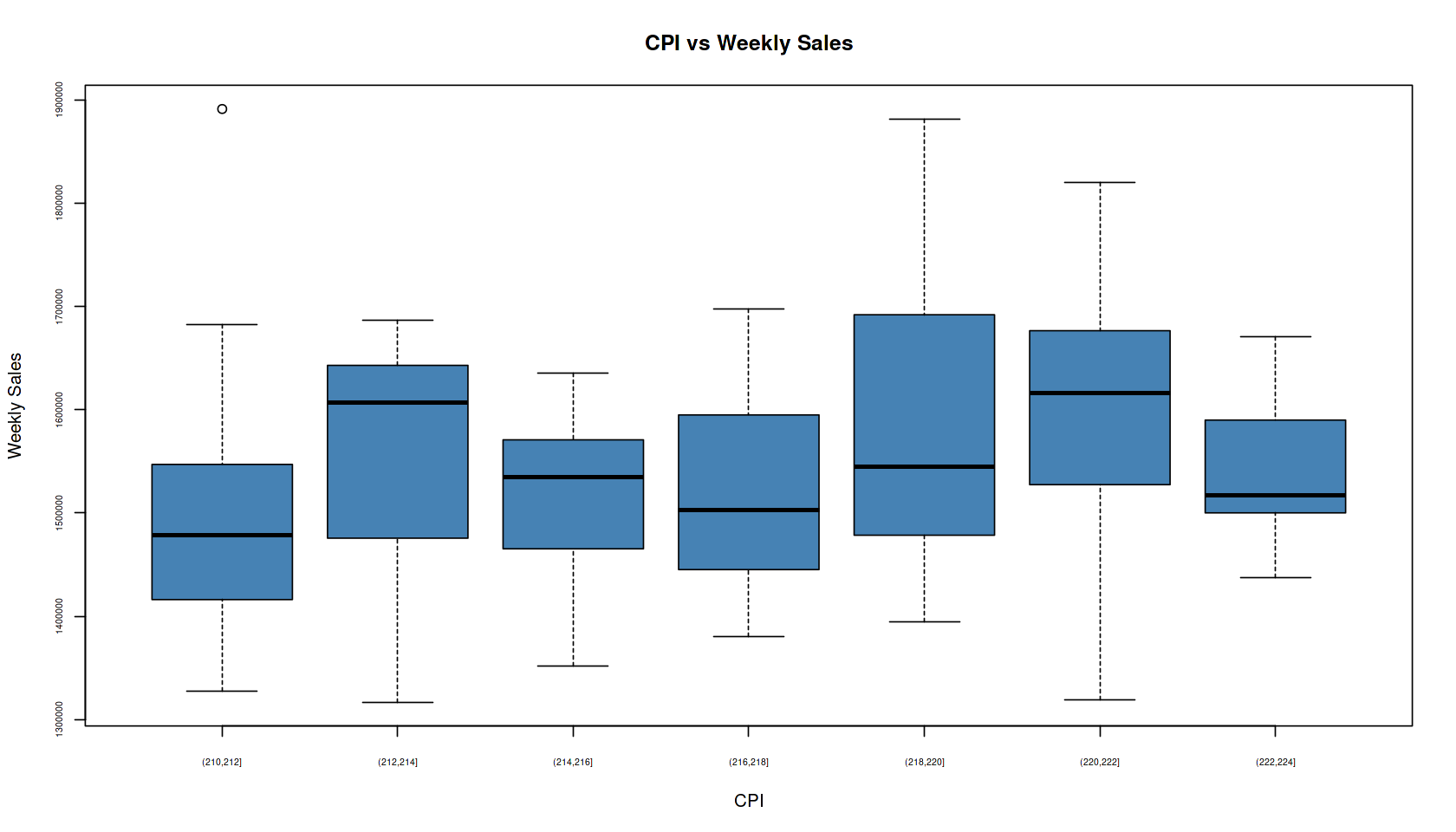
data5<- data5[-which(data5$Weekly\_Sales %in% outliers\_month),]

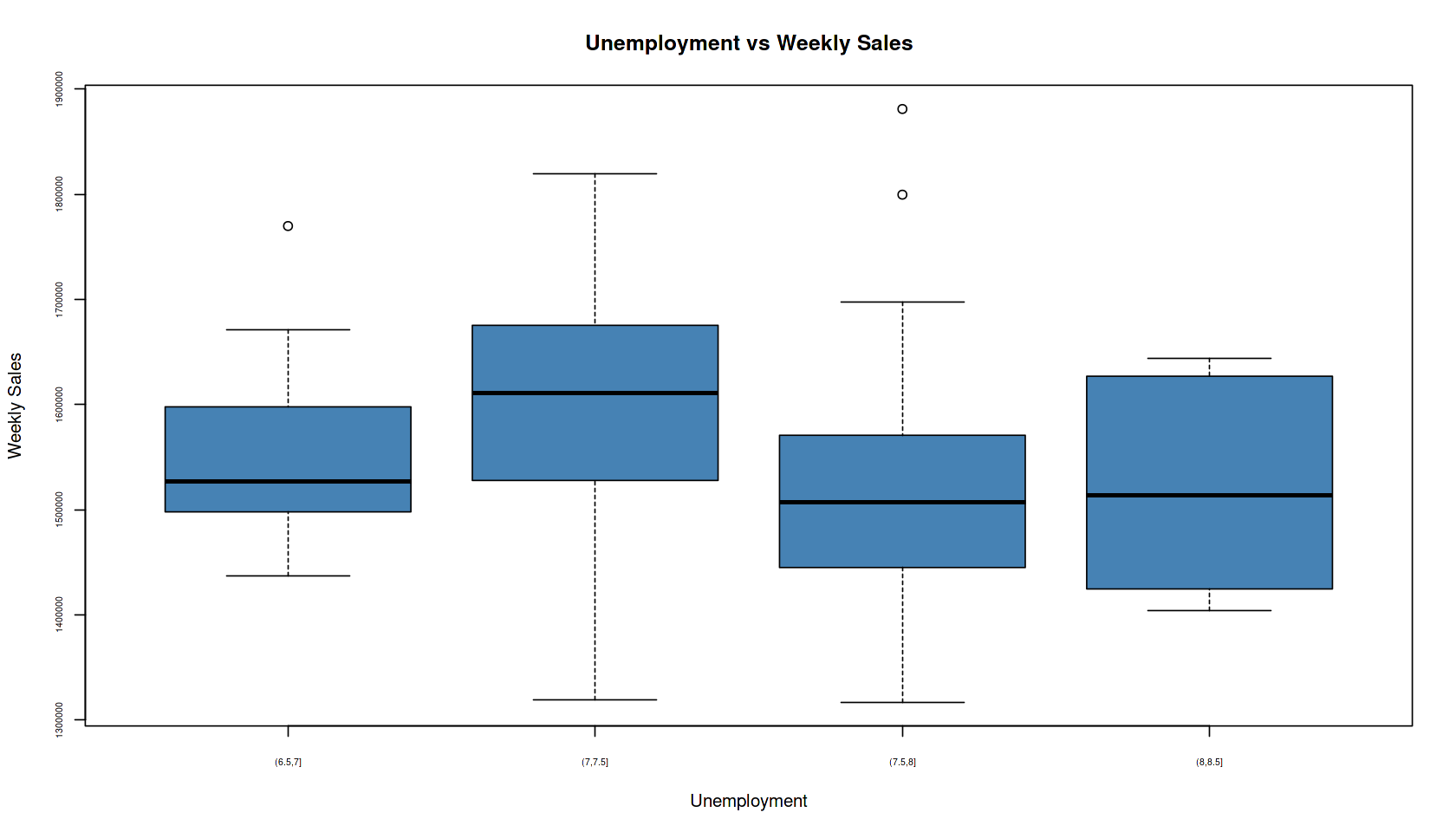
*#outlier treatment for quarter - 2 outliers found and removed*

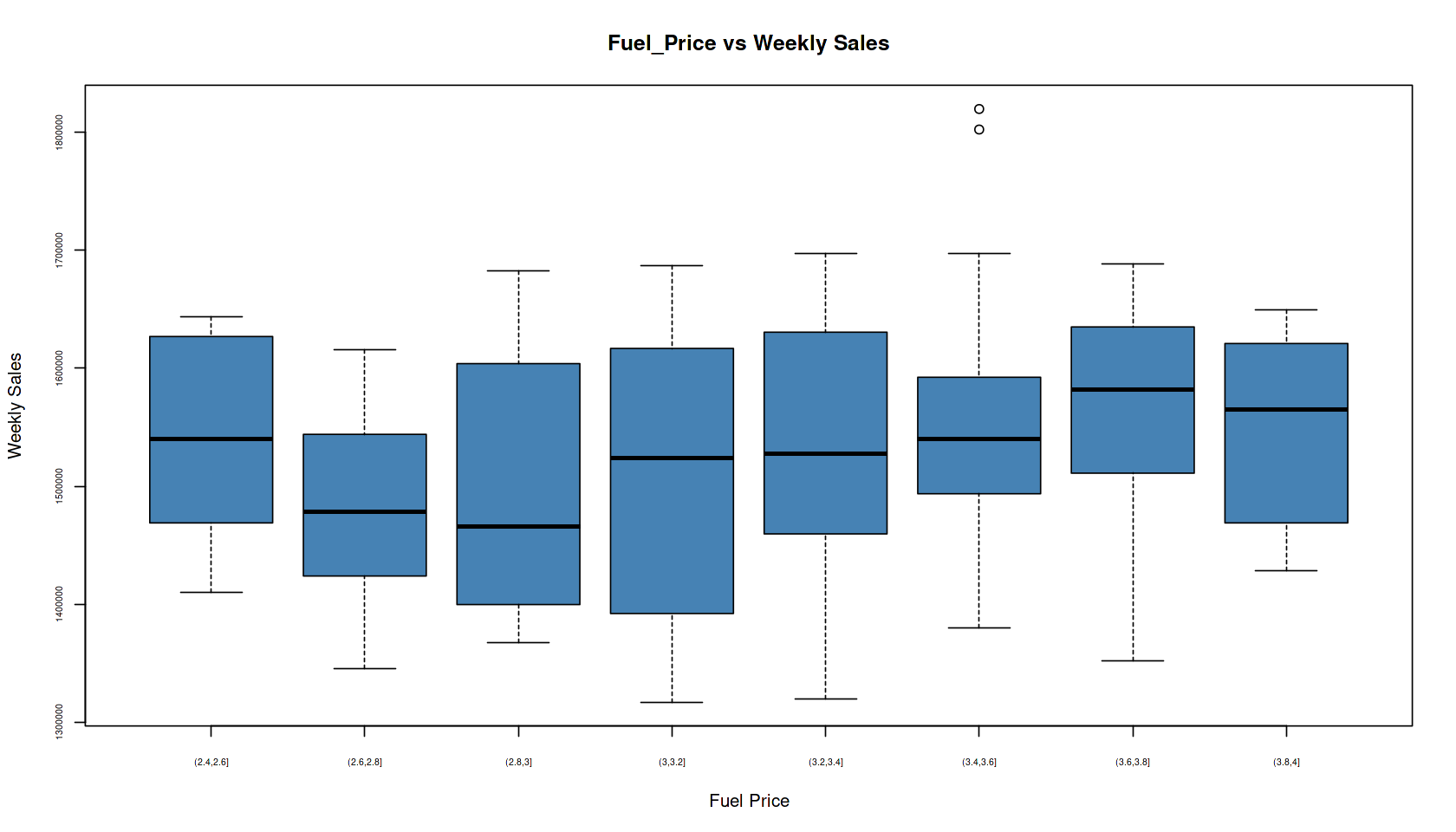
outliers\_quarter <- boxplot(data5$Weekly\_Sales ~ data5$quarter, main = 'Weekly Sales - quarter',xlab ="Quarters", ylab="Weekly Sales", col="Steel Blue")$out

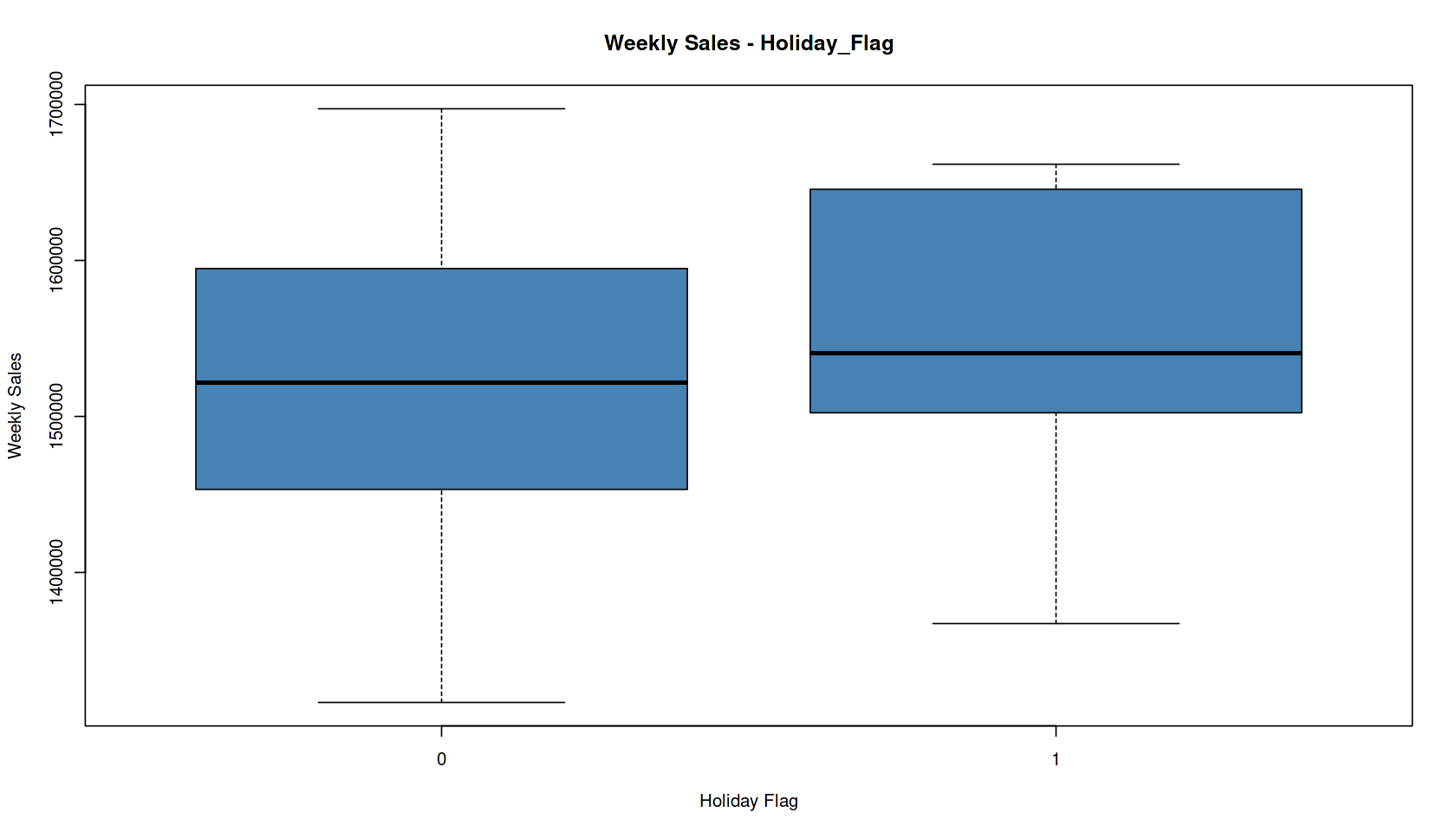
data5<- data5[-which(data5$Weekly\_Sales %in% outliers\_quarter),]

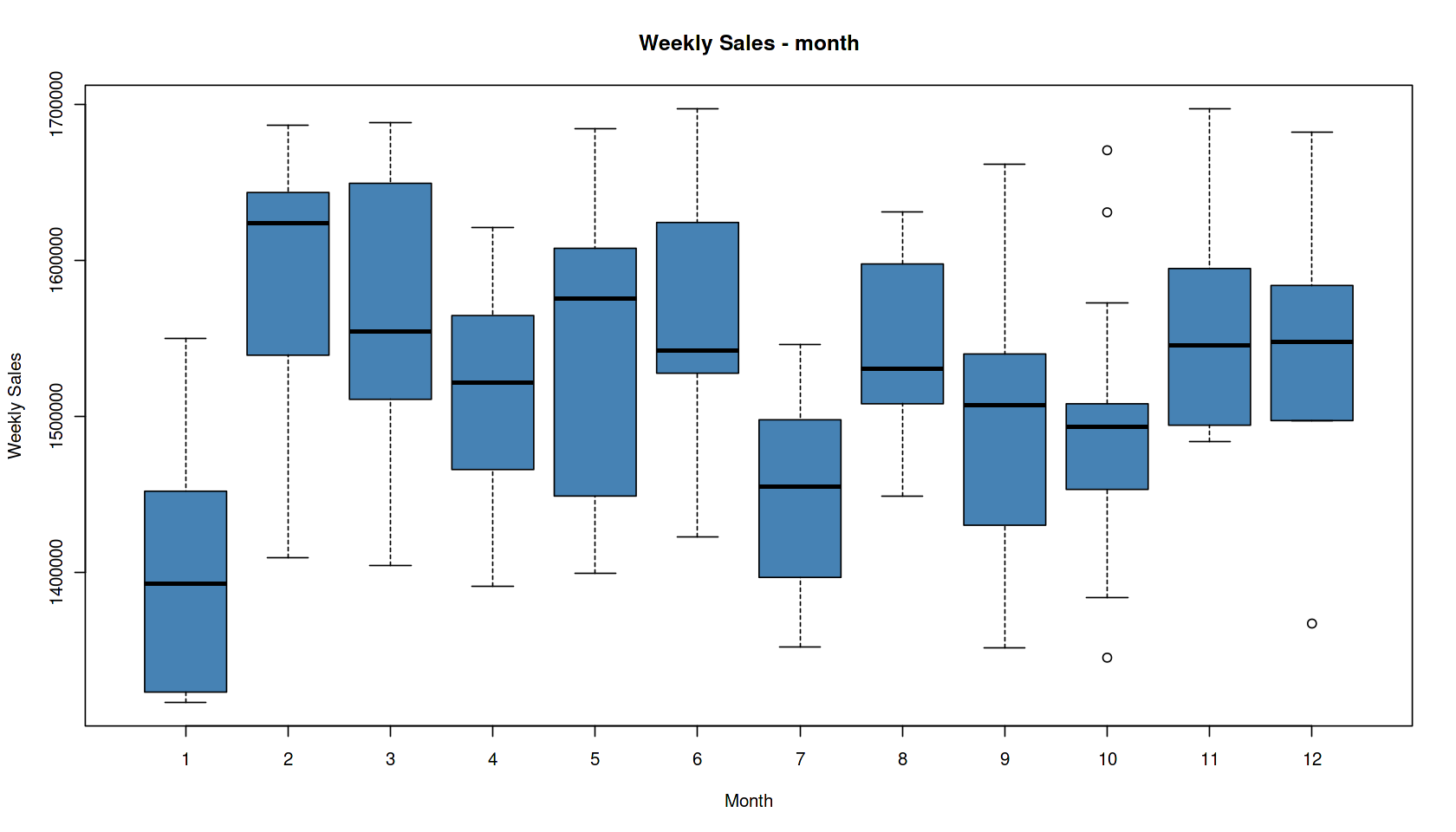














Note: Total 17 observations (11.8%) were removed from data frame

Outliers removed from data in order:

* Temperature 5 outliers
* CPI Outlier treatment 1 outliers
* Unemployment 3 outliers
* Fuel price 2 outliers
* Holiday Flag - No outliers
* Month - 4 outliers
* Quarter - 2 outliers

*#Removing unnecessary columns and changing structure of Events*

data5$Date <-NULL

data5$Store <- NULL

data5$Events <- as.factor(data5$Events)

str(data5)

data5$Holiday\_Flag <- as.numeric(data5$Holiday\_Flag)

data5$Week\_Number <- as.numeric(data5$Week\_Number)

data5$quarter <- as.numeric(data5$quarter)

'data.frame': 126 obs. of 10 variables:

$ Weekly\_Sales: num 1643691 1641957 1611968 1409728 1554807 ...

$ Holiday\_Flag: int 0 1 0 0 0 0 0 0 0 0 ...

$ Temperature : num 42.3 38.5 39.9 46.6 46.5 ...

$ Fuel\_Price : num 2.57 2.55 2.51 2.56 2.62 ...

$ CPI : num 211 211 211 211 211 ...

$ Unemployment: num 8.11 8.11 8.11 8.11 8.11 ...

$ Week\_Number : int 1 2 3 4 5 6 7 8 9 10 ...

$ month : num 2 2 2 2 3 3 3 3 4 4 ...

$ quarter : int 1 1 1 1 1 1 1 1 2 2 ...

$ Events : Factor w/ 4 levels "Christmas","Labour Day",..: 3 4 3 3 3 3 3 3 3 3 ...

**Correlation Plot & Correlation Matrix**

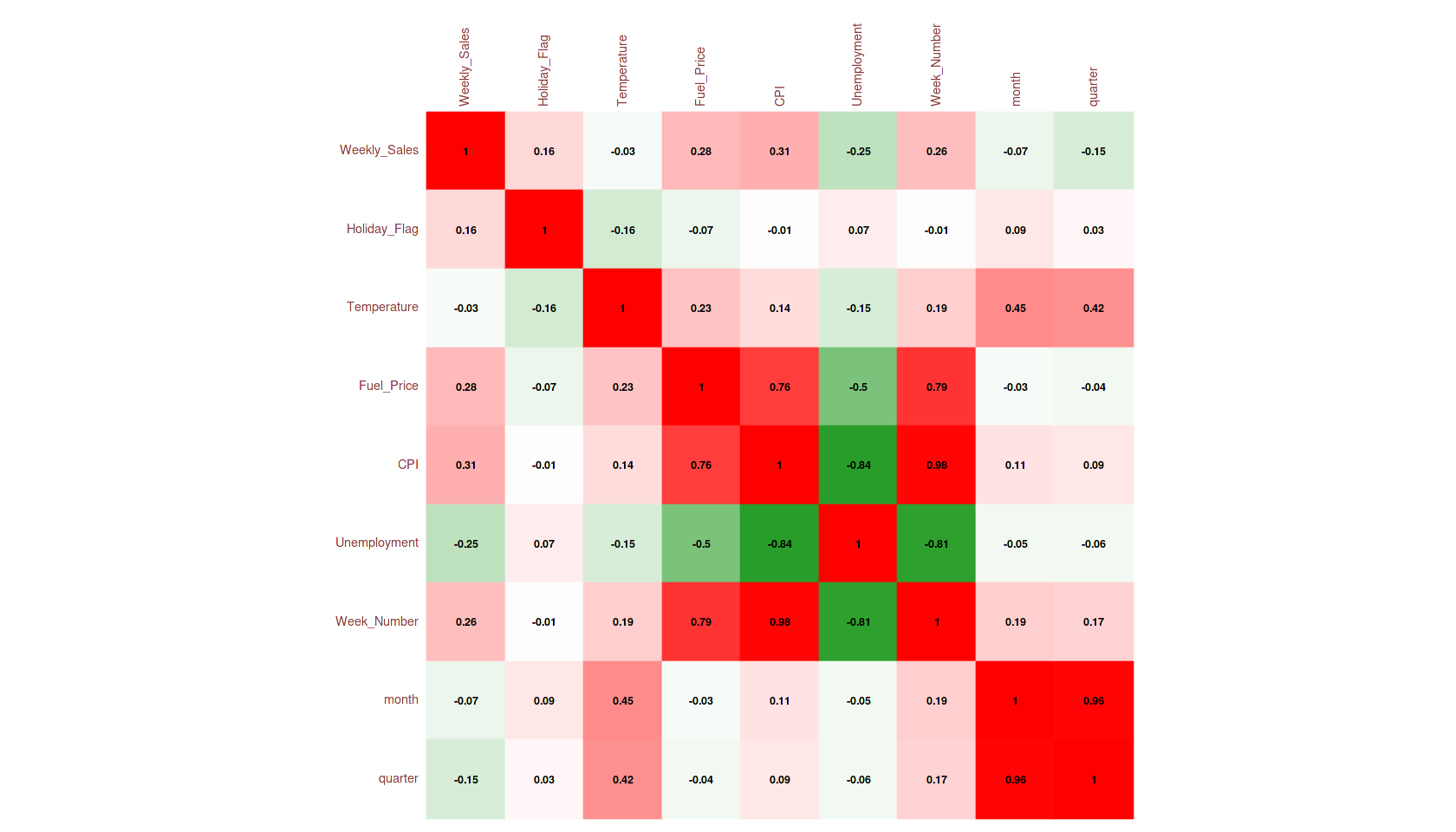
*#correlation matrix and corr plot*

corr = cor(data5[, c(1:9)])

View(corr)

corrplot(corr, method = "color", 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))

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A matrix: 9 × 9 of type dbl | | | | | | | | | |
|  | Weekly\_Sales | Holiday\_Flag | Temperature | Fuel\_Price | CPI | Unemployment | Week\_Number | month | quarter |
| Weekly\_Sales | 1.00000000 | 0.15936349 | -0.02610215 | 0.27923322 | 0.31318157 | -0.25204205 | 0.25822978 | -0.06518967 | -0.15353195 |
| Holiday\_Flag | 0.15936349 | 1.00000000 | -0.16250626 | -0.06936937 | -0.01493103 | 0.07009268 | -0.01255048 | 0.08527631 | 0.03363385 |
| Temperature | -0.02610215 | -0.16250626 | 1.00000000 | 0.22818320 | 0.13691384 | -0.15276761 | 0.18815375 | 0.45477027 | 0.42336802 |
| Fuel\_Price | 0.27923322 | -0.06936937 | 0.22818320 | 1.00000000 | 0.75511662 | -0.50237167 | 0.78665538 | -0.03296547 | -0.04492624 |
| CPI | 0.31318157 | -0.01493103 | 0.13691384 | 0.75511662 | 1.00000000 | -0.83636734 | 0.97533101 | 0.10978554 | 0.08863536 |
| Unemployment | -0.25204205 | 0.07009268 | -0.15276761 | -0.50237167 | -0.83636734 | 1.00000000 | -0.80935295 | -0.04646973 | -0.05850754 |
| Week\_Number | 0.25822978 | -0.01255048 | 0.18815375 | 0.78665538 | 0.97533101 | -0.80935295 | 1.00000000 | 0.18794303 | 0.17189320 |
| month | -0.06518967 | 0.08527631 | 0.45477027 | -0.03296547 | 0.10978554 | -0.04646973 | 0.18794303 | 1.00000000 | 0.96194027 |
| quarter | -0.15353195 | 0.03363385 | 0.42336802 | -0.04492624 | 0.08863536 | -0.05850754 | 0.17189320 | 0.96194027 | 1.00000000 |



**Insights:**

* Observed very low correlation between Temp and Weekly Sales – So can omit Temperature
* Observed low correlation between month, quarter, Holiday Flag with Weekly\_Sales may be due to considering categorical variables as continuous variables

**Creating Dummy Variables**

*#Creating Dummy Variables*

Events <- as.factor(data5$Events)

dummy\_Events <- data.frame(model.matrix(~Events))[,-1]

quarter <- as.factor(data5$quarter)

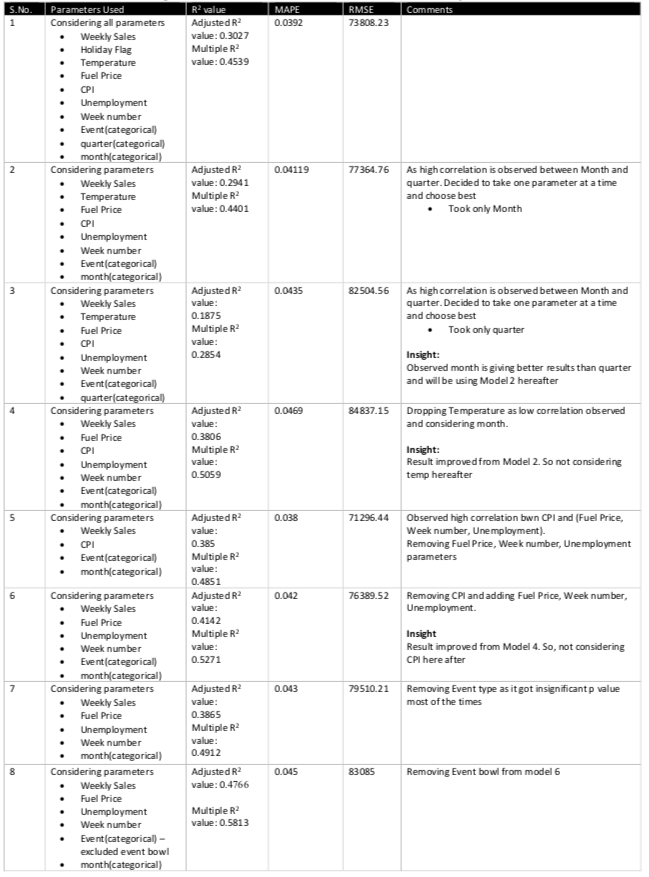
dummy\_quarter <- data.frame(model.matrix(~quarter))[,-1]

month <- as.factor(data5$month)

dummy\_month <- data.frame(model.matrix(~month))[,-1]

data5 <- cbind(data5,dummy\_Events,dummy\_quarter,dummy\_month)

**Summary of All models tried, their outputs and reason used to eliminate or consider a variable**



*############ Model- 8 ####################*

*#Considering parameters - Weekly Sales, Fuel Price, Week number, Unemployment,Event(categorical), month(categorical)*

*#removed Event bowl to Model 6 as it is causing NA's*

*# Splitting dataset into training set and test set*

set.seed(123) *# Seed initializes the randomness -- set.seed helps to fix randomness fixed everytime you open. you can write any number inside the set.seed()*

library(caTools)

*#Considering all parameters - Weekly Sales, Holiday FlagTemp, Fuel, CPI,Unemployment, Weeknumber, Event(categorical), quarter(categorical), month(categorical)*

dataset <- data5[, c(1,4,6,7,11:12, 17:27 )]

*#Creating a sample split and divided test & training sets in 30-70 ratio respectively*

sample = sample.split(dataset, SplitRatio = 0.7) *# Returns a vector with T for 70% of data*

trainingSet = subset(dataset, sample == T)

testSet = subset(dataset, sample == F)

*# Create model*

model = lm(formula = Weekly\_Sales ~ . , data = trainingSet)

summary(model)

Call:

lm(formula = Weekly\_Sales ~ ., data = trainingSet)

Residuals:

Min 1Q Median 3Q Max

-123134 -46522 1841 44510 114359

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1637841.6 345979.4 4.734 1.26e-05 \*\*\*

Fuel\_Price 49812.9 47710.2 1.044 0.300380

Unemployment -50517.0 50231.0 -1.006 0.318350

Week\_Number -146.0 757.6 -0.193 0.847794

EventsLabour.Day 79280.1 86347.1 0.918 0.361984

EventsNo\_Holiday -25662.8 72364.6 -0.355 0.724031

month2 260225.4 42181.7 6.169 5.19e-08 \*\*\*

month3 189405.7 39016.8 4.854 8.11e-06 \*\*\*

month4 137171.4 44248.7 3.100 0.002876 \*\*

month5 173484.8 43940.5 3.948 0.000199 \*\*\*

month6 206825.1 38475.7 5.375 1.15e-06 \*\*\*

month7 95704.1 37717.1 2.537 0.013614 \*

month8 172879.3 37006.9 4.672 1.58e-05 \*\*\*

month9 83160.4 40649.5 2.046 0.044891 \*

month10 102923.0 40265.4 2.556 0.012969 \*

month11 171773.5 46293.8 3.711 0.000435 \*\*\*

month12 115723.3 107500.7 1.076 0.285750

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 66040 on 64 degrees of freedom

Multiple R-squared: 0.5813, Adjusted R-squared: 0.4766

F-statistic: 5.554 on 16 and 64 DF, p-value: 3.221e-07

options(repr.plot.width = 10, repr.plot.height = 10)

*# Visualizing train set results*

y\_pred\_train = predict(model, newdata = trainingSet)

ggplot() +

geom\_point(aes(x=trainingSet$Weekly\_Sales,y=y\_pred\_train), size=3,colour = "Blue") +

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))+

scale\_y\_continuous(labels = label\_number(suffix = " K", scale = 1e-3))+

scale\_x\_continuous(labels = label\_number(suffix = " K", scale = 1e-3))+

ggtitle('Comparision of Actual Sales vs Predicted Sales - Train Data')+

theme(plot.title = element\_text(hjust = 0.5))+

xlab("Actual Sales") + ylab("Predicted Sales")

*# Visualizing the test set results*

y\_pred\_test = predict(model, newdata = testSet)

ggplot() +

geom\_point(aes(x=testSet$Weekly\_Sales,y=y\_pred\_test), size =3, colour = "Blue") +

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))+

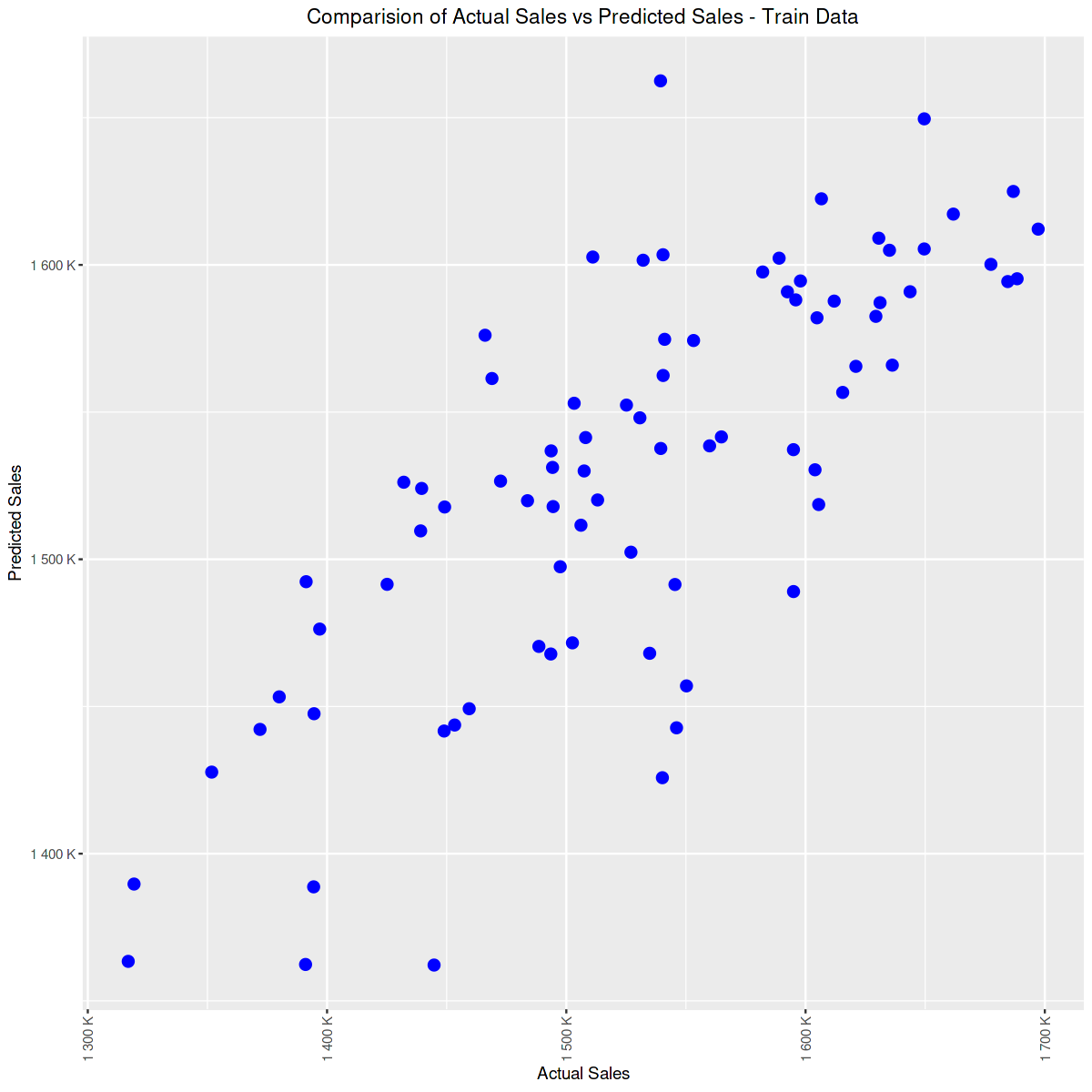
scale\_y\_continuous(labels = label\_number(suffix = " K", scale = 1e-3))+

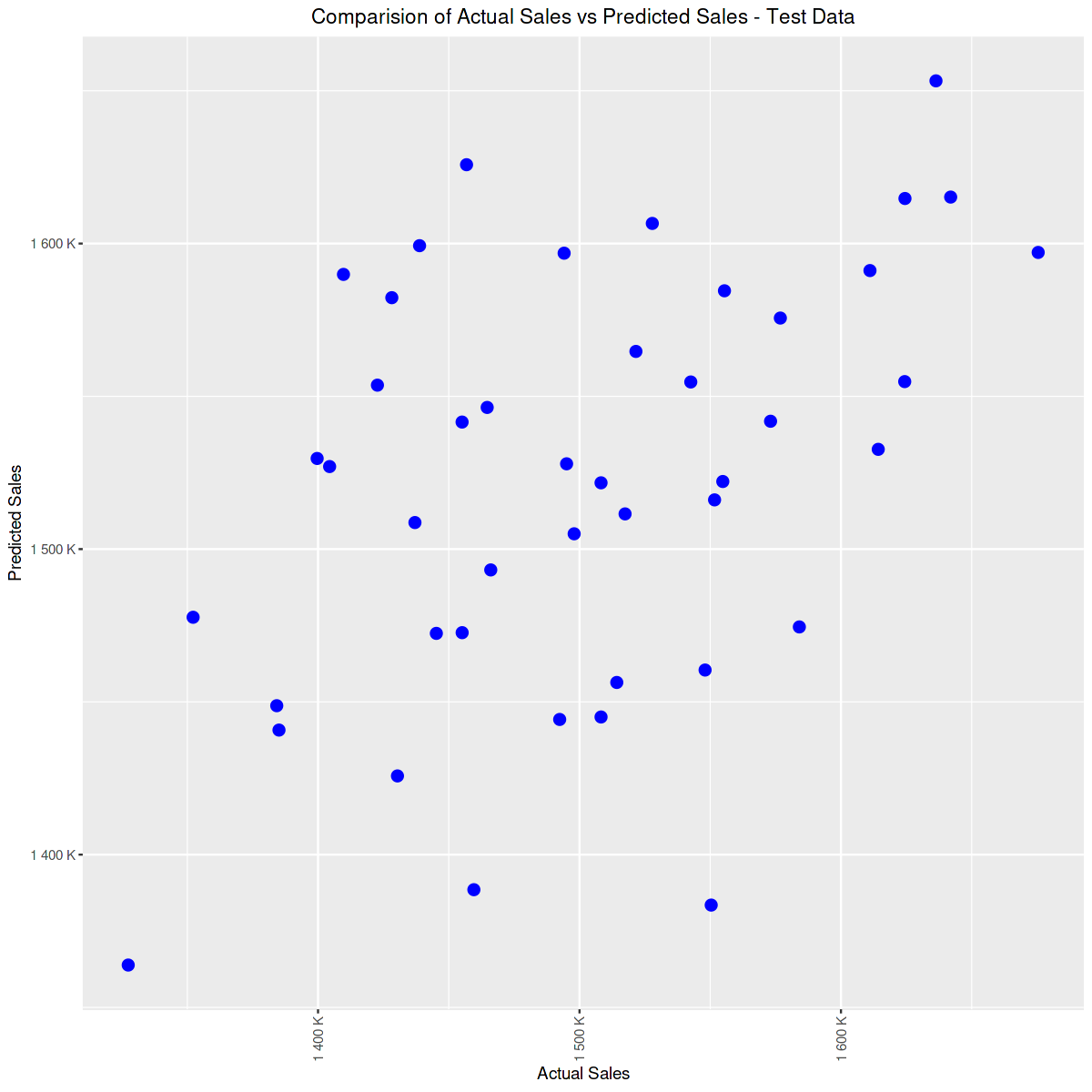
scale\_x\_continuous(labels = label\_number(suffix = " K", scale = 1e-3))+

ggtitle('Comparision of Actual Sales vs Predicted Sales - Test Data')+

theme(plot.title = element\_text(hjust = 0.5))+

xlab("Actual Sales") + ylab("Predicted Sales")





*### Parameters to validate the accuracy of the model and improvise.*

MAPE(y\_pred\_test,testSet$Weekly\_Sales)

RMSE(y\_pred\_test,testSet$Weekly\_Sales)

0.0448667359559854

83085.0503510467

*#checking multi collinearity*

car::vif(model)

**Fuel\_Price**

7.99477876429836

**Unemployment**

6.69257122962899

**Week\_Number**

18.6944083921967

**EventsLabour.Day**

4.93917243442149

**EventsNo\_Holiday**

5.63351325795102

**month2**

2.2667471596023

**month3**

2.79267638909497

**month4**

3.23711337840596

**month5**

3.19216380029003

**month6**

2.44754088980852

**month7**

2.3519733130973

**month8**

2.51237284616309

**month9**

3.03129962259417

**month10**

2.06547412390168

**month11**

1.86870092651086

**month12**

2.61731434210603

**Insights:**

* ‘Model 8’ is the best model among all due to following reasons
  1. Relatively high R square value
  2. VIF values low indicating Less multi Collinearity
  3. Avoiding rank deficient error warnings caused by NA rows
  4. MAPE value which is in acceptable range
* Month can be considered as important independent variable for prediction as its p value are highly significant
* Fuel Price, Unemployment together can be considered as important factors than CPI alone.
* Model can be further improvised by
  1. Considering all Stores data for prediction
  2. Using Advanced models like Decision Trees, Random Forest
  3. Using K cross validation techniques for Sampling data