Business Scenario:

One of the leading retail stores in the US, Walmart, would like to predict the sales 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 at different points of time covering seasonality ingest factors like economic conditions including CPI, Unemployment Index, etc.  
Historical sales data for 45 Walmart stores located in different regions are available.

**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

**Analysis Tasks**

**Basic Statistics tasks**

* Which store has maximum sales
* Which store has a maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of mean to standard deviation
* Which store/s has a good quarterly growth rate in Q3’2012
* Some holidays have a negative impact on sales. Find out holidays which have higher sales than the mean sales in a non-holiday season for all stores together
* Provide a monthly and semester view of sales in units and give insights

**Statistical Model**

* 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.
* Time series forecasting model –
  + Build ARIMA model to forecast 6 months i.e., input utilize only till April 2012.

Predict next 6 months i.e., June to Oct 2012. Check for MAPE.

CODES:

To call file in R environment

Walmart <- read.csv(file.choose())

#which store has Max sales

Store\_total\_sales <- aggregate(Walmart$Weekly\_Sales,by=list(Walmart$Store),sum)

names(Store\_total\_sales)[names(Store\_total\_sales)=="Group.1"] <- "Store\_Number"

names(Store\_total\_sales)[2] <- "Total\_sales"

max(Store\_total\_sales)

max\_sales\_index <- which.max(Store\_total\_sales$Total\_sales)

max\_sale\_store\_id <- Store\_total\_sales[max\_sales\_index,"Store\_Number"]

write (paste("store id number", max\_sale\_store\_id, "has maximum sales"),file= "a.txt")

print (paste("store id number", max\_sale\_store\_id, "has maximum sales"))

##########################################################################

#Which store has maximum statdered deviaion in sales

Sales\_SD <- aggregate(Walmart$Weekly\_Sales,by=list(Walmart$Store),sd)

max(Sales\_SD)

names(Sales\_SD)[1] <-"Store\_id"

names(Sales\_SD)[2] <- "SD"

max\_Sales\_SD\_Index <- which.max(Sales\_SD$SD)

max\_sales\_SD\_Store\_id <- Sales\_SD[max\_Sales\_SD\_Index,"Store\_id"]

print(paste("store id number", max\_sales\_SD\_Store\_id,"has maximum Staddard deviation"))

#####################################################################################

# Coefficient of mean to standard Deviation

CV <-function(x){sd(x)/mean(x)\*100}

Sales\_cv <- aggregate(Walmart$Weekly\_Sales,by=list(Walmart$Store),FUN=CV)

names(Sales\_cv)[1] <- "Store\_id"

names(Sales\_cv)[2] <- "Coefficient of Variation"

##################################################################################

library(lubridate)

#Quartly Growth rate

Date\_Quarters <- quarter(as.Date(Walmart$Date, "%d-%m-%Y"),with\_year = T)

Quartly\_Sales <- aggregate(Walmart$Weekly\_Sales,by=list(Walmart$Store,Date\_Quarters),sum)

names(Quartly\_Sales)[1] <- "Store\_id"

names(Quartly\_Sales)[2] <- "Yearwise\_Quarters"

names(Quartly\_Sales)[3] <- "Total\_Sales"

View(Quartly\_Sales)

Q3\_2012\_data <- Quartly\_Sales[Quartly\_Sales$Yearwise\_Quarters == 2012.3,]

names(Q3\_2012\_data)[2] <-"2012\_q3"

names(Q3\_2012\_data)[3] <-"2012\_q3\_sales"

Q2\_2012\_data <- Quartly\_Sales[Quartly\_Sales$Yearwise\_Quarters == 2012.2,]

names(Q2\_2012\_data)[2] <-"2012\_q2"

names(Q2\_2012\_data)[3] <-"2012\_q2\_sales"

merged\_Q3\_2012\_Q2\_2010 <- merge(Q3\_2012\_data,Q2\_2012\_data,by=intersect(names(Q2\_2012\_data), names(Q3\_2012\_data)))

merged\_Q3\_2012\_Q2\_2010["GR"] <- (merged\_Q3\_2012\_Q2\_2010$`2012\_q3\_sales` - merged\_Q3\_2012\_Q2\_2010$`2012\_q2\_sales`)/merged\_Q3\_2012\_Q2\_2010$`2012\_q2\_sales`\*100

store\_id\_with\_max\_gr <- merged\_Q3\_2012\_Q2\_2010 [which.max(merged\_Q3\_2012\_Q2\_2010$GR), "Store\_id"]

##################################################################################

# Semesterwise total sales

Date\_Semester <- semester(as.Date(Walmart$Date, "%d-%m-%Y"), with\_year = F)

Semester\_Sales <- aggregate(Walmart$Weekly\_Sales,by=list(Walmart$Store,Date\_Semester),sum)

names(Semester\_Sales)[1] <- "Store\_id"

names(Semester\_Sales)[2] <- "Semester"

names(Semester\_Sales)[3] <- "Total\_Sales"

#1st Semester Sales

first\_semester\_Sales <- Semester\_Sales[Semester\_Sales$Semester == "1",]

summary(first\_semester\_Sales$Total\_Sales)

plot(first\_semester\_Sales$Store\_id,first\_semester\_Sales$Total\_Sales,type ="b",

main = "First\_Semester\_Total\_Sale",col=("Red"))

#2nd Semester Sales

Second\_semester\_Sales <- Semester\_Sales[Semester\_Sales$Semester == "2",]

summary(Second\_semester\_Sales$Total\_Sales)

plot(Second\_semester\_Sales$Store\_id,Second\_semester\_Sales$Total\_Sales,type ="b",

main = "Second\_Semester\_Total\_Sale",col=("Royalblue"))

## Monthly and Yearly view of sales in units and give insights

Walmart$Date =as.Date(Walmart$Date,format=c("%d-%m-%Y"))

Walmart\_month\_year <- transform(Walmart,Year\_Sale =as.numeric(format(Date,"%Y"))

,Month\_Sale =as.numeric(format(Date,"%m")))

Monthly\_Yearly\_Sales <- aggregate(Walmart\_month\_year$Weekly\_Sales,

by=list(Walmart\_month\_year$Month\_Sale,Walmart\_month\_year$Year\_Sale),sum)

names(Monthly\_Yearly\_Sales)[1] <- "Month"

names(Monthly\_Yearly\_Sales)[2] <- "Year"

names(Monthly\_Yearly\_Sales)[3] <- "Total\_Sales"

Summarized\_View <- arrange(Monthly\_Yearly\_Sales,desc(Total\_Sales))

View(Summarized\_View)

summary(Monthly\_Yearly\_Sales$Total\_Sales)

############################################################################

##Holiday Sales & Non Holiday Sales

Holiday\_Sales <- aggregate(Walmart$Weekly\_Sales,by=list(Walmart$Store,Walmart$Holiday\_Flag),mean)

View(Holiday\_Sales)

names(Holiday\_Sales)[1] <- "Store\_id"

names(Holiday\_Sales)[2] <- "Holiday\_Week"

names(Holiday\_Sales)[3] <- "Average\_Sales"

Non\_Holiday\_Sales <- Holiday\_Sales[Holiday\_Sales$Holiday\_Week==0,]

names(Non\_Holiday\_Sales)[2] <- "Non\_holiday\_Week"

Non\_holiday\_Avg\_Sales <- mean(Non\_Holiday\_Sales$Average\_Sales)

Walmart\_Holiday <- Walmart[Walmart$Holiday\_Flag==1,]

View(Walmart\_Holiday)

##when we take average of storewiseaverage sale for holiday week

Holiday\_Week\_Avg\_Sales <- aggregate(Walmart\_Holiday$Weekly\_Sales,by=list(Walmart\_Holiday$Date),mean)

names(Holiday\_Week\_Avg\_Sales)[1] <- "Week"

names(Holiday\_Week\_Avg\_Sales)[2] <- "Avg\_Sales"

Holiday\_Higher\_Sales <- Holiday\_Week\_Avg\_Sales[Holiday\_Week\_Avg\_Sales$Avg\_Sales > Non\_holiday\_Avg\_Sales,]

View(Holiday\_Higher\_Sales)

## When we take Store wise Total sales for holiday weeks

Holiday\_Week\_Avg\_Sales <- aggregate(Walmart\_Holiday$Weekly\_Sales,by=list(Walmart\_Holiday$Date),sum)

names(Holiday\_Week\_Avg\_Sales)[1] <- "Week"

names(Holiday\_Week\_Avg\_Sales)[2] <- "Total\_Sales"

Holiday\_Higher\_Sales <- Holiday\_Week\_Avg\_Sales[Holiday\_Week\_Avg\_Sales$Total\_Sales > Non\_holiday\_Avg\_Sales,]

View(Holiday\_Higher\_Sales)

################################################################################

###Regression Analysis

Linear\_regression\_Walmart <- lm(Weekly\_Sales ~ Holiday\_Flag + Temperature + Fuel\_Price+ CPI + Unemployment , Walmart)

summary(Linear\_regression\_Walmart)

### droppping insignificant vars i.e temp and fuel price

Linear\_regression\_Walmart <- lm(Weekly\_Sales ~ Holiday\_Flag + CPI + Unemployment , Walmart)

summary(Linear\_regression\_Walmart)

###Time series model

# visually identifying if data is fit for time series

Time\_Series\_Model = aggregate(Weekly\_Sales~Date,Walmart,sum)

View(Time\_Series\_Model)

plot(Time\_Series\_Model,type='l',col=("red"))

options("scipen"=100,"digit"=4)

#################################################################################

###ARIMA Model

View(Monthly\_Yearly\_Sales)

Walmart\_month\_year\_filtered <- select(Monthly\_Yearly\_Sales,Total\_Sales,Year,Month)

View(Walmartretail\_month\_year\_filtered)

# rolling up sales at month level

Walmart\_Month <- aggregate(Walmart\_month\_year\_filtered$Total\_Sales,

by=list(Walmart\_month\_year\_filtered$Year,Walmart\_month\_year\_filtered$Month),sum)

names(Walmart\_Month)[1] <- "Year"

names(Walmart\_Month)[2] <- "Month"

names(Walmart\_Month)[3] <- "Total\_Sales"

# sorting in year and month order

Walmart\_sorted <- arrange(Walmart\_Month,Year,Month)

View(Walmart\_sorted)

# creating a Column with month and year of sale

Walmart\_TS <- transform(Walmart\_sorted,Time\_Of\_Sale = as.Date(paste(Year,"-",Month,"-",1,sep=""),

format="%Y-%m-%d"))[,c(4,3)]

#### Build up ARIMA model to forecast last 6 months i.e as in input utilize only till April 2012

# Predict next 6 months i.e June to Oct 2012. Check for MAPE

# Building ARIMA model

library(forecast)

Walmart\_ARIMA <- arima(Walmart\_TS[1:27,2],order=c(2,1,2))

Forecasted\_Sale <- forecast(Walmart\_ARIMA,h=6)

Forecasted\_Sale

plot(Forecasted\_Sale)

# 6 months forecast

Forecasted\_Sales <- as.data.frame(Forecasted\_Sale)

Forecasted\_Sales\_6m = Forecasted\_Sales[,1]

View(Forecasted\_Sales\_6m)

View(Walmart\_TS)

# 6 m actual

Actual\_Sales <-Walmart\_TS[28:33,]

# concatenating 6 m forecast and actual

Actual\_Forecst\_last\_6\_m <- cbind(Forecasted\_Sales,Actual\_Sales)

View(Actual\_vs\_Forecst\_last\_6\_m)

Actual\_Forecst\_last\_6\_m\_deviation <- transform(Actual\_Forecst\_last\_6\_m,

Errors = abs(Forecasted\_Sales\_6m-Total\_Sales)/Total\_Sales)

View(Actual\_vs\_Forecst\_last\_6\_m\_deviation)

### Accuracy Test

MAPE = mean(Actual\_vs\_Forecst\_last\_6\_m\_deviation$Errors)

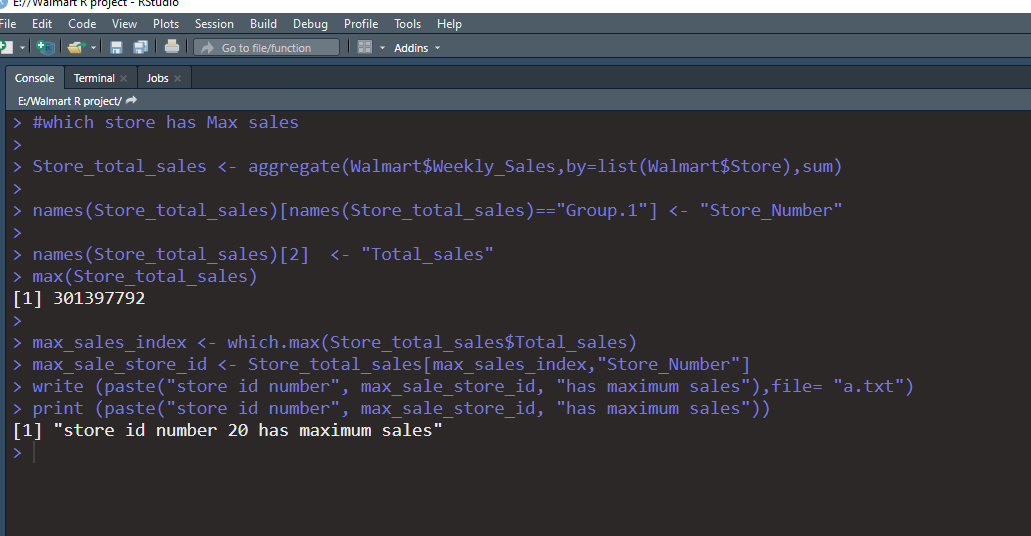
MAPE

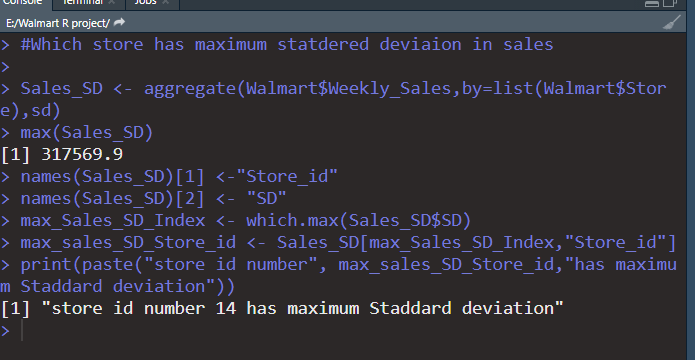
accuracy <- 1-MAPE

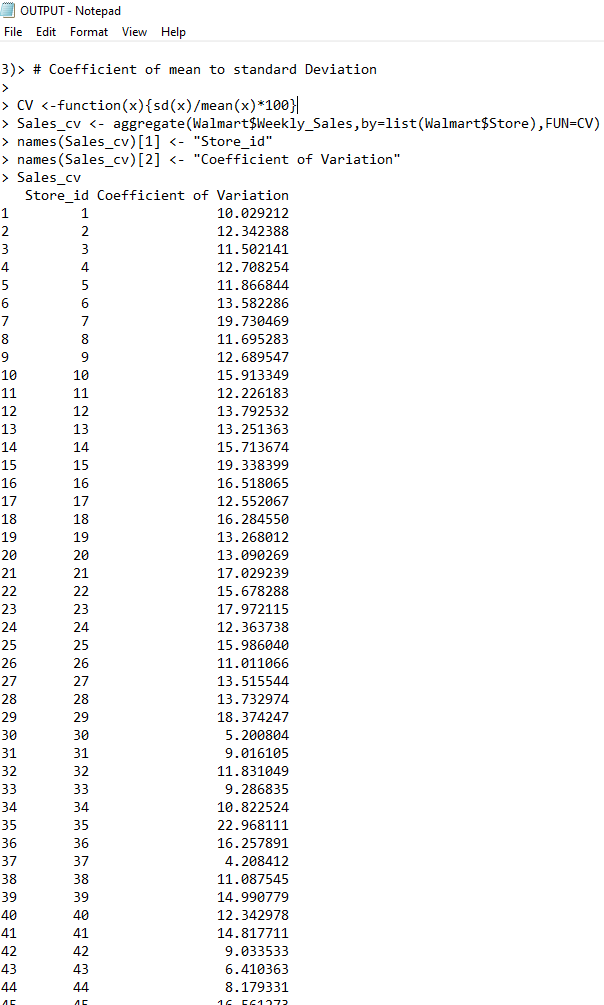
accuracy

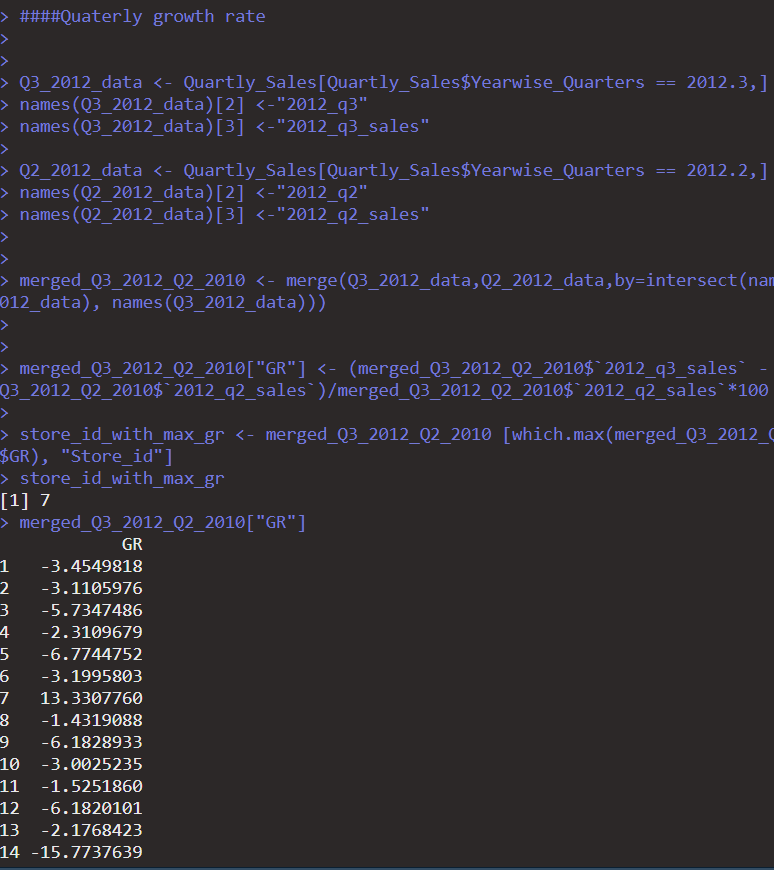
###The accuracy level of forecast of sales is 88%

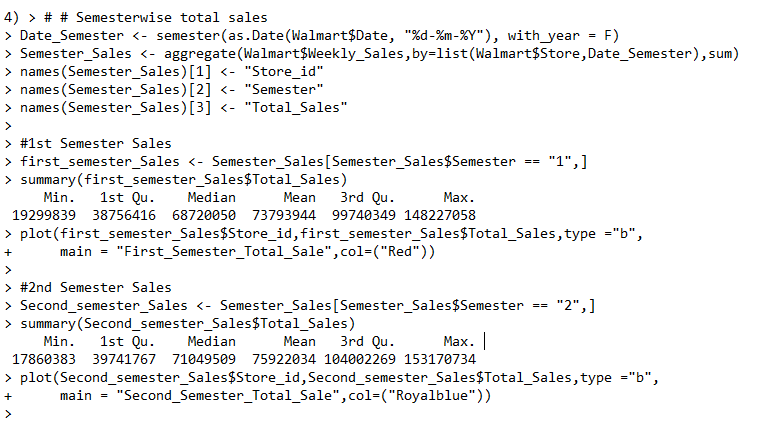
OUTPUT SCREENSHOT:

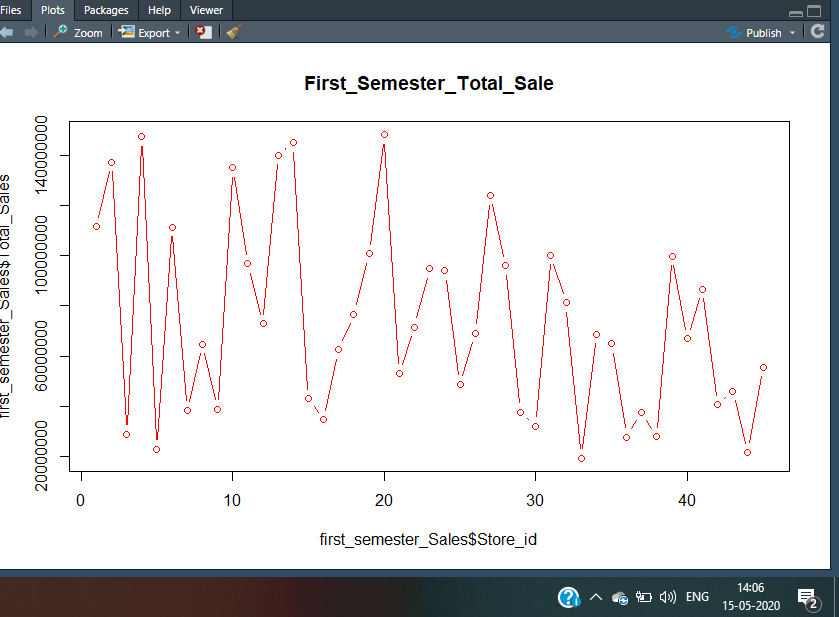


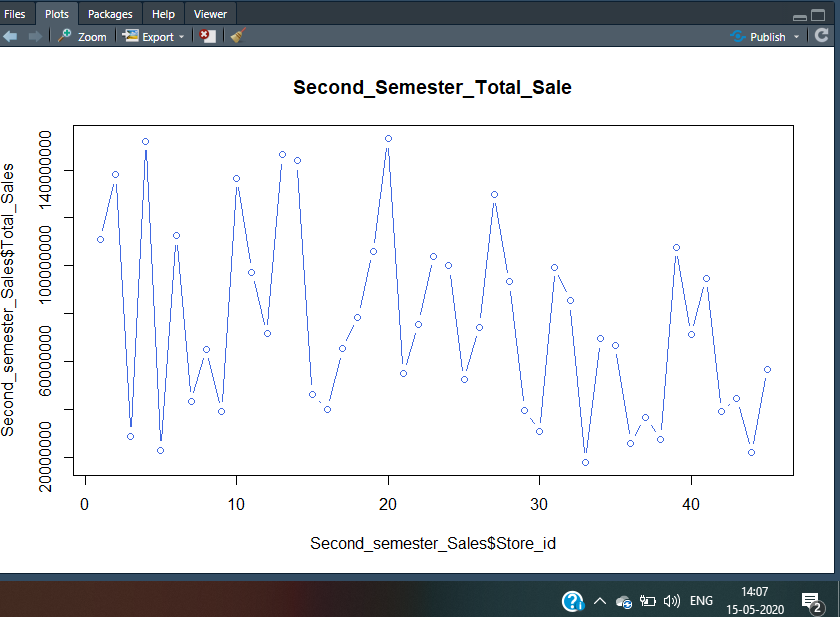


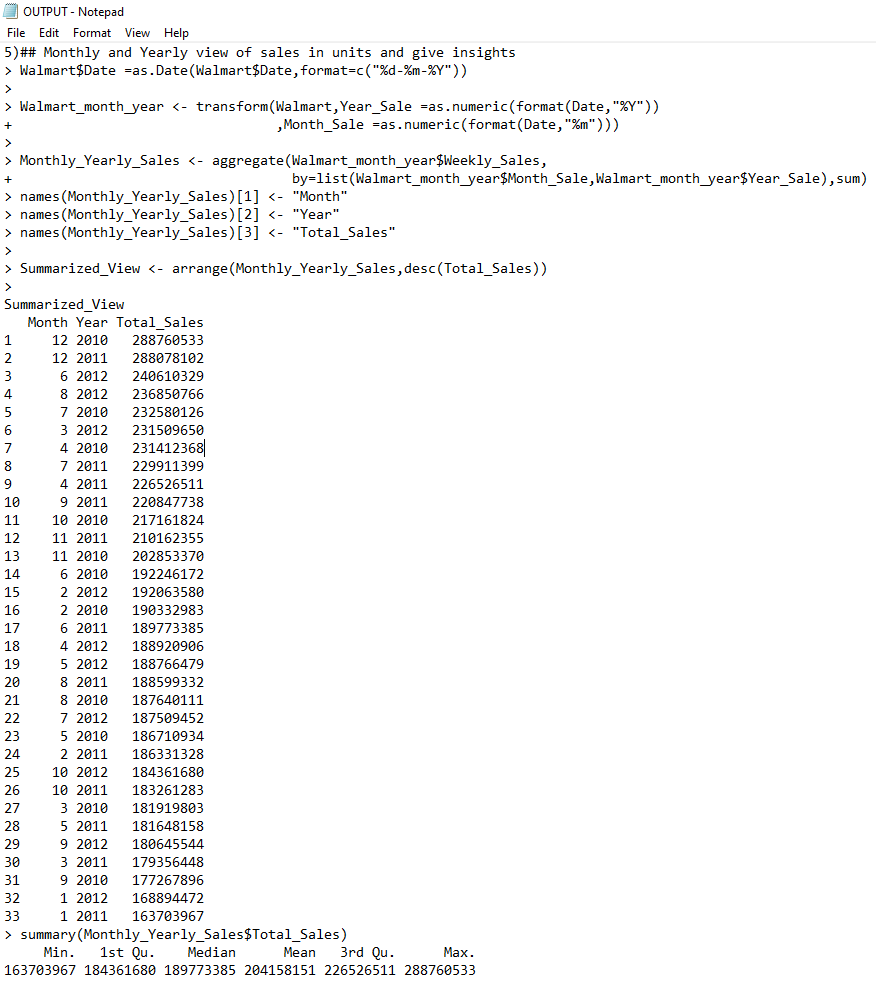


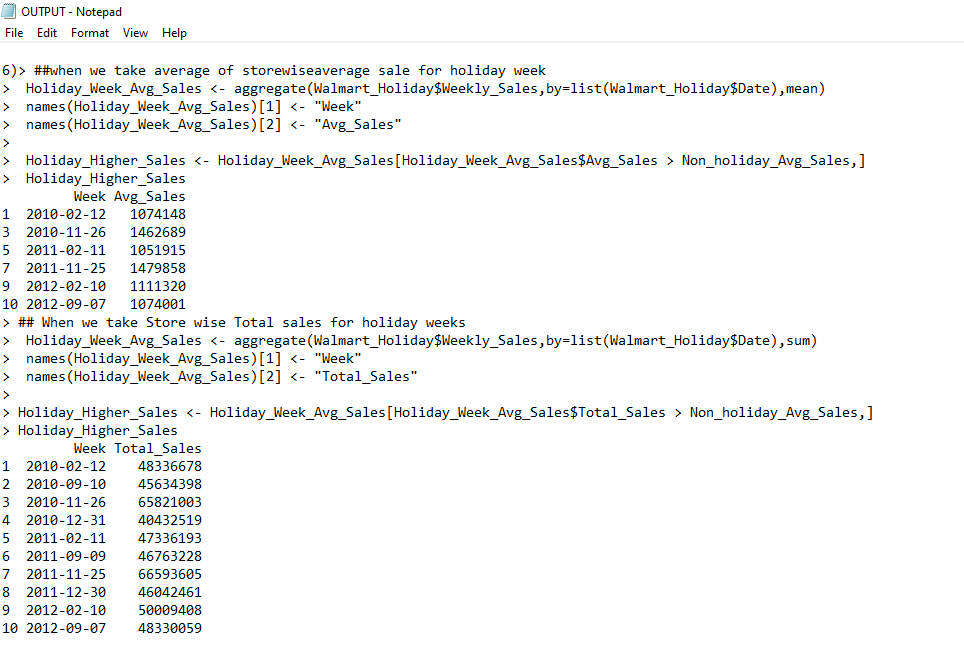


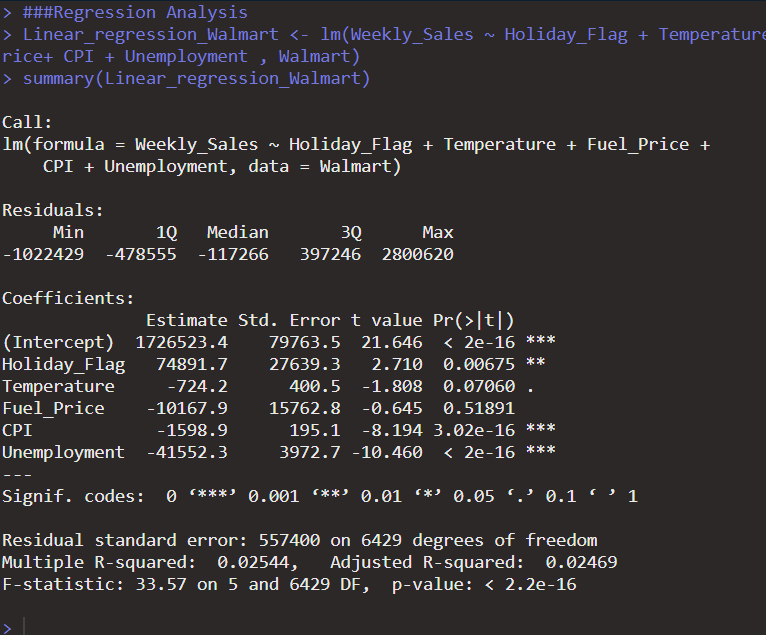


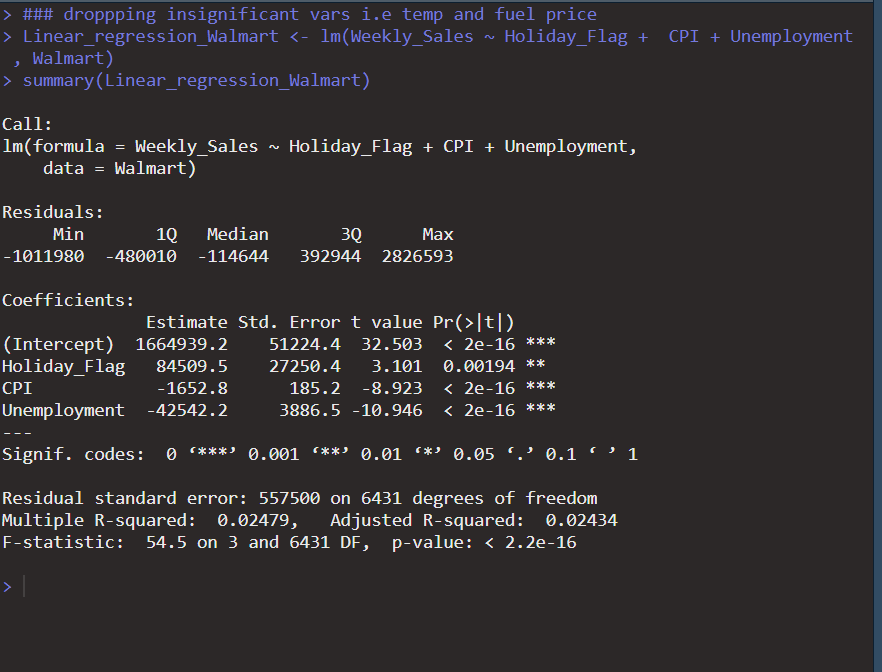


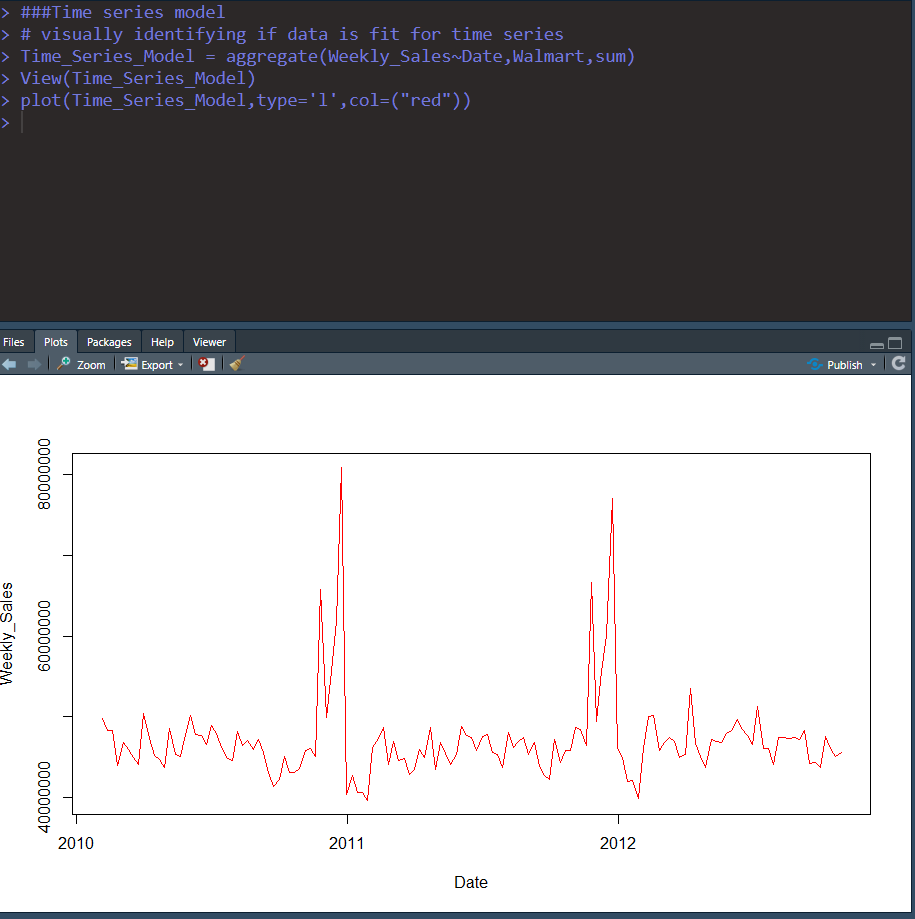


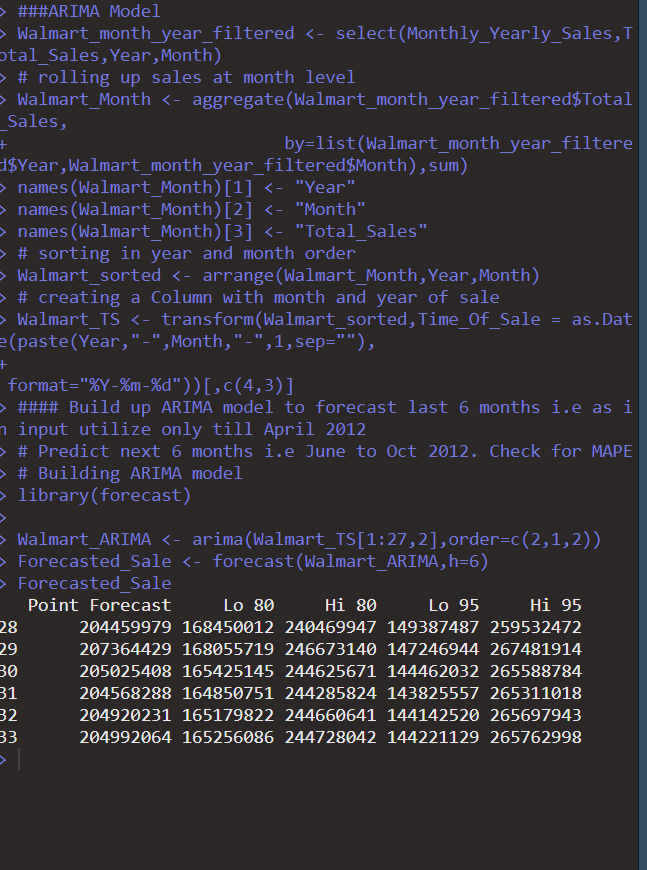


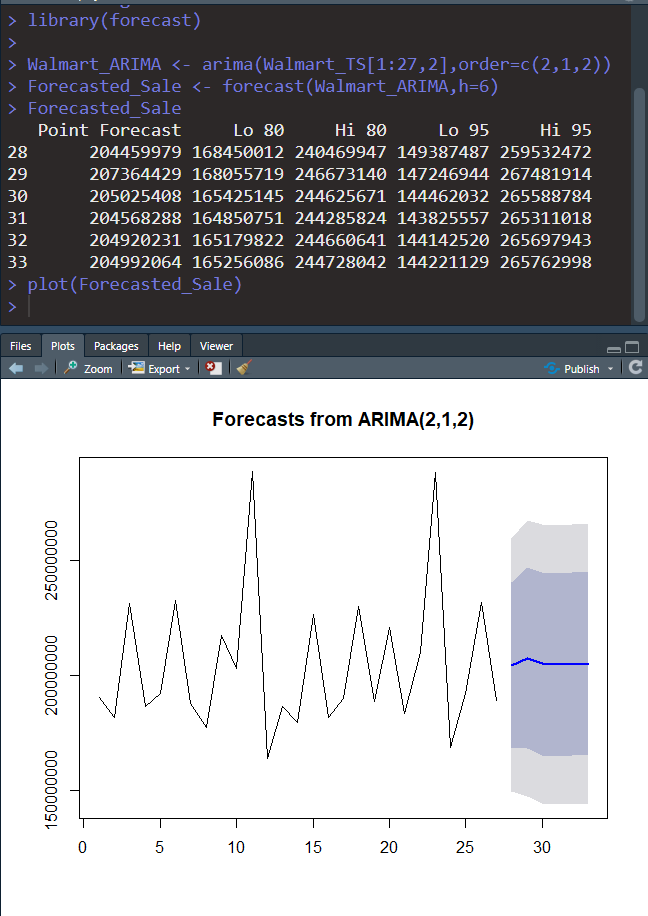


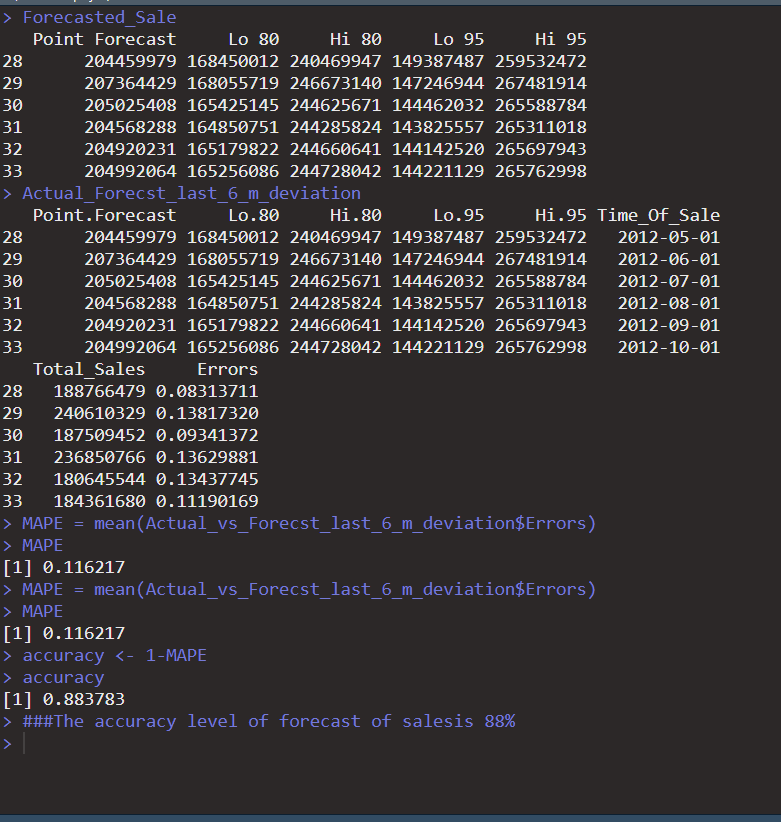












ANALYSIS:

* Maximum Total Weekly Sales is done by store 22 which is 301397792
* Store 14 is having maximus variation in weekly sale which is 317569.9
* Many stores are quarterly negative growth rate in Q32012 as compare to Q2 2012. Store 7 is highest growth rate in this quarter which is 13.3307760%.
* By analysing Monthly sales data, it shows December month is having maximum sale in every year. So Walmart should follow the same strategy which is working for December month.
* Second semester is having more sale than First Semester sale.
* When we compare sum of holiday week sales with mean of non-holiday week sales than all the 10 holidays having higher sales while when we compare weekly average sales of holiday weeks with mean sale of non-holiday weeks, following are the holiday weeks having higher sales:

Week Avg\_Sales

* 2010-02-12 1074148
* 2010-11-26 1462689
* 2011-02-11 1051915
* 2011-11-25 1479858
* 2012-02-10 1111320
* 2012-09-07 1074001
* First Linear regression model is showing Fuel\_prices and Temprature as an insignificant variable. So final regression model eliminates these two variables still it has very low R2 value as well as adjusted R2 is also very low. That shows Liner regression model is not fit to be good with given independent variables.
* Time series analysis is done for the given data set it is not stationary data. So to covert it into stationary ARIMA model is used with (p) =2 and (q)= 2 with d=1, which forecast the sale and MAPE value is 11.62% which shows 88.37% accuracy of forecast model.