**Project:** Retail Analysis with Walmart

**Submitted By:** Kumaran Kanniappan

**Dataset Description**

**Dataset:** Walmart Store Sales

**Description**: This is the historical data which covers sales from 2010-02-05 to 2012-11-01, in the file Walmart\_Store\_sales.

**Observations:** 6435

**Variables:** 8

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***

**dim – dimensions**

A screenshot of a social media post

Description automatically generated

Observations: 6435

Variables: 8

**str** – structure

A screenshot of a social media post

Description automatically generated

**summary** – summary of the data

A screenshot of a computer

Description automatically generated

* ***Which store has maximum sales***

Store 20 has maximum sales 3013397792

A screenshot of a cell phone

Description automatically generated

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

Store 14 has maximum standard deviation 317570

A close up of a logo

Description automatically generated

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

Store 7 has got the good quarterly growth rate (qoq) for Q3 2012 of 13.3 followed by 16,35, 26

A screenshot of a cell phone

Description automatically generated

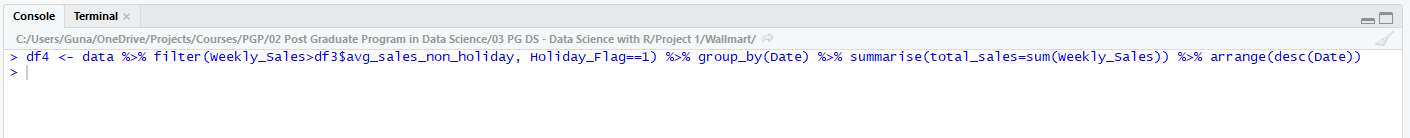
* Some holidays have a negative impact on sales. ***Find out holidays which have higher sales than the mean sales in non-holiday season for all stores together***

**Mean Sales in Non-Holiday Season** is 1041256

A screenshot of a cell phone

Description automatically generated

**Holidays which have higher sales than the mean sales in non-holiday season** – are as given below:



A screenshot of a cell phone

Description automatically generated

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

***A screenshot of a cell phone

Description automatically generated***

**Monthly Sales**

A picture containing fence

Description automatically generated

**Semester Sales**

A screenshot of a cell phone

Description automatically generated

**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.

**p-value is > 0.05 for Unemployment, Fuel\_Price.**

**There is no correlation to the Weekly\_Sales due to Unemployment, Fuel\_Price**

A screenshot of a social media post

Description automatically generated

* Select the model which gives best accuracy.

**Following model with Holiday\_Flag, Temperature, CPI, new\_day is selected based on their p\_value which can provide best accuracy**

A screenshot of a social media post

Description automatically generated

* Change dates into days by creating new variable.

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**R Script**

# import

data <- read.csv('Walmart\_Store\_sales.csv',stringsAsFactors = FALSE)

View(data)

# dimensions

dim(data)

# structure

str(data)

# summary

summary(data)

# libraries

library(lubridate)

library(dplyr)

library(ggplot2)

# store wise sum of sales in descending order of the sales

data %>%group\_by(Store) %>%  summarise(total\_sales=sum(Weekly\_Sales)) %>% arrange(desc(total\_sales))

# store wise standard deviation in descending order of the sales

data %>%group\_by(Store) %>%  summarise(sd\_sales=sd(Weekly\_Sales)) %>% arrange(desc(sd\_sales))

# store wise sum of sales and average of salaes

df1 <- data %>%group\_by(Store) %>%  summarise(sd\_sales=sd(Weekly\_Sales),avg\_sales=mean(Weekly\_Sales))

df1 %>% mutate(coef\_var=avg\_sales/sd\_sales)

# changing date format

data <- data %>% mutate(Date=dmy(Date))

# adding the quarter and year

data <- data %>% mutate(qtr=quarter(Date))

data <- data %>% mutate(yr=year(Date))

# concatenate the yr and qtr together

data <- data %>% mutate(quart=paste(yr,'Q',qtr,sep=''))

# store wise and quarter wise sum of sales

df2 <- data %>% group\_by(Store,quart) %>% summarise(total\_sales=sum(Weekly\_Sales))

# adding lag\_sales1, lag\_sales4

df2 <- df2 %>% mutate(lag\_sales1=lag(total\_sales,1),lag\_sales4=lag(total\_sales,4))

# find the growth rate and filter for Q3 2012

df2 %>%filter(quart=='2012Q3') %>%  mutate(qoq=((total\_sales/lag\_sales1)-1)\*100,yoy=((total\_sales/lag\_sales4)-1)\*100) %>% arrange(desc(qoq))

# find mean sales in non-holiday season for all stores together, find holidays with more sales than avg\_sales\_non\_holiday = 1041256

df3 <- data %>% filter(Holiday\_Flag==0) %>% summarise(avg\_sales\_non\_holiday=mean(Weekly\_Sales))

df4 <- data %>% filter(Weekly\_Sales>df3$avg\_sales\_non\_holiday, Holiday\_Flag==1) %>% group\_by(Date) %>%

  summarise(total\_sales=sum(Weekly\_Sales)) %>%

  arrange(desc(Date))

# add month, semester

data <- data %>% mutate(mon=month(Date),sem=semester(Date))

data <- data %>% mutate(mon\_yr=paste(yr,'M',mon,sep=''),sem\_yr=paste(yr,'S',sem,sep=''))

# monthly, semester summary and plot

mon\_sales <- data %>% group\_by(mon\_yr) %>% summarise(mon\_sales=sum(Weekly\_Sales))

sem\_sales <- data %>% group\_by(sem\_yr) %>% summarise(sem\_sales=sum(Weekly\_Sales))

View(mon\_sales)

View(sem\_sales)

ggplot(data=mon\_sales,mapping=aes(x=mon\_yr,y=mon\_sales))+geom\_bar(stat='identity')

ggplot(data=sem\_sales,mapping=aes(x=sem\_yr,y=sem\_sales))+geom\_bar(stat='identity')

# restructure dates, add new\_day, new\_date; change Store as factor, order as numeric; date order as a serial number from 1

data <- data %>% mutate(new\_day=day(Date))

df5 <- data.frame(new\_date=levels(as.factor(data$Date)))

df5$order <- rownames(df5)

df5$Date <- ymd(df5$new\_date)

View(df5)

str(df5)

data\_final <- inner\_join(data,df5,by='Date')

View(data\_final)

str(data\_final)

data\_final <- data\_final %>% mutate(Store=as.factor(Store),order=as.numeric(order))

str(data\_final)

# Prediction model to forecast demand for Store1 using linear regression model

store1\_data <- data\_final %>% filter(Store==1)

View(store1\_data)

linear\_model <- lm(Weekly\_Sales ~ Holiday\_Flag+Temperature+CPI+Unemployment+Fuel\_Price+new\_day,data=store1\_data)

summary(linear\_model)

linear\_model <- lm(Weekly\_Sales ~ Holiday\_Flag+Temperature+CPI+new\_day,data=store1\_data)

summary(linear\_model)