##### #################################### Activate libraries #####################################

# install packages, if needed, and then load the packages.

install.packages("pacman")

library(pacman) # No message.

pacman::p\_load(pacman, dplyr, ggplot2, lubridate, raster, psych, usdm, car, coefplot)

# install.packages(c("raster"))

# install.packages("lubridate")

# install.packages("dplyr")

# install.packages("psych", dependencies = TRUE)

library(psych)

library("dplyr")

library(sp)

library(raster)

library(usdm)

library("lubridate")

library(plotrix)

library(ggplot2)

library(coefplot)

detach("package:usdm", unload=TRUE)

library(car)

#Import the dataset

wss <- read.csv("~/R\_files/Project/Walmart/Walmart\_Store\_sales.csv")

View(wss)

wss$Store=as.factor(wss$Store)

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

str(wss)

# rm(wss)

########## Question 1 ##########

#find maximum Weekly Sales based on the Store category

top\_sales <- aggregate(Weekly\_Sales~Store,data=wss,max)

top\_sales

#find the maximum Weekly

top\_sales %>%

filter(Weekly\_Sales == max(Weekly\_Sales, na.rm = TRUE))

############## END ##############

########## Question 2 ##########

#find standard deviation of Weekly sales based on the Store category & put in object std\_dev

std\_dev <- aggregate(Weekly\_Sales~Store,data=wss,sd)

std\_dev

#rename column name

colnames(std\_dev)<-c("Store","Stdev")

std\_dev

#find the maximum standard deviation of weekly sales and pass to variable max\_std

max\_std <- std\_dev %>%

filter(Stdev == max(Stdev, na.rm = TRUE))

max\_std

#output the mean of std\_dev object variable Stdev

mean(std\_dev$Stdev)

#output summary std\_dev object

summary(std\_dev)

#find coefficient of standard variation

cosd <- max\_std$Stdev/mean(std\_dev$Stdev)

#output coefficient of standard variation

cosd

############## END ##############

########## Question 3 ##########

# Install & load lubridate

# Extract year and Quarter into a columns

wss$Year <- year(wss$Date)

wss$Quarter <- quarter(wss$Date)

# Install & load dplyr

wss %>% slice(1:220)

wss %>%

dplyr::select(Year, Quarter, Weekly\_Sales)

# df <- wss %>% filter(Year==2012, Quarter>=2 & Quarter<4)

# df

df2 <- wss %>% filter(Year==2012, Quarter==2)

df2

#find sum of Weekly Sales based on the Store category for data subset df2

top\_sales\_df2 <- aggregate(Weekly\_Sales~Store,data=df2,sum)

#rename column name

colnames(top\_sales\_df2)<-c("Store","Q2\_Sales\_Total")

top\_sales\_df2

df3 <- wss %>% filter(Year==2012, Quarter==3)

df3

#find sum of Weekly Sales based on the Store category

top\_sales\_df3 <- aggregate(Weekly\_Sales~Store,data=df3,sum)

#rename column name

colnames(top\_sales\_df3)<-c("Store","Q3\_Sales\_Total")

top\_sales\_df3

#Coloumn Bind df2 and df3

#df <- cbind(top\_sales\_df2,top\_sales\_df3)

#df

#Merge df2 and df3 by Store

df4 <- merge(top\_sales\_df2,top\_sales\_df3, by="Store", all=T)

df4

#Extract growth into a column

df4$Growth <- (df4$Q3\_Sales\_Total - df4$Q2\_Sales\_Total)/ df4$Q2\_Sales\_Total

df4

#Indentify positive growth

df5 <- df4 %>%

filter(Growth>0)

df5

#Plot the df5

barplot(df5$Growth, ylim =c(0,0.2), names = df5$Store, xlab = "Store", ylab = "Growth", main = "2012 Quarter 2 & 3 Positive Growth", col = c(2, 5, 7, 3))

############## END ##############

########## Question 4 ##########

#filter non-holidays sales

nh\_df <- wss %>%

filter(Holiday\_Flag == 0)

nh\_df

#find non-holidays sales mean by store

nh\_sm <- aggregate(Weekly\_Sales~Store,data=nh\_df,mean)

#rename column name for weekly\_sales to Sales\_Mean

colnames(nh\_sm)<-c("Store","Sales\_Mean\_nh")

nh\_sm

#filter holidays sales into data frame

h\_df <- wss %>%

filter(Holiday\_Flag == 1)

h\_df

#filter holiday sales mean by date

hsd\_sm <- aggregate(Weekly\_Sales~Date,data=h\_df,mean)

colnames(hsd\_sm)<-c("Date","Sales\_Mean\_h")

hsd\_sm

# Compare columns different dataframe with nested for loop

for(i in 1:length(nh\_sm)) {

for (j in 1:length(hsd\_sm)) {

if (nh\_sm$Sales\_Mean\_nh[i] < hsd\_sm$Sales\_Mean\_h[j]){

hsd\_sm$Status = "Y"

}

else{

hsd\_sm$Status[j] = "N"

}

}

}

hsd\_sm

#filter dates with higher means sales

hh\_df <- hsd\_sm %>%

filter(Status == "N")

hh\_df

#Plot the hh\_df

barplot(hh\_df$Sales\_Mean\_h, names = hh\_df$Date, xlab = "Holiday Dates", ylab = "Sales Means", main = "Holidays with higher Sales Means", col = c(2, 5, 7, 3))

############## END ##############

########## Question 5 ##########

# Extract Month and Semester into a columns

wss$Semester <- semester(wss$Date)

wss$Month <- month(wss$Date, label = TRUE)

# Select 3 columns from the modify wss df

mosm\_df <- wss %>%

dplyr::select(Month, Semester, Weekly\_Sales)

mosm\_df

# find sum of sales by Sememster

df\_sem\_sales <- aggregate(Weekly\_Sales~Semester,data=mosm\_df,sum)

colnames(df\_sem\_sales)<-c("Semester","Total\_Weekly\_Sales")

df\_sem\_sales

#find sum of sales by Month

df\_mon\_sales <- aggregate(Weekly\_Sales~Month,data=mosm\_df,sum)

colnames(df\_mon\_sales)<-c("Month","Total\_Weekly\_Sales")

df\_mon\_sales

#find semester with lowest sums of weekly sales

df\_lsm <- df\_sem\_sales %>%

filter(Total\_Weekly\_Sales == min(Total\_Weekly\_Sales))

df\_lsm

#find month with lowest sums of weekly sales

df\_lmo <-df\_mon\_sales %>%

filter(Total\_Weekly\_Sales == min(Total\_Weekly\_Sales))

df\_lmo

#find semester with highest sums of weekly sales

df\_lsm <- df\_sem\_sales %>%

filter(Total\_Weekly\_Sales == max(Total\_Weekly\_Sales))

df\_lsm

#find month with highest sums of weekly sales

df\_lmo <-df\_mon\_sales %>%

filter(Total\_Weekly\_Sales == max(Total\_Weekly\_Sales))

df\_lmo

#Bar plot the hh\_df

barplot(df\_mon\_sales$Total\_Weekly\_Sales, names = df\_mon\_sales$Month, xlab = "Months", ylab = "Total Sales", main = "Total Weekly Sales By Month", col = c(2, 5))

#3d pie chart plot

#pie3D(df\_sem\_sales$Total\_Weekly\_Sales,radius=1,col=rainbow(4),labels=c(df\_sem\_sales$Semester), main = "Total Weekly Sales By Semester")

#pie3D(df\_sem\_sales$Total\_Weekly\_Sales,radius=1,col=rainbow(4),labels=lbls, explode = 0.2, main = "Total Weekly Sales By Semester")

lbls <- c(df\_sem\_sales$Semester)

pct <- round(df\_sem\_sales$Total\_Weekly\_Sale/sum(df\_sem\_sales$Total\_Weekly\_Sale)\*100)

lbls <- paste(lbls, pct) # add percents to labels

lbls <- paste(lbls,"%",sep="") # ad % to labels

pie3D(df\_sem\_sales$Total\_Weekly\_Sale,labels = lbls, col=rainbow(length(lbls)),

explode = 0.3,

main="Total Weekly Sales By Semester")

########## Question 6 Statistical Model ###############

#Injest dataset afresh

wss.a <- read.csv("~/R\_files/Project/Walmart/Walmart\_Store\_sales.csv")

#Convert variables

wss.a$Store=as.factor(wss.a$Store)

wss.a$Holiday\_Flag=as.factor(wss.a$Holiday\_Flag)

wss.a$Date=as.Date(wss.a$Date, format="%d-%m-%Y")

str(wss.a)

View(wss.a)

summary(wss.a)

wss.a$WeekDay <- weekdays(wss.a$Date)

str(wss.a)

#Subset Store 1

df\_store1 <- wss.a %>%

filter(Store == 1)

df\_store1

summary(df\_store1)

str(df\_store1)

##histogram of Weekly Sales

ggplot(df\_store1,aes(x=Weekly\_Sales))+

geom\_histogram(bins=30)

#histogram of Weekly Sales showing Holiday

ggplot(df\_store1,aes(x=Weekly\_Sales,fill=Holiday\_Flag))+

geom\_histogram(bins=30)+

facet\_wrap(~Holiday\_Flag)

##scatterplot of Weekly Sales vs Fuel Price on holidays and non-holdays

ggplot(df\_store1,aes(y=Weekly\_Sales,x=Fuel\_Price,col=Holiday\_Flag))+

geom\_point()+

facet\_wrap(~Holiday\_Flag)

##scatterplot of Weekly Sales vs Temperature on holidays and non-holdays

ggplot(df\_store1,aes(y=Weekly\_Sales,x=Temperature,col=Holiday\_Flag))+

geom\_point()+

facet\_wrap(~Holiday\_Flag)

##scatterplot of Weekly Sales vs Unemployment on holidays and non-holdays

ggplot(df\_store1,aes(y=Weekly\_Sales,x=Unemployment,col=Holiday\_Flag))+

geom\_point()+

facet\_wrap(~Holiday\_Flag)

##### While developing the model, iteratively analyse variables for

# - Normality of distribution

# - Extreme values

# - Multiple collinearity

# - Homoscedasticity (even distribution of residuals)

# - p-value of coefficients and R2/F-statistic of the model

# fit1=lm(SellingPrice000s~., data=Housing)

# summary(fit1)

#Remove non-numeric variables

# df\_store1=df\_store1[,-1]

# df\_store1=df\_store1[,-1]

df\_store1=df\_store1[,-2]

str(df\_store1)

pairs.panels(df\_store1, col="red")

#Transform variables

# log10(5)

# 10 ^ 0.69897

df\_store1$Weekly\_Sales <- log10(df\_store1$Weekly\_Sales)

df\_store1$CPI <- log10(df\_store1$CPI)

pairs.panels(df\_store1, col="red")

#Fit the model (1)

# Check for multi-collinearity with Variance Inflation Factor

# Correlated: none VIF=1, moderately 1<VIF<5, \*\* highly 5<VIF<10, ...

### Refit the model (1)

fit1=lm(Weekly\_Sales~Holiday\_Flag+Temperature+Fuel\_Price+Unemployment, data=df\_store1)

summary(fit1) #temp is significant

vif(fit1)

### Refit the model (2)

fit2=lm(Weekly\_Sales~Holiday\_Flag+Temperature+Fuel\_Price+Unemployment+CPI, data=df\_store1)

summary(fit2)

vif(fit2)

### Refit the model (3)

fit3=lm(Weekly\_Sales~Holiday\_Flag+Temperature+Fuel\_Price+CPI, data=df\_store1)

summary(fit3)

vif(fit3)

### Refit the model (4)

fit4=lm(Weekly\_Sales~Holiday\_Flag+Temperature+CPI, data=df\_store1)

summary(fit4)

vif(fit4)

coefplot(fit4)

##Note##

# All the predictors are signification but CPI s most significant,

# We can reject the null hypothesis that says the model has no significant predictor

# based on the P-vlue of the Overall F-test

#which tells us that our model is effective even tho it might be the best#

##visualize the models

coefplot(fit4)

###model comparison

anova(fit1,fit2)

#####Note#####

# Model 42 (Fit2) that include the CPI predictor is slightly better than Model

# High p-value means the CPI predictor is not however significant to completely reject the null hypothesis

# i.e Fit1 model is as well as Fit2

anova(fit2,fit4)

# High p-value means we cannot reject the null hypothesis

# i.e Fit2 model is as well as Fit4 - Unemployment as predictor had no significance

###################################### Models with interactions ##############################################

### Refit the model (1)

fit5=lm(Weekly\_Sales~Holiday\_Flag+Temperature+I(Fuel\_Price^2)+Unemployment, data=df\_store1)

summary(fit5)

### Refit the model (1)

fit6=lm(Weekly\_Sales~Holiday\_Flag+Temperature\*Unemployment+I(Fuel\_Price^2), data=df\_store1)

summary(fit6)

anova(fit5,fit6)

# High p-value means we cannot reject the null hypothesis

# i.e Fit5 model is as well as Fit6 - Different interactions on Unemployment and Fuel Price as predictor had no significance