Import Data

rm(list = ls()) library(dplyr) data <- read.csv('Walmart_Store_sales.csv', header = T) data1 <- data.frame(data)

Analysis Tasks

Basic Statistics tasks

#

Which store has maximum sales

#

Code-

Sales_Stores <- data1[order(data1$Weekly_Sales,decreasing = T),] Sales_Stores

Output-

Store Date Weekly_Sales Holiday_Flag Temperature Fuel_Price CPI

1906 14 24-12-2010 3818686 0 30.59 3.141 182.5446

2764 20 24-12-2010 3766687 0 25.17 3.141 204.6377

1334 10 24-12-2010 3749058 0 57.06 3.236 126.9836

14th store has maximum weeklysale

#

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

#

standard deviation mean of store

Code-

summarise(group_by(data1, Store), mean(Weekly_Sales)) tapply(data1$Weekly_Sales, data1$Store,mean) aggregate(Weekly_Sales ~ Store, data1, sd) aggregate(Weekly_Sales ~ Store, data1, var)

data1[which.max(data1$sd),]

Output-

Store Date Weekly_Sales Holiday_Flag Temperature Fuel_Price CPI Unemployment

14 07-05-2010 1603955 0 72.55 2.835 210.34 7.808

sd

14 317569.9

14th store has max SD

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

Creating Quaters

head(data1) Q2_2012 <- mutate(data1, start_time =1-04-2012, end_time =30-06-2012) Q3_2012 <- mutate(data1, start_time =1-07-2012, end_time =30-09-2012)

#

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

#

Nonholidaysales <- data1%>%group_by(Weekly_Sales)%>%filter(Holiday_Flag==0) Avg_Nonholidaysales <- mean(Nonholidaysales$Weekly_Sales) Avg_Nonholidaysales

filter(data1,Weekly_Sales>Avg_Nonholidaysales & Holiday_Flag==1)

Output-

Store Date Weekly_Sales Holiday_Flag Temperature Fuel_Price CPI

1 12-02-2010 1641957 1 38.51 2.548 211.2422

1 10-09-2010 1507461 1 78.69 2.565 211.4952

1 26-11-2010 1955624 1 64.52 2.735 211.7484

1 31-12-2010 1367320 1 48.43 2.943 211.4049

1 11-02-2011 1649615 1 36.39 3.022 212.9367

1 09-09-2011 1540471 1 76.00 3.546 215.8611

#

Change dates into days by creating new variable.

#

convert date to YY-MM-DD

as.Date(data1$Date, format = "%d-%m-%Y") data1$Date <- as.Date(data1$Date, format = "%d-%m-%Y") data1

Creating days variable by the help of baseline date

data1$Date <- as.character(data1$Date) baseline_date <- as.Date('2010-02-05') data1$Days <- as.numeric(as.Date(data1$Date) - baseline_date) data1

Output

Days

0

7

14

21

28

35

42

Split Date into Year/Month/Day

data1$Date <- as.character(data1$Date) # convert date to cher d <- strsplit(data1$Date, '-') d <- as.numeric(unlist(d)) d <- matrix(d, dim(data1)[1], 3, byrow=T) data1$Year <- d[,1] data1$Month <- d[,2] data1$Day <- d[,3] data1

Output-

Days Year Month Day

0 2010 2 5

7 2010 2 12

14 2010 2 19

21 2010 2 26

28 2010 3 5

#

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

#

weekly Sales by month

data1%>%group_by(Month)%>% summarise(Mean_Weekly_Sales = mean(Weekly_Sales))

Output-

Month Mean_Weekly_Sales

1 923885.

2 1053200.

3 1013309.

4 1026762.

5 1031714.

6 1064325.

weekly Sales by year

data1%>%group_by(Year)%>% summarise(Mean_Weekly_Sales = mean(Weekly_Sales))

Output

Year Mean_Weekly_Sales

2010 1059670.

2011 1046239.

2012 1033660.

#

Statistical Model

#

#

Hypothesize if CPI, unemployment, and fuel price have any impact on sales.

#

#

Relation between Weekly_Sales and CPI

#

Ho: There is no linear relationship between Weekly_Sales and CPI

Ha: There is linear relationship between Weekly_Sales and CPI

model1 <- lm (Weekly_Sales ~ CPI, data=data1) summary(model1) p_value = 5.438e-09 alpha = 0.05 p_value < alpha

Output-

TRUE ,so their is relationship

#

Relation between Weekly_Sales and Unemployment

#

Ho: There is no linear relationship between Weekly_Sales and Unemployment

Ha: There is linear relationship between Weekly_Sales and Unemployment

model2 <- lm (Weekly_Sales ~ Unemployment, data=data1) summary(model2) p_value = 2.2e-16 alpha = 0.05 p_value < alpha

Output-

TRUE ,so their is relationship

#

Relation between Weekly_Sales and Fuel price

#

Ho: There is no linear relationship between Weekly_Sales and Fuel price

Ha: There is linear relationship between Weekly_Sales and Fuel price

model3 <- lm (Weekly_Sales ~ Fuel_Price, data=data1) summary(model3) p_value = 0.4478 alpha = 0.05 p_value < alpha # FALSE ,so their is no relationship

Output-

FALSE ,so their is no relationship

#

Build prediction models to forecast demand

#

Creating New coulmn for model building By Droping Store and Date Coulmn

col.vars <- c('Holiday_Flag','Temperature', 'Fuel_Price', 'CPI', 'Unemployment','Weekly_Sales') datamodel <- data1[,col.vars]

Model Building

model4 <- lm (Weekly_Sales ~ ., datamodel) summary(model4) Rsqd1 <- summary(model4)$r.squared Rsqd1 # 0.02544366

predicted_y1 <- predict(model4, datamodel) RMSE1 = sqrt(mean((data1$Weekly_Sales - predicted_y1)^2)) RMSE1 # 557097.3

Summary-In this model we use all varriables to check the impact

considering those independent where the value is higher

model5 <- lm(Weekly_Sales ~ Unemployment + CPI+ Temperature, datamodel) summary(model5) Rsqd2 <- summary(model5)$r.squared Rsqd2 # 0.02423897

predicted_y2 <- predict(model5, datamodel)

RMSE2 = sqrt(mean((datamodel$Weekly_Sales - predicted_y2)^2)) RMSE2 # 557441.5

Summary in model5 we are taken the high value that the co-related to Weekly_Sales like Unemployment & CPI & Temperature

model6 <- lm(Weekly_Sales ~ log(Unemployment) + CPI+ Temperature, datamodel) summary(model6) Rsqd3 <- summary(model6)$r.squared Rsqd3 # 0.02270446

predicted_y3 <- predict(model6, datamodel)

RMSE3 = sqrt(mean((datamodel$Weekly_Sales - predicted_y3)^2)) RMSE3 # 557879.7

Summary- In model6 we take the log of Unemployment & CPI & Temperature for better linearity

model7 <- lm(Weekly_Sales ~ Unemployment + CPI+ Temperature + Holiday_Flag , datamodel) summary(model7) Rsqd4 <- summary(model7)$r.squared Rsqd4 # 0.02538059

predicted_y2 <- predict(model7, datamodel)

RMSE4 = sqrt(mean((datamodel$Weekly_Sales - predicted_y2)^2)) RMSE4 # 557115.3

Summary- In model7 we take Unemployment & CPI & Temperature & Holiday_Flag for linear model

===

Comparing all models

===

Rsqd_ <- c(Rsqd1,Rsqd2,Rsqd3,Rsqd4) RMSE_ <- c(RMSE1,RMSE2,RMSE3,RMSE4)

Model_Validation <- cbind(Rsqd_,RMSE_) rownames(Model_Validation) <- c("model4 - (all variables)", "model5 - (Weekly_Sales on CPI & Temperature)", "model6 - (Weekly_Sales on log(Unemployment) & CPI & Temperature)", "model7 - (Weekly_Sales on Unemployment + CPI+ Temperature + Holiday_Flag)")

Model_Validation

Output-

Model_Validation

Rsqd_

model4 - (all variables) 0.02544366

model5 - (Weekly_Sales on CPI & Temperature) 0.02423897

model6 - (Weekly_Sales on log(Unemployment) & CPI & Temperature) 0.02270446

model7 - (Weekly_Sales on Unemployment + CPI+ Temperature + Holiday_Flag) 0.02538059

RMSE_

model4 - (all variables) 557097.3

model5 - (Weekly_Sales on CPI & Temperature) 557441.5

model6 - (Weekly_Sales on log(Unemployment) & CPI & Temperature) 557879.7

model7 - (Weekly_Sales on Unemployment + CPI+ Temperature + Holiday_Flag) 557115.3

Summery- By model validation technique we can see which model would perfome best between all models