**Project 1 - Retail Analysis with Walmart Data**

**Q1 - Which store has maximum sales?**

**Ans 1 :**

Code:

getwd()

setwd('F:/PGP Data Science - Purdue University/PG DS - Data Science with R - Course 3/Project/Assessment Project -1/')

library(csv)

my\_walmart = read.csv('Walmart\_Store\_sales.csv')

### changed date format in data frame since it was in character format before ###

my\_walmart$Date = as.Date(my\_walmart$Date,format = "%d-%m-%Y")

library(dplyr)

max\_sales = aggregate(Weekly\_Sales~Store,my\_walmart,sum)

class(max\_sales)

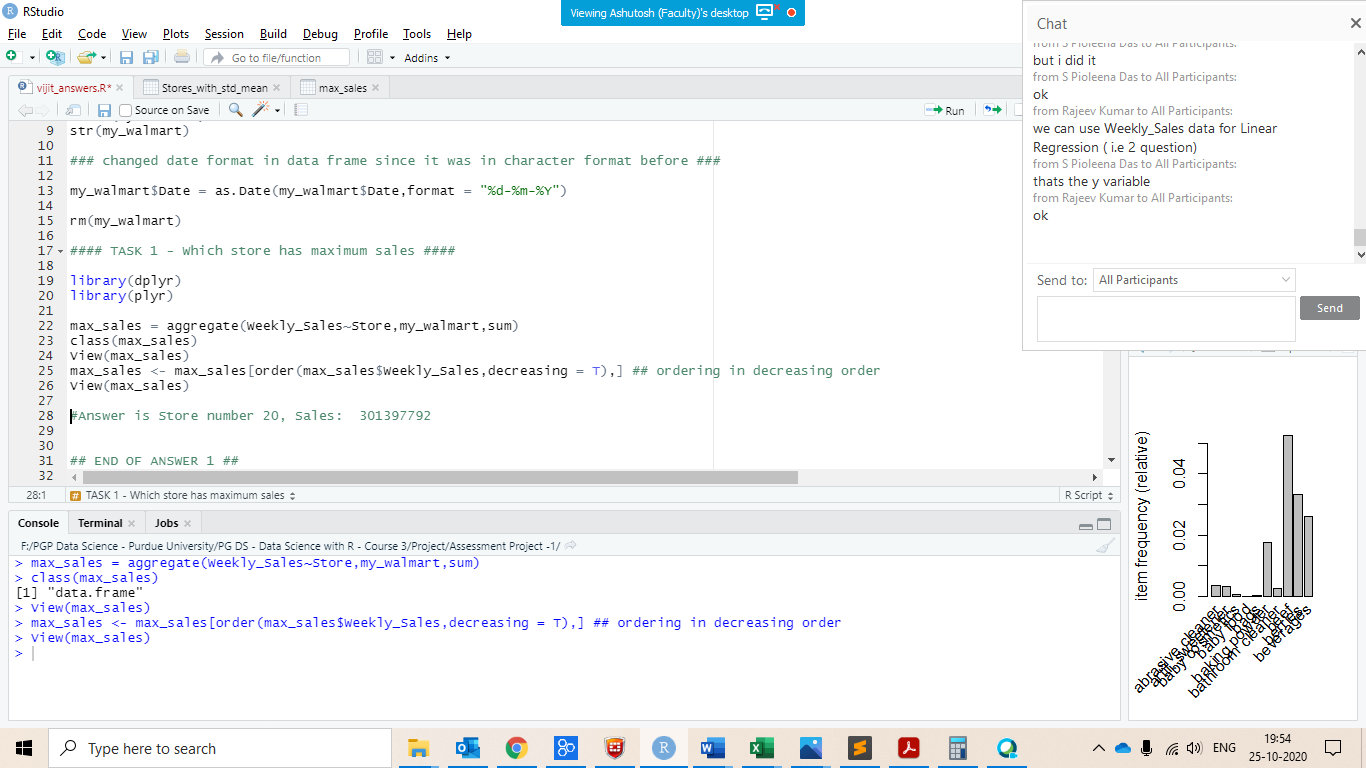
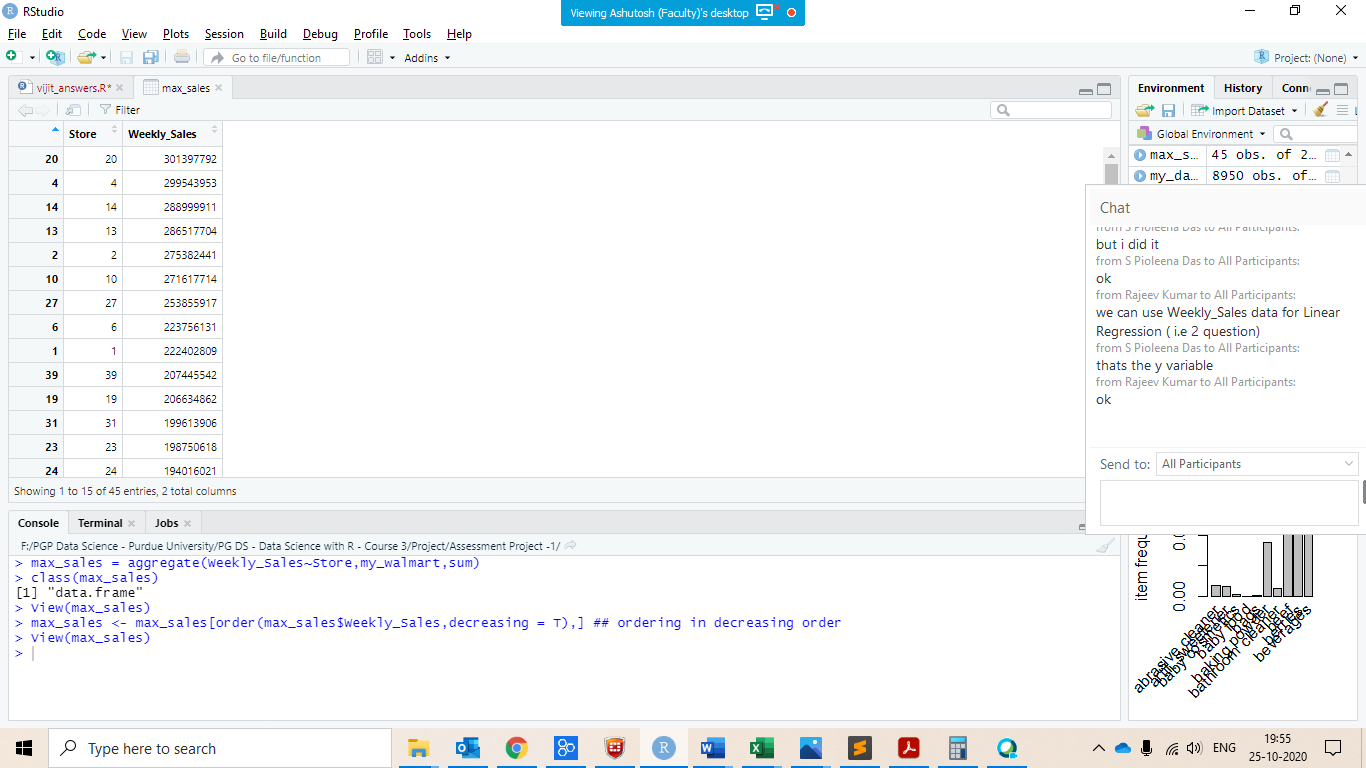
View(max\_sales)

max\_sales <- max\_sales[order(max\_sales$Weekly\_Sales,decreasing = T),] ## ordering in decreasing order

View(max\_sales)

#Answer is Store number 20, Sales: 301397792

Screenshot:

(i) (ii)

#Answer is Store number 20, Sales: 301397792

**--------------------------------------------------------------------------------------------------------------------------------------------------------------**

**Q2 - Which store has maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of mean to standard deviation?**

**Ans 2 –**

Code :

Stores\_with\_mean = aggregate(Weekly\_Sales~Store,my\_walmart,mean) #mean

Stores\_with\_mean

Stores\_with\_std= aggregate(Weekly\_Sales~Store,my\_walmart,sd) #standard deviation

Stores\_with\_std

Stores\_with\_std <- Stores\_with\_std[order(Stores\_with\_std$Weekly\_Sales,decreasing = T),]

View(Stores\_with\_std ) ## ANSWER - STORE 14 has maximum maximum standard deviation 317569.95

Stores\_with\_std\_mean = cbind(Stores\_with\_mean,Stores\_with\_std)

str(Stores\_with\_std\_mean)

View(Stores\_with\_std\_mean)

colnames(Stores\_with\_std\_mean)[2] = "Mean\_value" # renaming columns

colnames(Stores\_with\_std\_mean)[4] = "Std\_value" # renaming columns

View(Stores\_with\_std\_mean)

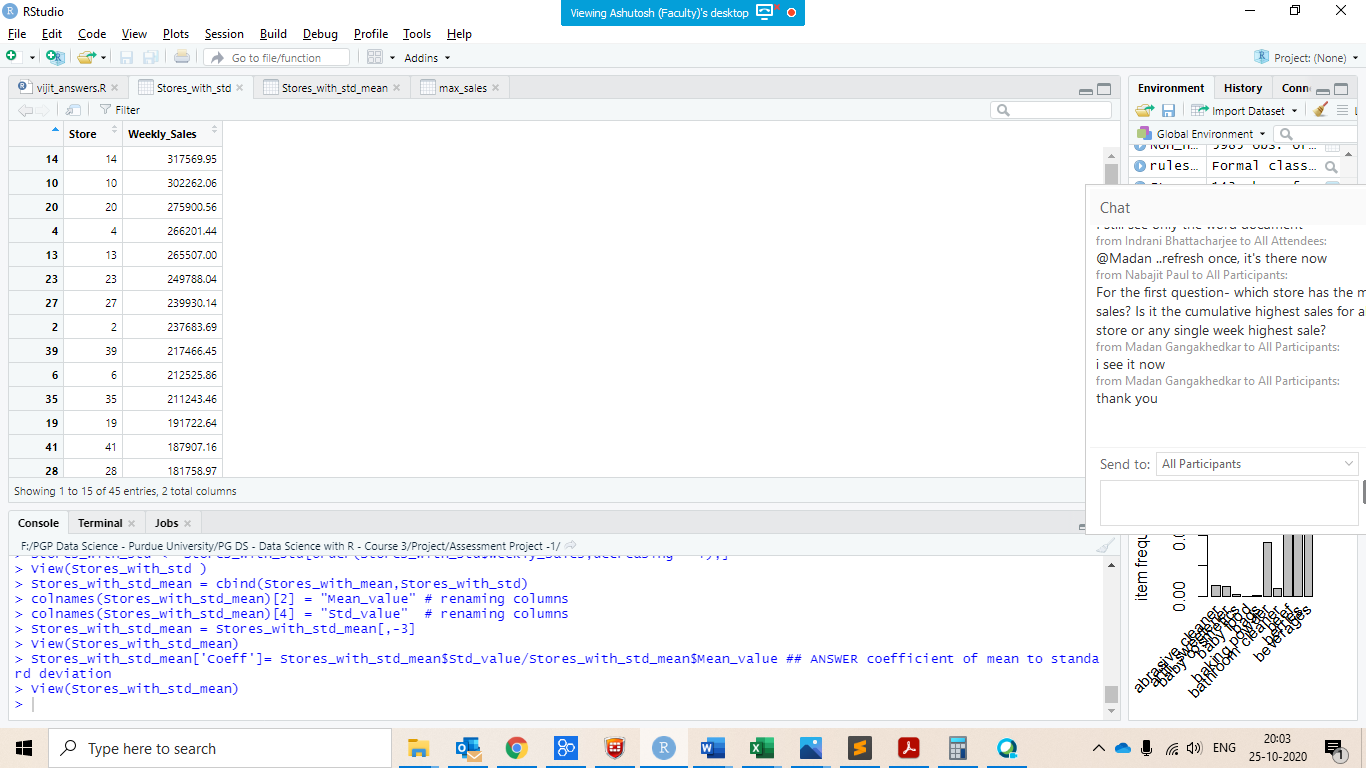
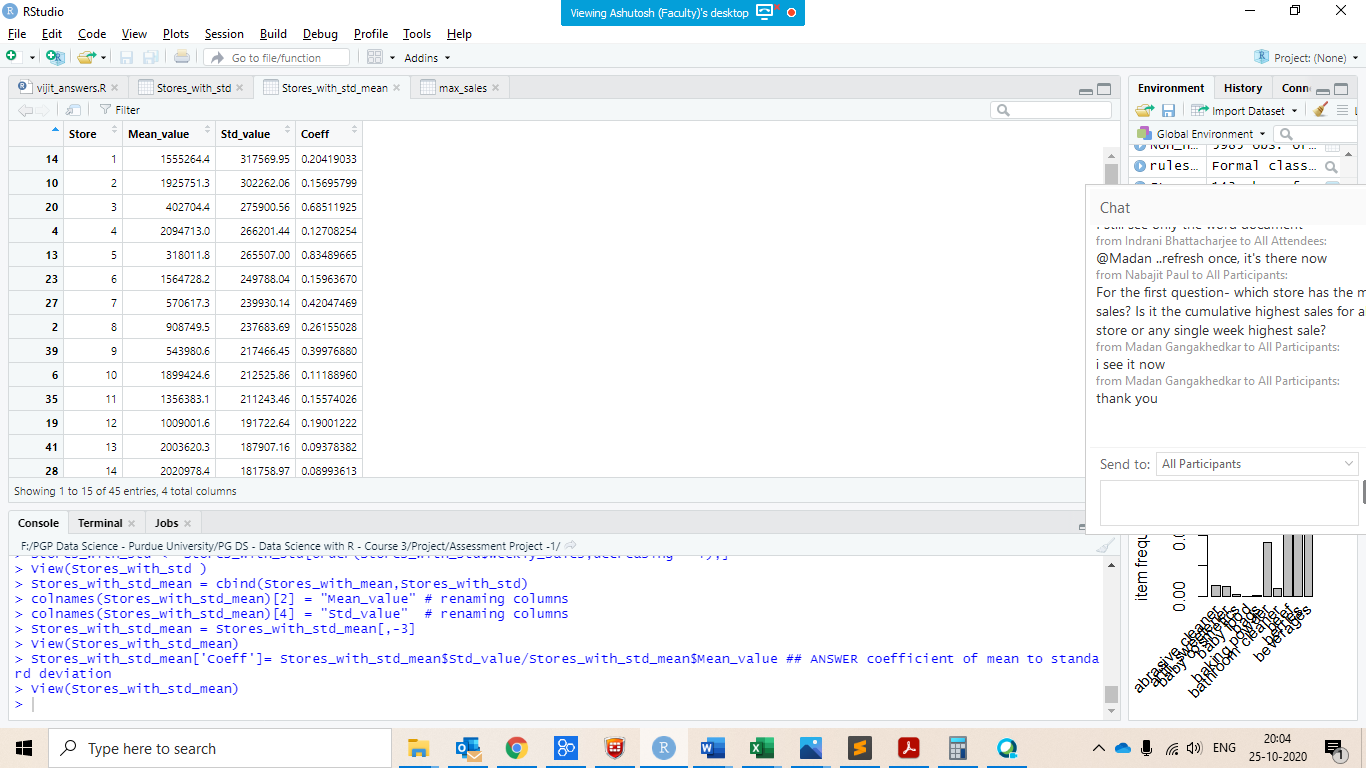
Stores\_with\_std\_mean = Stores\_with\_std\_mean[,-3]

View(Stores\_with\_std\_mean)

Stores\_with\_std\_mean['Coeff’] = Stores\_with\_std\_mean$Std\_value/Stores\_with\_std\_mean$Mean\_value ## ANSWER coefficient of mean to standard deviation

View(Stores\_with\_std\_mean)

Screenshot:

1.  (ii) 

**----------------------------------------------------------------------------------------------------------------------------------------------------------------**

**Q 3 - Which store/s has good quarterly growth rate in Q3’2012?**

**Ans 3 -**

Code :

## taken FINANACIAL YEAR FROM JAN - DEC

Stores\_Quarter\_Flag = transform(my\_walmart,Q\_Flag= ifelse((Date>='2012-04-01' & Date<= '2012-06-30'),"Q2\_2012",

ifelse((Date>='2012-07-01' & Date<= '2012-09-30'),"Q3\_2012","-")))

View(Stores\_Quarter\_Flag)

# summarizing

Stores\_with\_Q\_Flag\_sum = aggregate(Weekly\_Sales~Store+Q\_Flag,Stores\_Quarter\_Flag,sum)

View(Stores\_with\_Q\_Flag\_sum)

str(Stores\_with\_Q\_Flag\_sum)

##then transposing using reshaping

Stores\_with\_Q\_Flag\_sum\_t = reshape(Stores\_with\_Q\_Flag\_sum,idvar="Store",timevar ='Q\_Flag',direction="wide")

View(Stores\_with\_Q\_Flag\_sum\_t)

class(Stores\_with\_Q\_Flag\_sum\_t)

## calculating growth rate Q3’2012

Stores\_with\_Q\_Flag\_sum\_t\_GR = transform(Stores\_with\_Q\_Flag\_sum\_t,

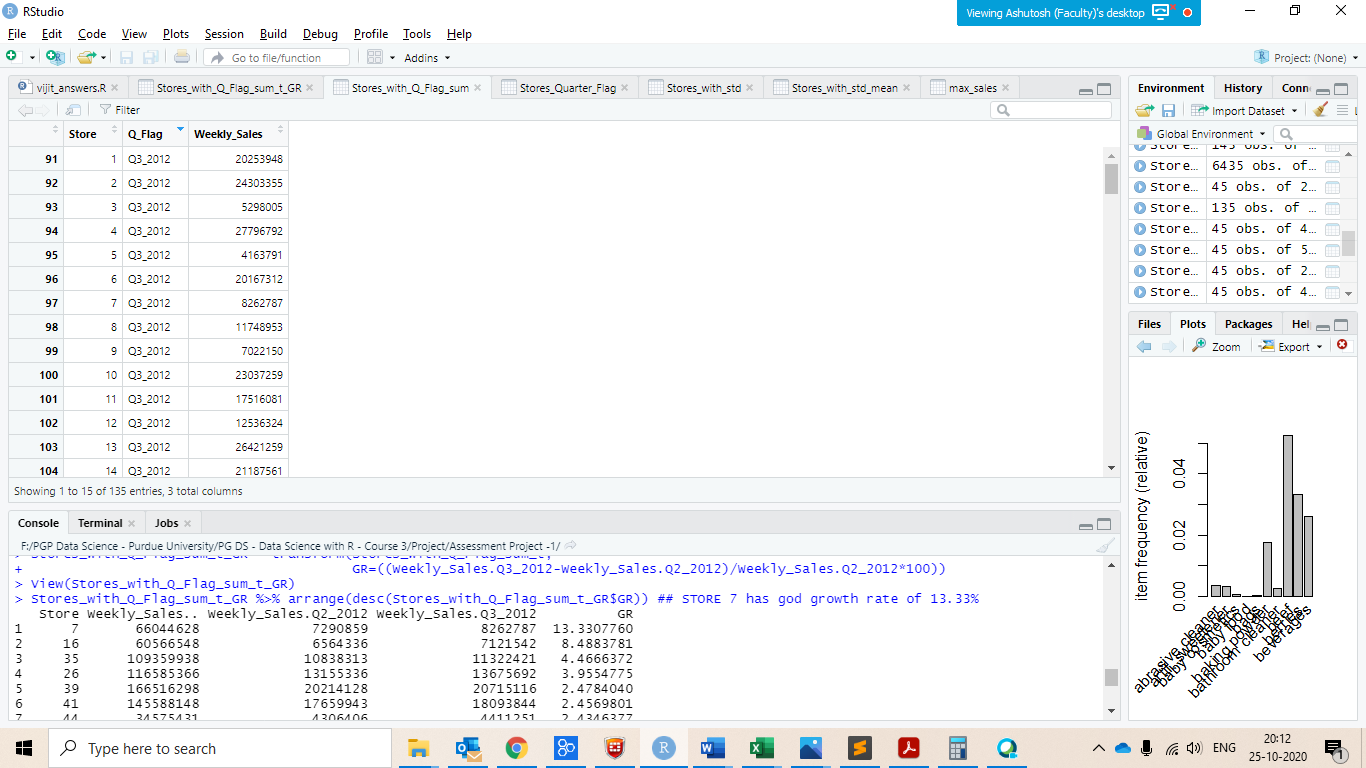
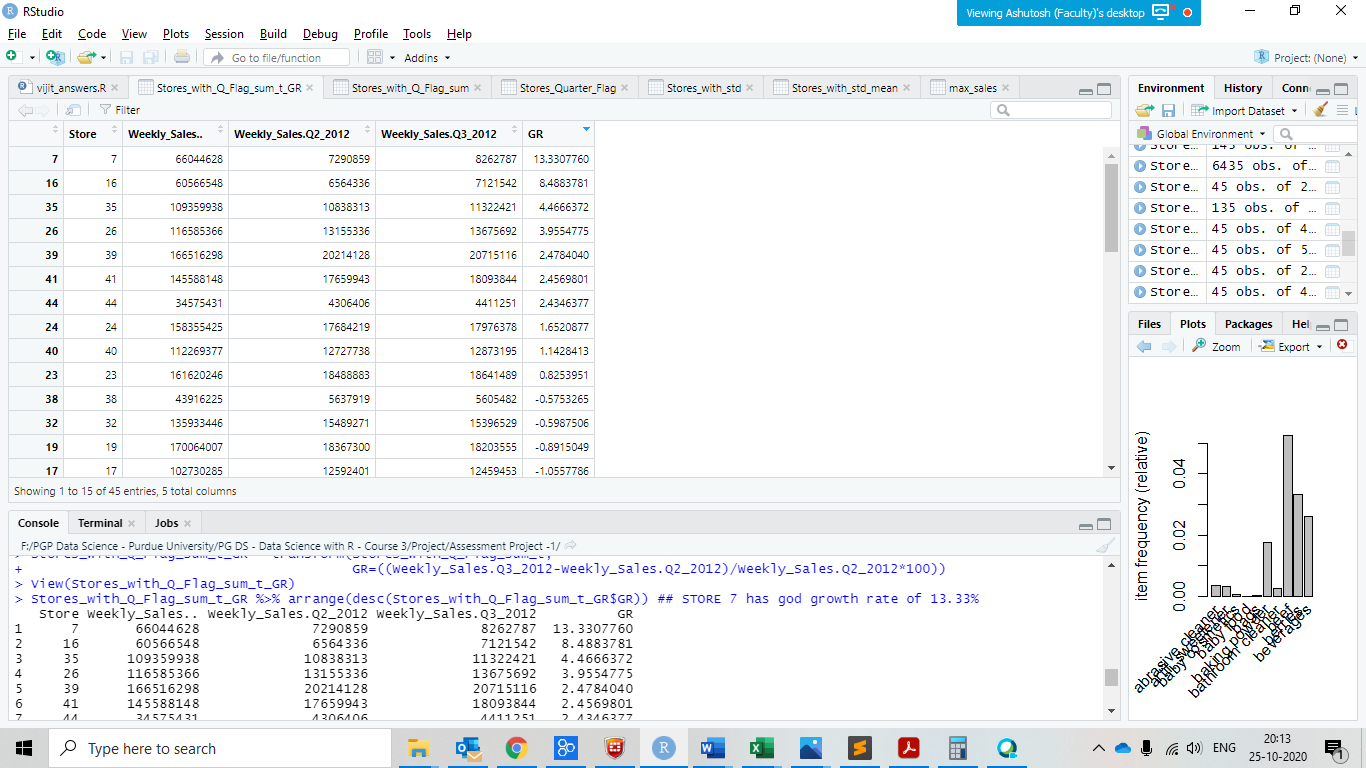
GR=((Weekly\_Sales.Q3\_2012-Weekly\_Sales.Q2\_2012)/Weekly\_Sales.Q2\_2012\*100))

View(Stores\_with\_Q\_Flag\_sum\_t\_GR)

Stores\_with\_Q\_Flag\_sum\_t\_GR %>% arrange(desc(Stores\_with\_Q\_Flag\_sum\_t\_GR$GR))

## STORE 7 has highest growth rate of 13.33%

Screenshot:

**(i)** **(ii)** 

Insights :  
Store 7,16,35,26,39,41,44,24 ,40 , 23 has positive growth rate, Store 7 has highest growth rate of 13.33%

**----------------------------------------------------------------------------------------------------------------------------------------------------**

**Q4 - 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?**

**Ans 4-**

Code:

View(my\_walmart)

Non\_Holiday\_sales = my\_walmart%>%group\_by(Weekly\_Sales)%>%filter(Holiday\_Flag==0)# filtering non holidays weekly\_sales

View(Non\_Holiday\_sales)

avg\_non\_holiday\_sales = mean(Non\_Holiday\_sales$Weekly\_Sales) #mean of weekly sales of non holidays sales

avg\_non\_holiday\_sales

## holiday name added and holidays which have higher sales than the mean sales in non-holiday season

high\_sales = filter(my\_walmart,Weekly\_Sales>avg\_non\_holiday\_sales & Holiday\_Flag==1)

high\_sales = transform(high\_sales,Holiday\_name = ifelse((Date=='2010-02-12' | Date== '2011-02-11' | Date== '2012-02-10' | Date=='2013-02-8'),"Super Bowl"

,ifelse((Date=='2010-09-10' | Date== '2011-09-09'| Date== '2012-09-7'| Date=='2013-09-6'),"Labour Day"

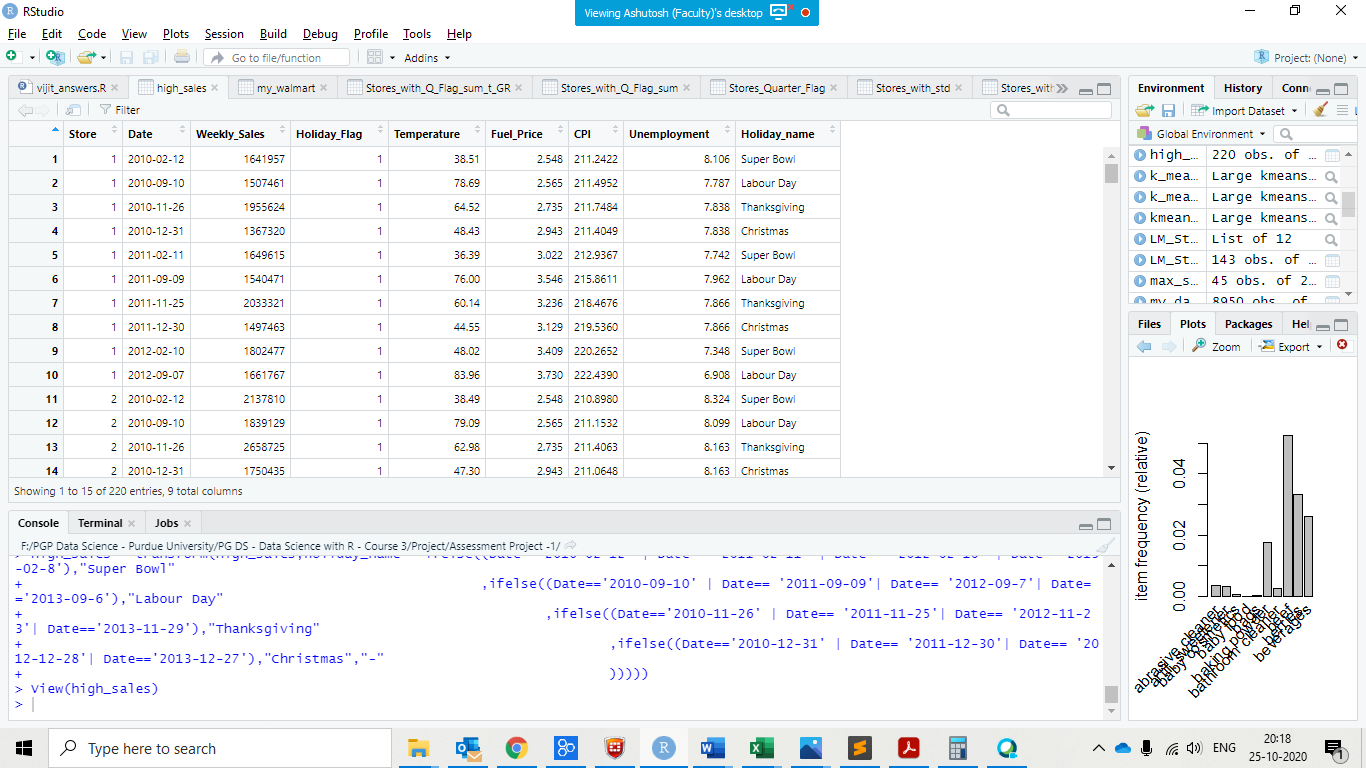
,ifelse((Date=='2010-11-26' | Date== '2011-11-25'| Date== '2012-11-23'| Date=='2013-11-29'),"Thanksgiving"

,ifelse((Date=='2010-12-31' | Date== '2011-12-30'| Date== '2012-12-28'| Date=='2013-12-27'),"Christmas","-"

)))))

View(high\_sales) ## ANSWER

Screenshot :

(i)

**--------------------------------------------------------------------------------------------------------------------------------------------------------------**

**Q5 - Provide a monthly and semester view of sales in units and give insights?**

**Ans 5 –**

**Code:**

View(my\_walmart)

my\_walmart$Date = as.Date(my\_walmart$Date,format = "%d-%m-%Y")

install.packages('lubridate')

library(lubridate)

my\_walmart = transform(my\_walmart, Months = months(as.Date(Date,format="%B"))) ## Months column extracted and added

View(my\_walmart)

my\_walmart = transform(my\_walmart, Year = year(as.Date(Date,format="%Y"))) ## Year column extracted and added

View(my\_walmart)

#mean of monthly Sales

install.packages('dplyr')

library(dplyr)

library(plyr)

Monthly\_Sales = my\_walmart%>%group\_by(Months)%>%summarise(sum\_Monthly\_Sales=sum(Weekly\_Sales))

View(Monthly\_Sales) ## Overall July month has most sales and January has lowest sales

Monthly\_mean\_Sales = my\_walmart%>%group\_by(Months)%>%summarise(Mean\_Monthly\_Sales=mean(Weekly\_Sales))

View(Monthly\_mean\_Sales) ## DEC month as highest average sales, January Month has lowest average sales

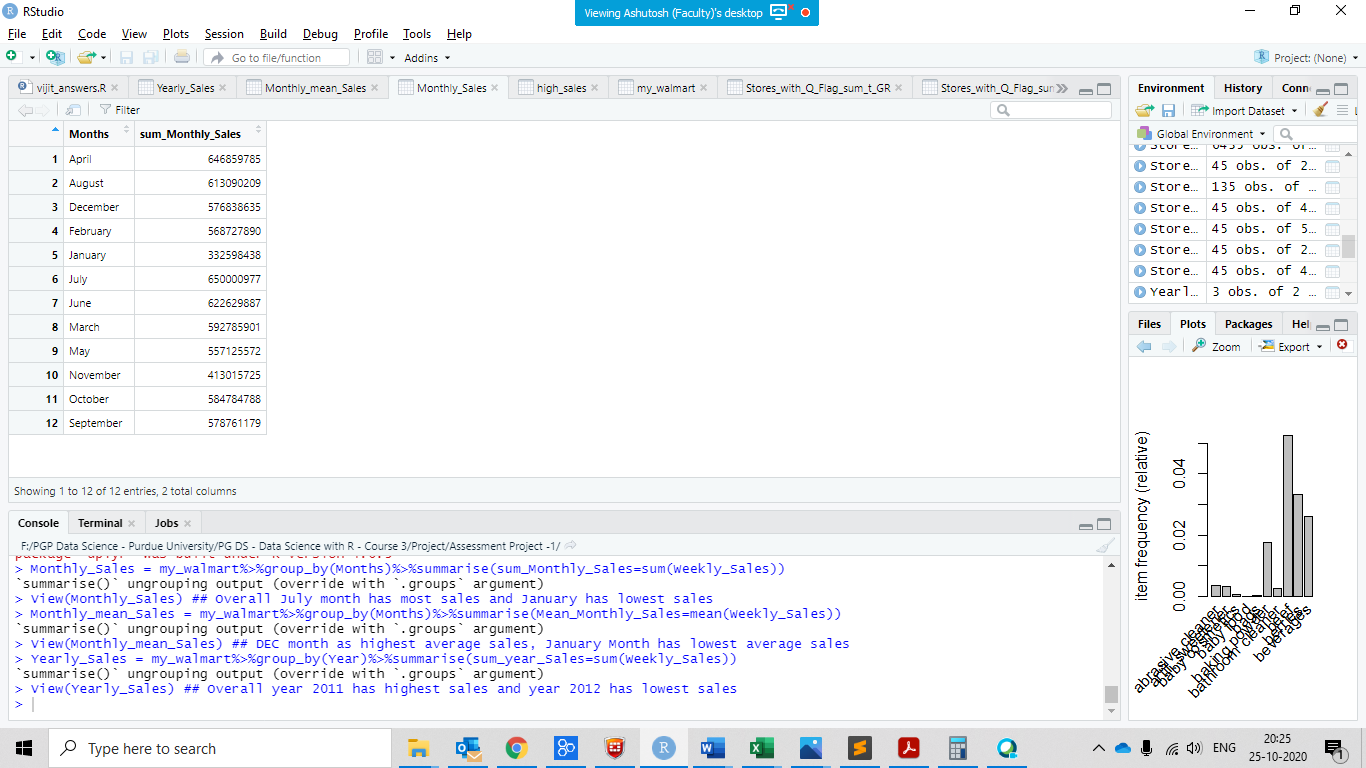
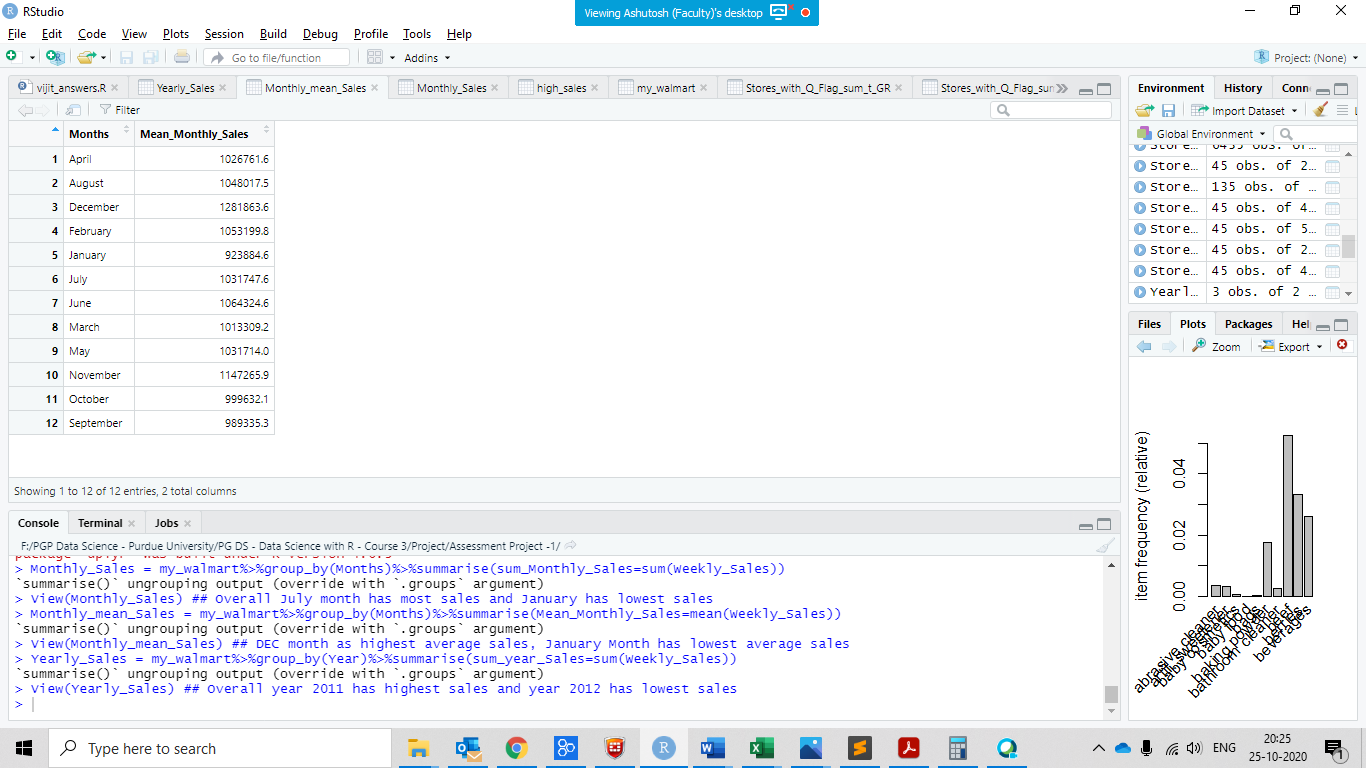
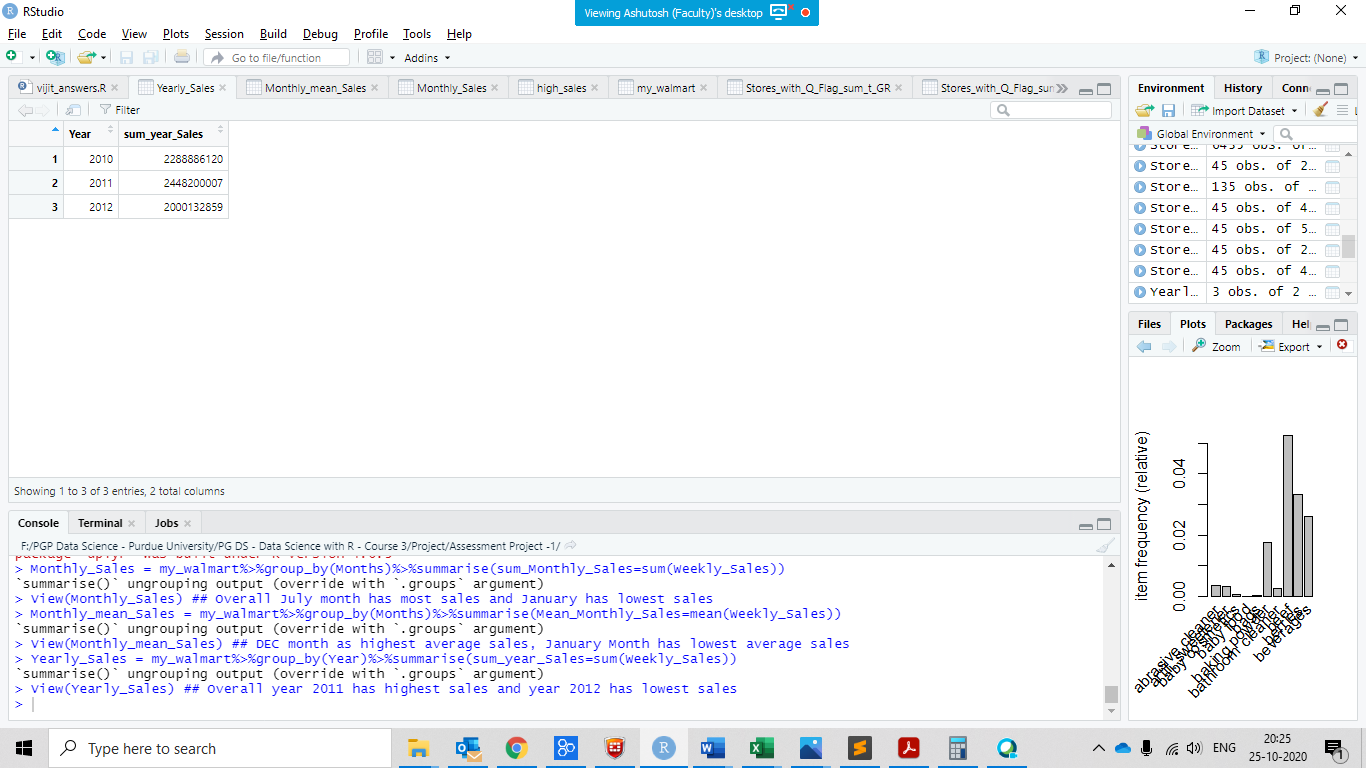
Yearly\_Sales = my\_walmart%>%group\_by(Year)%>%summarise(sum\_year\_Sales=sum(Weekly\_Sales))

View(Yearly\_Sales) ## Overall year 2011 has highest sales and year 2012 has lowest sales

#Yearly\_mean\_Sales = my\_walmart%>%group\_by(Year)%>%summarize(Mean\_Year\_Sales=mean(Weekly\_Sales))

#View(Yearly\_mean\_Sales) ## year 2010 has highest average sales, year 2012 has lowest average sales

Screenshots:

(i) (ii)  (iii) 

Insights:

1 - Overall July month has the most sales and January Month has the lowest sales

2- DEC month as highest average sales, January Month has lowest average sales

3- Overall year 2011 has highest sales and year 2012 has lowest sales

**----------------------------------------------------------------------------------------------------------------------------------------------------------------**

**Q 6 - 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.**

**Ans 6 –**

Code:

library(lubridate)

my\_walmart = read.csv('Walmart\_Store\_sales.csv')

View(my\_walmart)

Str(my\_walmart)

# Extracting Store 1 data from my\_walmart

Store\_one\_data = my\_walmart %>% filter(Store == 1)

View(Store\_one\_data)

days = c(1:143) ## creating vector for dates

Store\_one\_data = cbind(Store\_one\_data,day\_id = days) ## restructure dates as 1 for 5 Feb 2010

View(Store\_one\_data)

str(Store\_one\_data)

summary(Store\_one\_data)

**## Linear model**

LM\_Store\_one = lm(Weekly\_Sales~Fuel\_Price+CPI+Unemployment ,Store\_one\_data)

summary(LM\_Store\_one)

# Fuel price is insignificant variable here and can be dropped

LM\_Store\_one = lm(Weekly\_Sales~ CPI+Unemployment ,Store\_one\_data)

summary(LM\_Store\_one)

# Unemployment is also insignificant variable here and can be dropped

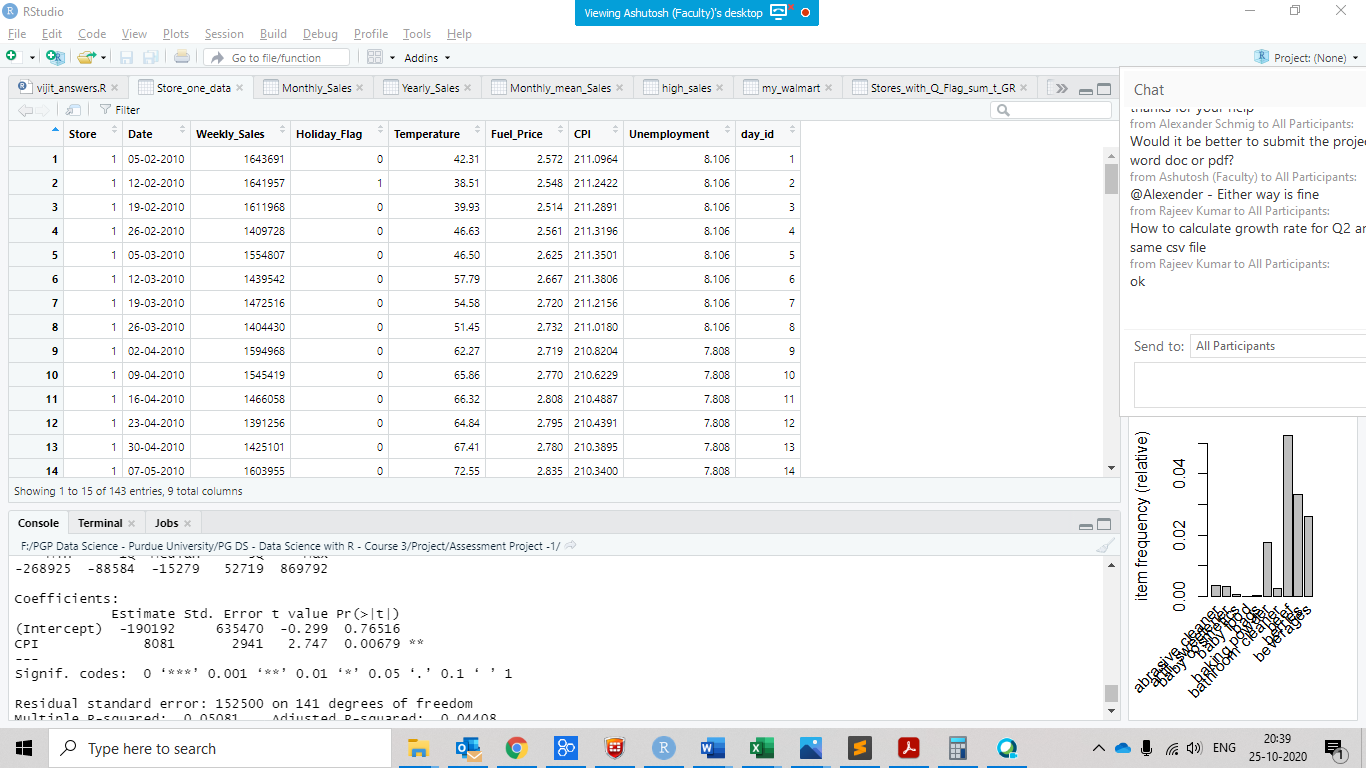
LM\_Store\_one = lm(Weekly\_Sales~ CPI,Store\_one\_data)

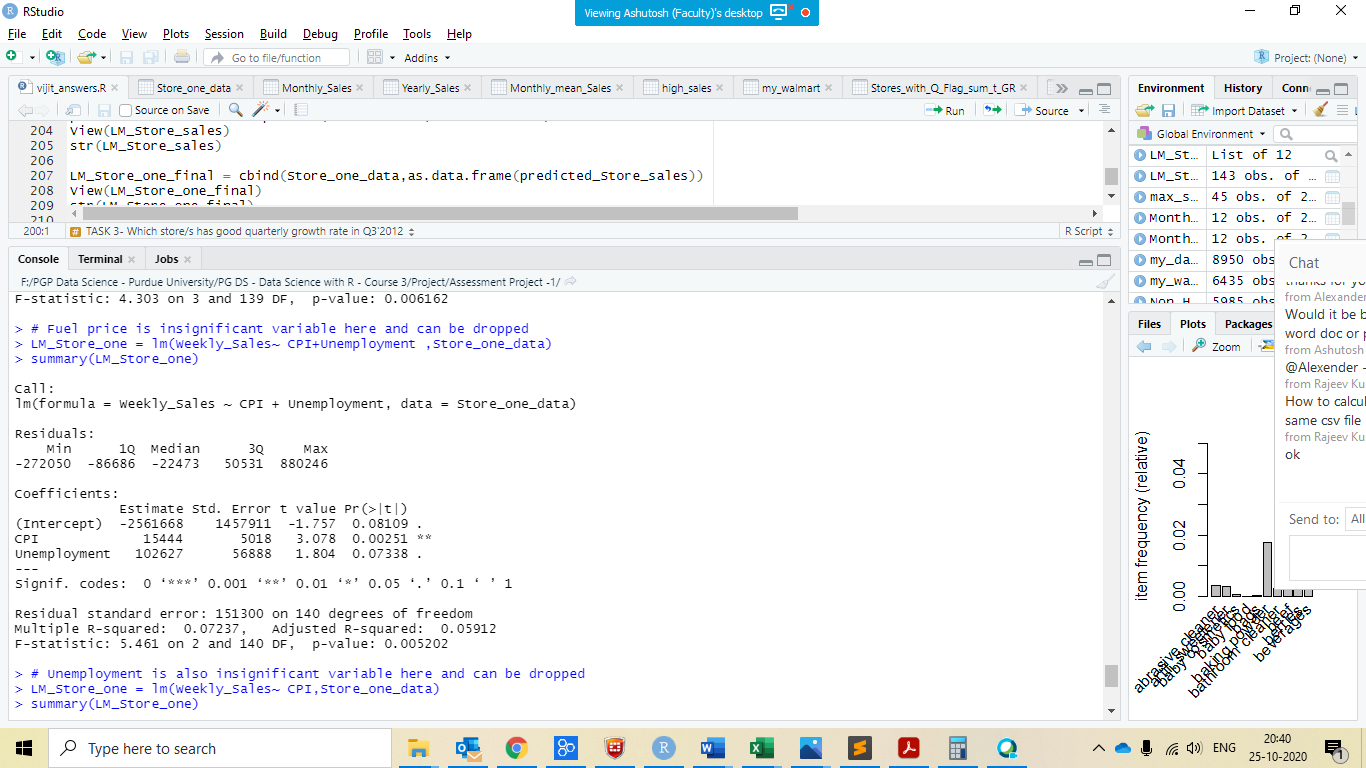
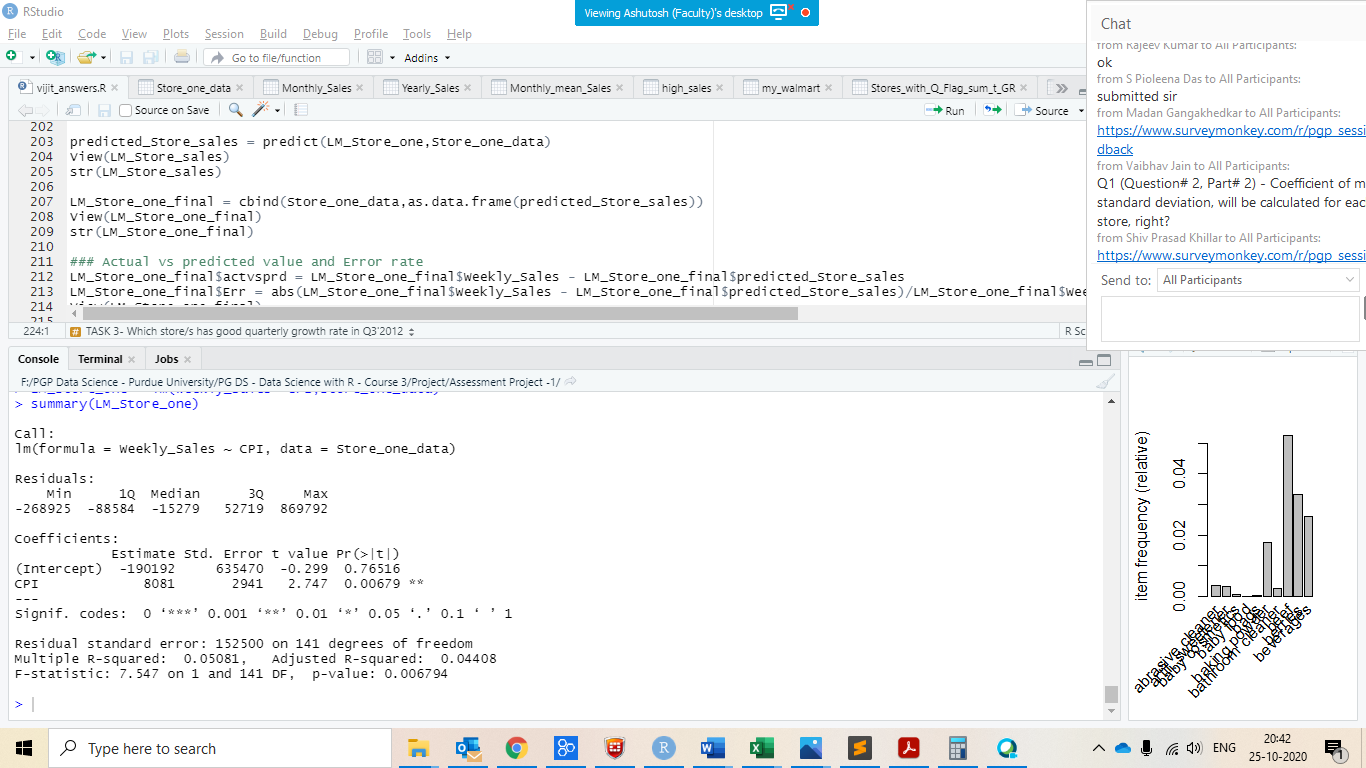
summary(LM\_Store\_one)

## intercept : -190192 , R square : 0.05, Adjusted R square : 0.044, p value : 0.006 < 0.05

#Consumer Price Index has impact on Weekly Sales

Screenshots:

(i)

(ii) (iii) 

Insights:

1 - Fuel price and Unemployment are insignificant variables here

2 – p value of CPI is < 0.05 , Consumer Price Index has impact on Weekly Sales

3 – better the CPI the weekly sales is high

4 - intercept : -190192 , R square : 0.05, Adjusted R square : 0.044, p value : 0.006 < 0.05