walmart <- read.csv(file.choose(), stringsAsFactors = FALSE)

str(walmart)

library(lubridate)

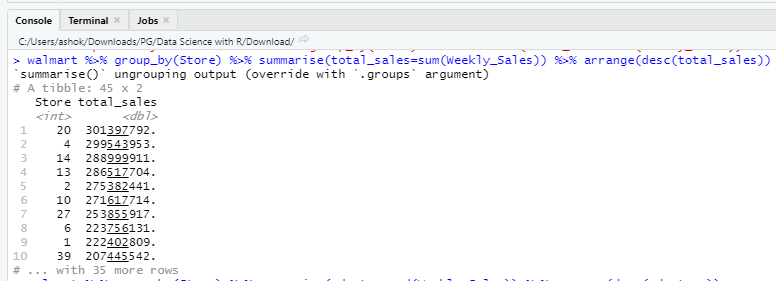
walmart$Date <- dmy(walmart$Date)

str(walmart)

# 1. which store has max sales

library(dplyr)

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



Result: Store 20

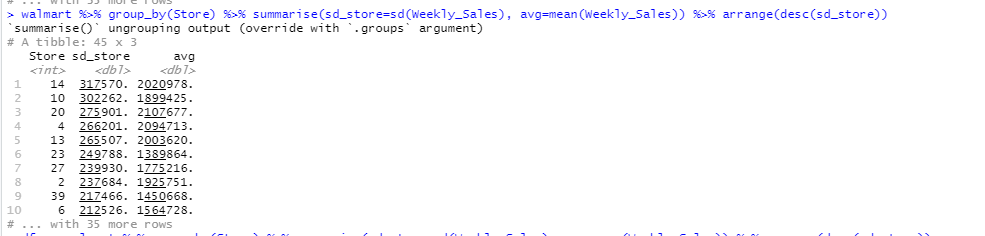
#2. max standard deviation

df <- walmart %>% group\_by(Store) %>% summarise(sd\_store=sd(Weekly\_Sales), avg=mean(Weekly\_Sales)) %>% arrange(desc(sd\_store))

#coefficient of mean to sd = (sd/mean)\*100

df$cv <- (df$sd\_store/df$avg)\*100

df



Result : Store 14

#3. which stores has good quarterly growth rate

walmart$qtr <- quarter(wallmart$Date, with\_year = TRUE)

#aggregate by quarter

qtr\_sales <- walmart %>% group\_by(Store, qtr) %>% summarise(qtrly\_sales=sum(Weekly\_Sales))

View(qtr\_sales)

qtr\_sales$lag\_4 <- lag(qtr\_sales$qtrly\_sales,4)

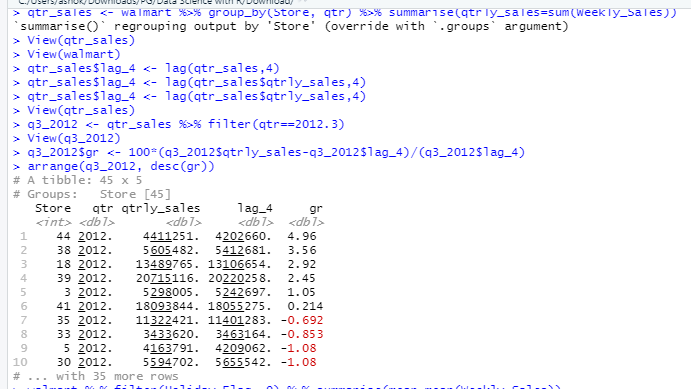
View(qtr\_sales)

q3\_2012 <- qtr\_sales %>% filter(qtr==2012.3)

View(q3\_2012)

q3\_2012$gr <- 100\*(q3\_2012$qtrly\_sales-q3\_2012$lag\_4)/(q3\_2012$lag\_4)

arrange(q3\_2012, desc(gr))



#4. impact of holidays

walmart %>% filter(Holiday\_Flag==0) %>% summarise(mean=mean(Weekly\_Sales))

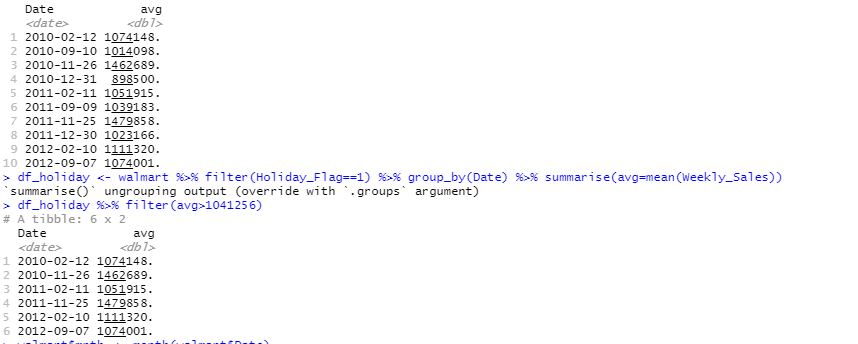
#mean sales for all stores = 1041256

#holidays which have higher sales

#those weeks which have holidays

df\_holiday <- walmart %>% filter(Holiday\_Flag==1) %>% group\_by(Date) %>% summarise(avg=mean(Weekly\_Sales))

df\_holiday %>% filter(avg>1041256)



#5. provide monthly and semester view of sales

walmart$mnth <- month(walmart$Date)

walmart$yr <- year(walmart$Date)

walmart$sem <- semester(walmart$Date, with\_year = TRUE)

#concatenate date and year

walmart <- walmart %>% mutate(mon\_yr= paste(yr, mnth))

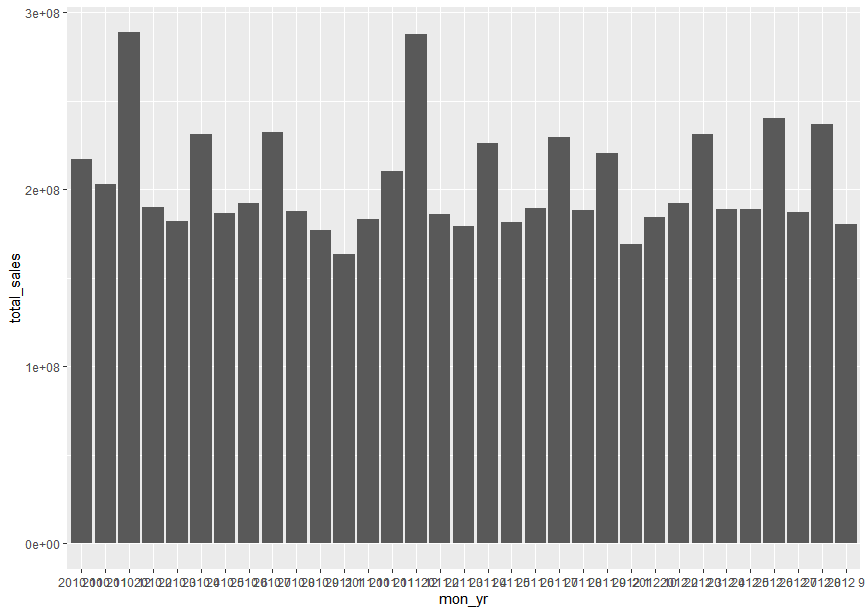
#provide monthly view

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

#bar plot to describe data

library(ggplot2)

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

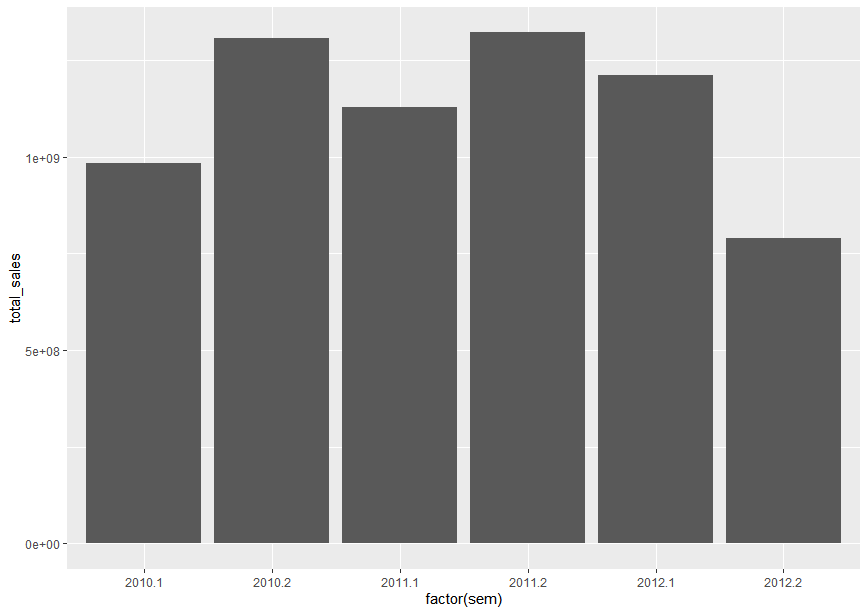


#semester aggregate

semyr <- walmart %>% group\_by(sem) %>% summarise(total\_sales=sum(Weekly\_Sales))

#bar plot to describe data

ggplot(data = semyr, mapping = aes(factor(sem), total\_sales)) +geom\_bar(stat = 'identity')



#statistical model

#for store 1

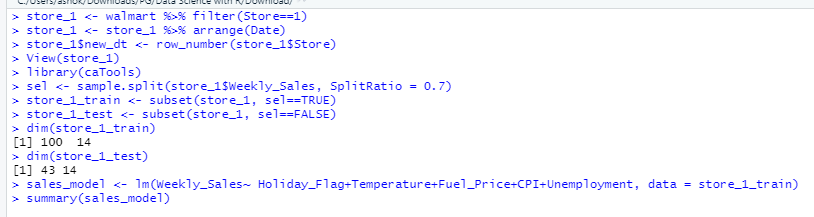
store\_1 <- walmart %>% filter(Store==1)

store\_1 <- store\_1 %>% arrange(Date)

#restructure dates 1 for fst date

store\_1$new\_dt <- row\_number(store\_1$Store)

View(store\_1)



#hypothesis if CPI, unemployment, fuel price have any impact on sales

library(caTools)

sel <- sample.split(store\_1$Weekly\_Sales, SplitRatio = 0.7)

#training data set

store\_1\_train <- subset(store\_1, sel==TRUE)

#test data

store\_1\_test <- subset(store\_1, sel==FALSE)

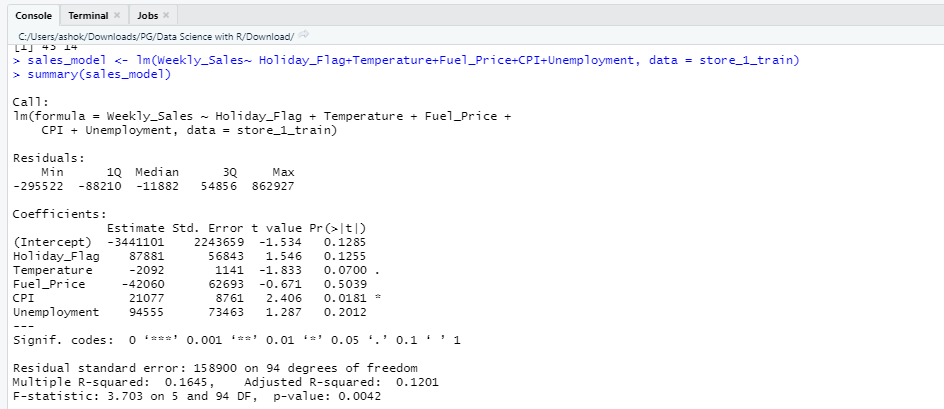
dim(store\_1\_train)

dim(store\_1\_test)

#train model on training data

sales\_model <- lm(Weekly\_Sales~ Holiday\_Flag+Temperature+Fuel\_Price+CPI+Unemployment, data = store\_1\_train)

summary(sales\_model)



#F stats checks if all variables are simultaneously=0

#h0 : all coefficient for the model = 0

#since p <0.05 , so we reject the h0

#look at individual variables now

#H0: coefficient for each variable = 0

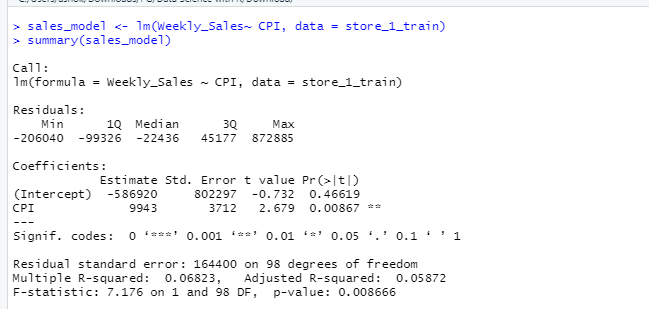
# holiday flag : p>alpha fail to reject, remove from model

# fuel\_price, temp, unemployment : p>alpha fail to reject, remove from model

# CPI : p<alpha , reject h0, use in model.

sales\_model <- lm(Weekly\_Sales~ CPI, data = store\_1\_train)

summary(sales\_model)



#change dates into days

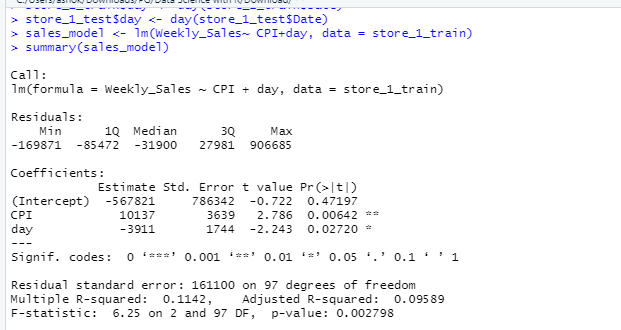
day(store\_1$Date)

store\_1\_train$day <- day(store\_1\_train$Date)

store\_1\_test$day <- day(store\_1\_test$Date)

sales\_model <- lm(Weekly\_Sales~ CPI+day, data = store\_1\_train)

summary(sales\_model)



predict(sales\_model,store\_1\_test)