ROI on Hand Picked Stocks 2007-2020

Janis Corona

2/17/2020

portfolio <- read.csv('all\_portfolio\_prices.csv', header=TRUE, na.strings=c('',' '),  
 row.names=1)

portfolio$Date <- row.names(portfolio)

Vol <- grep('Volume', colnames(portfolio))  
close <- grep('Close',colnames(portfolio))  
Close <- portfolio[,close]  
Volume <- portfolio[,Vol]  
colnames(Close)

## [1] "TGT.Close" "FTR.Close" "UBSI.Close" "HD.Close" "JPM.Close"   
## [6] "XOM.Close" "CVX.Close" "NSANY.Close" "GNBT.Close" "MGM.Close"   
## [11] "TEVA.Close" "HST.Close" "FCAU.Close" "WFC.Close" "WWE.Close"   
## [16] "INO.Close" "QSR.Close" "GRPN.Close" "SCE.PB.Close" "FFIN.Close"   
## [21] "GOOG.Close" "WM.Close" "ONCY.Close" "S.Close" "GM.Close"   
## [26] "F.Close" "ASCCY.Close" "ARWR.Close" "COST.Close" "AAL.Close"   
## [31] "JWN.Close" "CSSEP.Close" "NUS.Close" "AMC.Close" "ADDYY.Close"   
## [36] "KSS.Close" "MSFT.Close" "LUV.Close" "HMC.Close" "PCG.Close"   
## [41] "DLTR.Close" "KGJI.Close" "NKE.Close" "AMZN.Close" "ROST.Close"   
## [46] "TMUS.Close" "WMT.Close" "TJX.Close" "TM.Close" "PBYI.Close"   
## [51] "T.Close" "JNJ.Close" "C.Close" "EPD.Close" "VZ.Close"   
## [56] "HRB.Close" "NFLX.Close" "AAP.Close" "HOFT.Close" "SIG.Close"   
## [61] "SDC.Close" "RRGB.Close" "M.Close" "JBLU.Close" "YELP.Close"

Remove NAs from the data. The colSums(is.na(Close)) isn’t returning the columns with NAs, so this must be done manually.

Close\_noNAs <- Close[,-c(9,13,17,18,25,27,32,34,46,50,61,65)]  
Volume\_noNAs <- Volume[,-c(9,13,17,18,25,27,32,34,46,50,61,65)]  
  
Close\_noNAs$SCE.PB.Close <- as.numeric(Close\_noNAs$SCE.PB.Close)  
Volume\_noNAs$SCE.PB.Volume <- as.numeric(Volume\_noNAs$SCE.PB.Volume)

Add in a value of the portfolio column for each day’s closing price of all stock that don’t have NAs.

Close\_noNAs$DailyValue <- rowSums(Close\_noNAs,na.rm=TRUE)

Add in a daily change column of the portfolio closing prices.

dayVal <- as.data.frame(Close\_noNAs$DailyValue)  
colnames(dayVal) <- 'previousDayValue'  
zero <- as.data.frame(as.numeric(dayVal$previousDayValue[1]))  
colnames(zero) <- 'previousDayValue'  
prevDay <- rbind(zero,dayVal)  
Close\_noNAs$prevDay <- prevDay[1:3303,1]  
dailyChange <- as.data.frame(Close\_noNAs$DailyValue-Close\_noNAs$prevDay)  
colnames(dailyChange) <- 'dailyValueChange'  
  
Close1 <- cbind(Close\_noNAs,dailyChange)

Add a column that gives the return in dollars on initial dollars invested.

Close1$ROI\_dollars <- Close1$DailyValue-Close1$DailyValue[1]

Add some date fields to look at the values by date, day of the week, month, and year in analyzing this data.

Close1$Date <- as.Date.character(row.names(Close1))

Close1$DayOfWeek <- weekdays(as.Date(Close1$Date))

month <- month(as.Date(Close1$Date))  
Month <- month.abb[month]  
Close1$Month <- Month

Add in the year of the Date column.

Year <- year(as.Date(Close1$Date))  
  
Close1$Year <- Year  
  
Close1$MonthYear <- paste(Close1$Month, Close1$Year, sep='-')  
Close1$MonthYear <- as.factor(Close1$MonthYear)

Add in some [unemployment](https://data.bls.gov/pdq/SurveyOutputServlet) information as a column to see how the portfolio is doing by date.

ue <- read.delim('BLS\_unemploymentRates2007-2020.txt', sep=',',header=TRUE,   
 na.strings=c('',' '))  
UE <- ue[,-14]#remove the empty 'Annual' column

Use tidyr to gather the month fields with their respective unemployment rates per month.

gatherMonths <- gather(UE, 'UE\_Month', 'UE\_monthlyRate',2:13)  
  
gatherMonths$MonthYear <- paste(gatherMonths$UE\_Month, gatherMonths$Year, sep='-')  
gatherMonths$MonthYear <- as.factor(gatherMonths$MonthYear)

UE2 <- gatherMonths[,3:4]  
Close2 <- merge(Close1, UE2, by.x='MonthYear', by.y='MonthYear')  
row.names(Close2) <- Close2$Date

write.csv(Close2, 'ROI\_UE\_2007\_2020.csv', row.names=FALSE)

Lets add in the volume of trades per day from the Volume\_noNAs data set. But lets add in some fields for total portfolio trades per day,

Volume1 <- Volume\_noNAs  
Volume1$DailyVolume <- rowSums(Volume1, na.rm=TRUE)  
  
dayVol <- as.data.frame(Volume1$DailyVolume)  
colnames(dayVol) <- 'previousDayVolume'  
zero <- as.data.frame(as.numeric(dayVol$previousDayVolume[1]))  
colnames(zero) <- 'previousDayVolume'  
prevDay1 <- rbind(zero,dayVol)  
Volume1$prevDayVolume <- prevDay1[1:3303,1]  
  
dailyVolumeChange <- as.data.frame(Volume1$DailyVolume-Volume1$prevDayVolume)  
colnames(dailyVolumeChange) <- 'dailyVolumeChange'  
  
Volume2 <- cbind(Volume1,dailyVolumeChange)  
Volume2$VolumeRatioDaily2Initial <- Volume2$DailyVolume/Volume2$prevDayVolume[1]  
  
Volume2$Date <- as.Date(row.names(Volume2))

stocks <- cbind(Close2, Volume2)  
  
Stocks <- stocks[,c(2:54,64:116,1,55:63,117:120)]  
colnames(Stocks)

## [1] "TGT.Close" "FTR.Close"   
## [3] "UBSI.Close" "HD.Close"   
## [5] "JPM.Close" "XOM.Close"   
## [7] "CVX.Close" "NSANY.Close"   
## [9] "MGM.Close" "TEVA.Close"   
## [11] "HST.Close" "WFC.Close"   
## [13] "WWE.Close" "INO.Close"   
## [15] "SCE.PB.Close" "FFIN.Close"   
## [17] "GOOG.Close" "WM.Close"   
## [19] "ONCY.Close" "S.Close"   
## [21] "F.Close" "ARWR.Close"   
## [23] "COST.Close" "AAL.Close"   
## [25] "JWN.Close" "NUS.Close"   
## [27] "ADDYY.Close" "KSS.Close"   
## [29] "MSFT.Close" "LUV.Close"   
## [31] "HMC.Close" "PCG.Close"   
## [33] "DLTR.Close" "KGJI.Close"   
## [35] "NKE.Close" "AMZN.Close"   
## [37] "ROST.Close" "WMT.Close"   
## [39] "TJX.Close" "TM.Close"   
## [41] "T.Close" "JNJ.Close"   
## [43] "C.Close" "EPD.Close"   
## [45] "VZ.Close" "HRB.Close"   
## [47] "NFLX.Close" "AAP.Close"   
## [49] "HOFT.Close" "SIG.Close"   
## [51] "RRGB.Close" "M.Close"   
## [53] "JBLU.Close" "TGT.Volume"   
## [55] "FTR.Volume" "UBSI.Volume"   
## [57] "HD.Volume" "JPM.Volume"   
## [59] "XOM.Volume" "CVX.Volume"   
## [61] "NSANY.Volume" "MGM.Volume"   
## [63] "TEVA.Volume" "HST.Volume"   
## [65] "WFC.Volume" "WWE.Volume"   
## [67] "INO.Volume" "SCE.PB.Volume"   
## [69] "FFIN.Volume" "GOOG.Volume"   
## [71] "WM.Volume" "ONCY.Volume"   
## [73] "S.Volume" "F.Volume"   
## [75] "ARWR.Volume" "COST.Volume"   
## [77] "AAL.Volume" "JWN.Volume"   
## [79] "NUS.Volume" "ADDYY.Volume"   
## [81] "KSS.Volume" "MSFT.Volume"   
## [83] "LUV.Volume" "HMC.Volume"   
## [85] "PCG.Volume" "DLTR.Volume"   
## [87] "KGJI.Volume" "NKE.Volume"   
## [89] "AMZN.Volume" "ROST.Volume"   
## [91] "WMT.Volume" "TJX.Volume"   
## [93] "TM.Volume" "T.Volume"   
## [95] "JNJ.Volume" "C.Volume"   
## [97] "EPD.Volume" "VZ.Volume"   
## [99] "HRB.Volume" "NFLX.Volume"   
## [101] "AAP.Volume" "HOFT.Volume"   
## [103] "SIG.Volume" "RRGB.Volume"   
## [105] "M.Volume" "JBLU.Volume"   
## [107] "MonthYear" "DailyValue"   
## [109] "prevDay" "dailyValueChange"   
## [111] "ROI\_dollars" "Date"   
## [113] "DayOfWeek" "Month"   
## [115] "Year" "UE\_monthlyRate"   
## [117] "DailyVolume" "prevDayVolume"   
## [119] "dailyVolumeChange" "VolumeRatioDaily2Initial"

Add a value of stock daily to the initial value as a ratio.

Stocks$ValueRatioDaily2Initial <- Stocks$DailyValue/Stocks$DailyValue[1]

Add a field that multiplies the daily value and daily volume ratios compared to the initial value and volume by the unemployment rate.

Stocks$DailyRatios\_X\_UE <- Stocks$ValueRatioDaily2Initial\*Stocks$VolumeRatioDaily2Initial\*Stocks$UE\_monthlyRate

Add an exponential calculation field based on the unemployment rate for rate, and using numeric day of the month for t, and k as the month.

Stocks$dayOfMonth <- day(Stocks$Date)  
dayOfMonth <- day(Stocks$Date)  
ue1 <- Stocks$UE\_monthlyRate  
  
Stocks$poisson <- (exp(-(ue1))\*(ue1)^dayOfMonth)/(factorial(dayOfMonth))

write.csv(Stocks, 'StocksStats.csv', row.names=TRUE)

Make a daily ROI dollars column for each of the stocks in this set.

stocks1 <- Stocks[,1:53]  
colnames(stocks1)

## [1] "TGT.Close" "FTR.Close" "UBSI.Close" "HD.Close" "JPM.Close"   
## [6] "XOM.Close" "CVX.Close" "NSANY.Close" "MGM.Close" "TEVA.Close"   
## [11] "HST.Close" "WFC.Close" "WWE.Close" "INO.Close" "SCE.PB.Close"  
## [16] "FFIN.Close" "GOOG.Close" "WM.Close" "ONCY.Close" "S.Close"   
## [21] "F.Close" "ARWR.Close" "COST.Close" "AAL.Close" "JWN.Close"   
## [26] "NUS.Close" "ADDYY.Close" "KSS.Close" "MSFT.Close" "LUV.Close"   
## [31] "HMC.Close" "PCG.Close" "DLTR.Close" "KGJI.Close" "NKE.Close"   
## [36] "AMZN.Close" "ROST.Close" "WMT.Close" "TJX.Close" "TM.Close"   
## [41] "T.Close" "JNJ.Close" "C.Close" "EPD.Close" "VZ.Close"   
## [46] "HRB.Close" "NFLX.Close" "AAP.Close" "HOFT.Close" "SIG.Close"   
## [51] "RRGB.Close" "M.Close" "JBLU.Close"

stocks1$TGT\_ROI\_dollars <- stocks1$TGT.Close-stocks1$TGT.Close[1]  
stocks1$FTR\_ROI\_dollars <- stocks1$FTR.Close-stocks1$FTR.Close[1]  
stocks1$UBSI\_ROI\_dollars <- stocks1$UBSI.Close-stocks1$UBSI.Close[1]  
stocks1$HD\_ROI\_dollars <- stocks1$HD.Close-stocks1$HD.Close[1]  
stocks1$JPM\_ROI\_dollars <- stocks1$JPM.Close-stocks1$JPM.Close[1]  
  
stocks1$XOM\_ROI\_dollars <- stocks1$XOM.Close-stocks1$XOM.Close[1]  
stocks1$CVX\_ROI\_dollars <- stocks1$CVX.Close-stocks1$CVX.Close[1]  
stocks1$NSANY\_ROI\_dollars <- stocks1$NSANY.Close-stocks1$NSANY.Close[1]  
stocks1$MGM\_ROI\_dollars <- stocks1$MGM.Close-stocks1$MGM.Close[1]  
stocks1$TEVA\_ROI\_dollars <- stocks1$TEVA.Close-stocks1$TEVA.Close[1]  
  
stocks1$HST\_ROI\_dollars <- stocks1$HST.Close-stocks1$HST.Close[1]  
stocks1$WFC\_ROI\_dollars <- stocks1$WFC.Close-stocks1$WFC.Close[1]  
stocks1$WWE\_ROI\_dollars <- stocks1$WWE.Close-stocks1$WWE.Close[1]  
stocks1$INO\_ROI\_dollars <- stocks1$INO.Close-stocks1$INO.Close[1]  
stocks1$SCE.PB\_ROI\_dollars <- stocks1$SCE.PB.Close-stocks1$SCE.PB.Close[1]  
  
stocks1$FFIN\_ROI\_dollars <- stocks1$FFIN.Close-stocks1$FFIN.Close[1]  
stocks1$GOOG\_ROI\_dollars <- stocks1$GOOG.Close-stocks1$GOOG.Close[1]  
stocks1$WM\_ROI\_dollars <- stocks1$WM.Close-stocks1$WM.Close[1]  
stocks1$ONCY\_ROI\_dollars <- stocks1$ONCY.Close-stocks1$ONCY.Close[1]  
stocks1$S\_ROI\_dollars <- stocks1$S.Close-stocks1$S.Close[1]  
  
stocks1$F\_ROI\_dollars <- stocks1$F.Close-stocks1$F.Close[1]  
stocks1$ARWR\_ROI\_dollars <- stocks1$ARWR.Close-stocks1$ARWR.Close[1]  
stocks1$COST\_ROI\_dollars <- stocks1$COST.Close-stocks1$COST.Close[1]  
stocks1$AAL\_ROI\_dollars <- stocks1$AAL.Close-stocks1$AAL.Close[1]  
stocks1$JWN\_ROI\_dollars <- stocks1$JWN.Close-stocks1$JWN.Close[1]  
  
stocks1$NUS\_ROI\_dollars <- stocks1$NUS.Close-stocks1$NUS.Close[1]  
stocks1$HMC\_ROI\_dollars <- stocks1$HMC.Close-stocks1$HMC.Close[1]  
stocks1$AMZN\_ROI\_dollars <- stocks1$AMZN.Close-stocks1$AMZN.Close[1]  
stocks1$T\_ROI\_dollars <- stocks1$T.Close-stocks1$T.Close[1]  
stocks1$HRB\_ROI\_dollars <- stocks1$HRB.Close-stocks1$HRB.Close[1]  
stocks1$RRGB\_ROI\_dollars <- stocks1$RRGB.Close-stocks1$RRGB.Close[1]  
  
stocks1$ADDYY\_ROI\_dollars <- stocks1$ADDYY.Close-stocks1$ADDYY.Close[1]  
stocks1$PCG\_ROI\_dollars <- stocks1$PCG.Close-stocks1$PCG.Close[1]  
stocks1$ROST\_ROI\_dollars <- stocks1$ROST.Close-stocks1$ROST.Close[1]  
stocks1$JNJ\_ROI\_dollars <- stocks1$JNJ.Close-stocks1$JNJ.Close[1]  
stocks1$NFLX\_ROI\_dollars <- stocks1$NFLX.Close-stocks1$NFLX.Close[1]  
stocks1$M\_ROI\_dollars <- stocks1$M.Close-stocks1$M.Close[1]  
  
stocks1$KSS\_ROI\_dollars <- stocks1$KSS.Close-stocks1$KSS.Close[1]  
stocks1$DLTR\_ROI\_dollars <- stocks1$DLTR.Close-stocks1$DLTR.Close[1]  
stocks1$WMT\_ROI\_dollars <- stocks1$WMT.Close-stocks1$WMT.Close[1]  
stocks1$C\_ROI\_dollars <- stocks1$C.Close-stocks1$C.Close[1]  
stocks1$AAP\_ROI\_dollars <- stocks1$AAP.Close-stocks1$AAP.Close[1]  
stocks1$JBLU\_ROI\_dollars <- stocks1$JBLU.Close-stocks1$JBLU.Close[1]  
  
stocks1$MSFT\_ROI\_dollars <- stocks1$MSFT.Close-stocks1$MSFT.Close[1]  
stocks1$KGJI\_ROI\_dollars <- stocks1$KGJI.Close-stocks1$KGJI.Close[1]  
stocks1$EPD\_ROI\_dollars <- stocks1$EPD.Close-stocks1$EPD.Close[1]  
stocks1$TJX\_ROI\_dollars <- stocks1$TJX.Close-stocks1$TJX.Close[1]  
stocks1$HOFT\_ROI\_dollars <- stocks1$HOFT.Close-stocks1$HOFT.Close[1]  
  
stocks1$LUV\_ROI\_dollars <- stocks1$LUV.Close-stocks1$LUV.Close[1]  
stocks1$NKE\_ROI\_dollars <- stocks1$NKE.Close-stocks1$NKE.Close[1]  
stocks1$TM\_ROI\_dollars <- stocks1$TM.Close-stocks1$TM.Close[1]  
stocks1$VZ\_ROI\_dollars <- stocks1$VZ.Close-stocks1$VZ.Close[1]  
stocks1$SIG\_ROI\_dollars <- stocks1$SIG.Close-stocks1$SIG.Close[1]

These are the values of the stock the previous day that will be subtracted from each day to get the daily change from the day before in dollars.

TGTa <- c(0,stocks1$TGT.Close[1:3302])  
FTRa <- c(0, stocks1$FTR.Close[1:3302])  
UBSIa <- c(0,stocks1$UBSI.Close[1:3302])  
HDa <- c(0,stocks1$HD.Close[1:3302])  
JPMa <- c(0,stocks1$JPM.Close[1:3302])  
XOMa <- c(0,stocks1$XOM.Close[1:3302])  
CVXa <- c(0,stocks1$CVX.Close[1:3302])  
NSANYa <- c(0,stocks1$NSANY.Close[1:3302])  
MGMa <- c(0,stocks1$MGM.Close[1:3302])  
TEVAa <- c(0, stocks1$TEVA.Close[1:3302])  
HSTa <- c(0, stocks1$HST.Close[1:3302])  
WFCa <- c(0, stocks1$WFC.Close[1:3302])  
WWEa <- c(0, stocks1$WWE.Close[1:3302])  
INOa <- c(0,stocks1$INO.Close[1:3302])  
SCEa <- c(0,stocks1$SCE.PB.Close[1:3302])  
FFINa <- c(0,stocks1$FFIN.Close[1:3302])  
GOOGa <- c(0,stocks1$GOOG.Close[1:3302])  
WMa <- c(0,stocks1$WM.Close[1:3302])  
ONCYa <- c(0,stocks1$ONCY.Close[1:3302])  
Sa <- c(0,stocks1$S.Close[1:3302])  
Fa <- c(0,stocks1$F.Close[1:3302])  
ARWRa <- c(0,stocks1$ARWR.Close[1:3302])  
COSTa <- c(0,stocks1$COST.Close[1:3302])  
AALa <- c(0,stocks1$AAL.Close[1:3302])  
JWNa <- c(0,stocks1$JWN.Close[1:3302])  
NUSa <- c(0,stocks1$NUS.Close[1:3302])  
ADDYYa <- c(0,stocks1$ADDYY.Close[1:3302])  
KSSa <- c(0,stocks1$KSS.Close[1:3302])  
MSFTa <- c(0,stocks1$MSFT.Close[1:3302])  
LUVa <- c(0,stocks1$LUV.Close[1:3302])  
HMCa <- c(0,stocks1$HMC.Close[1:3302])  
PCGa <- c(0,stocks1$PCG.Close[1:3302])  
DLTRa <- c(0,stocks1$DLTR.Close[1:3302])  
KGJIa <- c(0,stocks1$KGJI.Close[1:3302])  
NKEa <- c(0,stocks1$NKE.Close[1:3302])  
AMZNa <- c(0,stocks1$AMZN.Close[1:3302])  
ROSTa <- c(0,stocks1$ROST.Close[1:3302])  
WMTa <- c(0,stocks1$WMT.Close[1:3302])  
TJXa <- c(0,stocks1$TJX.Close[1:3302])  
TMa <- c(0,stocks1$TM.Close[1:3302])  
Ta <- c(0,stocks1$T.Close[1:3302])  
JNJa <- c(0,stocks1$JNJ.Close[1:3302])  
Ca <- c(0,stocks1$C.Close[1:3302])  
EPDa <- c(0,stocks1$EPD.Close[1:3302])  
VZa <- c(0,stocks1$VZ.Close[1:3302])  
HRBa <- c(0,stocks1$HRB.Close[1:3302])  
NFLXa <- c(0,stocks1$NFLX.Close[1:3302])  
AAPa <- c(0,stocks1$AAP.Close[1:3302])  
HOFTa <- c(0,stocks1$HOFT.Close[1:3302])  
SIGa <- c(0,stocks1$SIG.Close[1:3302])  
RRGBa <- c(0,stocks1$RRGB.Close[1:3302])  
Ma <- c(0,stocks1$M.Close[1:3302])  
JBLUa <- c(0,stocks1$JBLU.Close[1:3302])

This creates the DailyChange per stock columns.

stocks1$TGT\_dailyChange <- stocks1$TGT.Close-TGTa  
stocks1$FTR\_dailyChange <- stocks1$FTR.Close-FTRa  
stocks1$UBSI\_dailyChange <- stocks1$UBSI.Close-UBSIa  
stocks1$HD\_dailyChange <- stocks1$HD.Close-HDa  
stocks1$JPM\_dailyChange <- stocks1$JPM.Close-JPMa  
  
stocks1$XOM\_dailyChange <- stocks1$XOM.Close-XOMa  
stocks1$CVX\_dailyChange <- stocks1$CVX.Close-CVXa  
stocks1$NSANY\_dailyChange <- stocks1$NSANY.Close-NSANYa  
stocks1$MGM\_dailyChange <- stocks1$MGM.Close-MGMa  
stocks1$TEVA\_dailyChange <- stocks1$TEVA.Close-TEVAa  
  
stocks1$HST\_dailyChange <- stocks1$HST.Close-HSTa  
stocks1$WFC\_dailyChange <- stocks1$WFC.Close-WFCa  
stocks1$WWE\_dailyChange <- stocks1$WWE.Close-WWEa  
stocks1$INO\_dailyChange <- stocks1$INO.Close-INOa  
stocks1$SCE.PB\_dailyChange <- stocks1$SCE.PB.Close-SCEa  
  
stocks1$FFIN\_dailyChange <- stocks1$FFIN.Close-FFINa  
stocks1$GOOG\_dailyChange <- stocks1$GOOG.Close-GOOGa  
stocks1$WM\_dailyChange <- stocks1$WM.Close-WMa  
stocks1$ONCY\_dailyChange <- stocks1$ONCY.Close-ONCYa  
stocks1$S\_dailyChange <- stocks1$S.Close-Sa  
  
stocks1$F\_dailyChange <- stocks1$F.Close-Fa  
stocks1$ARWR\_dailyChange <- stocks1$ARWR.Close-ARWRa  
stocks1$COST\_dailyChange <- stocks1$COST.Close-COSTa  
stocks1$AAL\_dailyChange <- stocks1$AAL.Close-AALa  
stocks1$JWN\_dailyChange <- stocks1$JWN.Close-JWNa  
  
stocks1$NUS\_dailyChange <- stocks1$NUS.Close-NUSa  
stocks1$HMC\_dailyChange <- stocks1$HMC.Close-HMCa  
stocks1$AMZN\_dailyChange <- stocks1$AMZN.Close-AMZNa  
stocks1$T\_dailyChange <- stocks1$T.Close-Ta  
stocks1$HRB\_dailyChange <- stocks1$HRB.Close-HRBa  
stocks1$RRGB\_dailyChange <- stocks1$RRGB.Close-RRGBa  
  
stocks1$ADDYY\_dailyChange <- stocks1$ADDYY.Close-ADDYYa  
stocks1$PCG\_dailyChange <- stocks1$PCG.Close-PCGa  
stocks1$ROST\_dailyChange <- stocks1$ROST.Close-ROSTa  
stocks1$JNJ\_dailyChange <- stocks1$JNJ.Close-JNJa  
stocks1$NFLX\_dailyChange <- stocks1$NFLX.Close-NFLXa  
stocks1$M\_dailyChange <- stocks1$M.Close-Ma  
  
stocks1$KSS\_dailyChange <- stocks1$KSS.Close-KSSa  
stocks1$DLTR\_dailyChange <- stocks1$DLTR.Close-DLTRa  
stocks1$WMT\_dailyChange <- stocks1$WMT.Close-WMTa  
stocks1$C\_dailyChange <- stocks1$C.Close-Ca  
stocks1$AAP\_dailyChange <- stocks1$AAP.Close-AAPa  
stocks1$JBLU\_dailyChange <- stocks1$JBLU.Close-JBLUa  
  
stocks1$MSFT\_dailyChange <- stocks1$MSFT.Close-MSFTa  
stocks1$KGJI\_dailyChange <- stocks1$KGJI.Close-KGJIa  
stocks1$EPD\_dailyChange <- stocks1$EPD.Close-EPDa  
stocks1$TJX\_dailyChange <- stocks1$TJX.Close-TJXa  
stocks1$HOFT\_dailyChange <- stocks1$HOFT.Close-HOFTa  
  
stocks1$LUV\_dailyChange <- stocks1$LUV.Close-LUVa  
stocks1$NKE\_dailyChange <- stocks1$NKE.Close-NKEa  
stocks1$TM\_dailyChange <- stocks1$TM.Close-TMa  
stocks1$VZ\_dailyChange <- stocks1$VZ.Close-VZa  
stocks1$SIG\_dailyChange <- stocks1$SIG.Close-SIGa

Combine the stocks1 stats of ROI and daily change in dollars per stock to the stocks stats data table.

stocks2 <- stocks1[,-c(1:53)]  
StocksSTATS <- cbind(Stocks,stocks2)

write.csv(StocksSTATS, 'STOCKS\_STATS.csv', row.names=TRUE)

All the columns we now have are:

colnames(StocksSTATS)

## [1] "TGT.Close" "FTR.Close"   
## [3] "UBSI.Close" "HD.Close"   
## [5] "JPM.Close" "XOM.Close"   
## [7] "CVX.Close" "NSANY.Close"   
## [9] "MGM.Close" "TEVA.Close"   
## [11] "HST.Close" "WFC.Close"   
## [13] "WWE.Close" "INO.Close"   
## [15] "SCE.PB.Close" "FFIN.Close"   
## [17] "GOOG.Close" "WM.Close"   
## [19] "ONCY.Close" "S.Close"   
## [21] "F.Close" "ARWR.Close"   
## [23] "COST.Close" "AAL.Close"   
## [25] "JWN.Close" "NUS.Close"   
## [27] "ADDYY.Close" "KSS.Close"   
## [29] "MSFT.Close" "LUV.Close"   
## [31] "HMC.Close" "PCG.Close"   
## [33] "DLTR.Close" "KGJI.Close"   
## [35] "NKE.Close" "AMZN.Close"   
## [37] "ROST.Close" "WMT.Close"   
## [39] "TJX.Close" "TM.Close"   
## [41] "T.Close" "JNJ.Close"   
## [43] "C.Close" "EPD.Close"   
## [45] "VZ.Close" "HRB.Close"   
## [47] "NFLX.Close" "AAP.Close"   
## [49] "HOFT.Close" "SIG.Close"   
## [51] "RRGB.Close" "M.Close"   
## [53] "JBLU.Close" "TGT.Volume"   
## [55] "FTR.Volume" "UBSI.Volume"   
## [57] "HD.Volume" "JPM.Volume"   
## [59] "XOM.Volume" "CVX.Volume"   
## [61] "NSANY.Volume" "MGM.Volume"   
## [63] "TEVA.Volume" "HST.Volume"   
## [65] "WFC.Volume" "WWE.Volume"   
## [67] "INO.Volume" "SCE.PB.Volume"   
## [69] "FFIN.Volume" "GOOG.Volume"   
## [71] "WM.Volume" "ONCY.Volume"   
## [73] "S.Volume" "F.Volume"   
## [75] "ARWR.Volume" "COST.Volume"   
## [77] "AAL.Volume" "JWN.Volume"   
## [79] "NUS.Volume" "ADDYY.Volume"   
## [81] "KSS.Volume" "MSFT.Volume"   
## [83] "LUV.Volume" "HMC.Volume"   
## [85] "PCG.Volume" "DLTR.Volume"   
## [87] "KGJI.Volume" "NKE.Volume"   
## [89] "AMZN.Volume" "ROST.Volume"   
## [91] "WMT.Volume" "TJX.Volume"   
## [93] "TM.Volume" "T.Volume"   
## [95] "JNJ.Volume" "C.Volume"   
## [97] "EPD.Volume" "VZ.Volume"   
## [99] "HRB.Volume" "NFLX.Volume"   
## [101] "AAP.Volume" "HOFT.Volume"   
## [103] "SIG.Volume" "RRGB.Volume"   
## [105] "M.Volume" "JBLU.Volume"   
## [107] "MonthYear" "DailyValue"   
## [109] "prevDay" "dailyValueChange"   
## [111] "ROI\_dollars" "Date"   
## [113] "DayOfWeek" "Month"   
## [115] "Year" "UE\_monthlyRate"   
## [117] "DailyVolume" "prevDayVolume"   
## [119] "dailyVolumeChange" "VolumeRatioDaily2Initial"  
## [121] "ValueRatioDaily2Initial" "DailyRatios\_X\_UE"   
## [123] "dayOfMonth" "poisson"   
## [125] "TGT\_ROI\_dollars" "FTR\_ROI\_dollars"   
## [127] "UBSI\_ROI\_dollars" "HD\_ROI\_dollars"   
## [129] "JPM\_ROI\_dollars" "XOM\_ROI\_dollars"   
## [131] "CVX\_ROI\_dollars" "NSANY\_ROI\_dollars"   
## [133] "MGM\_ROI\_dollars" "TEVA\_ROI\_dollars"   
## [135] "HST\_ROI\_dollars" "WFC\_ROI\_dollars"   
## [137] "WWE\_ROI\_dollars" "INO\_ROI\_dollars"   
## [139] "SCE.PB\_ROI\_dollars" "FFIN\_ROI\_dollars"   
## [141] "GOOG\_ROI\_dollars" "WM\_ROI\_dollars"   
## [143] "ONCY\_ROI\_dollars" "S\_ROI\_dollars"   
## [145] "F\_ROI\_dollars" "ARWR\_ROI\_dollars"   
## [147] "COST\_ROI\_dollars" "AAL\_ROI\_dollars"   
## [149] "JWN\_ROI\_dollars" "NUS\_ROI\_dollars"   
## [151] "HMC\_ROI\_dollars" "AMZN\_ROI\_dollars"   
## [153] "T\_ROI\_dollars" "HRB\_ROI\_dollars"   
## [155] "RRGB\_ROI\_dollars" "ADDYY\_ROI\_dollars"   
## [157] "PCG\_ROI\_dollars" "ROST\_ROI\_dollars"   
## [159] "JNJ\_ROI\_dollars" "NFLX\_ROI\_dollars"   
## [161] "M\_ROI\_dollars" "KSS\_ROI\_dollars"   
## [163] "DLTR\_ROI\_dollars" "WMT\_ROI\_dollars"   
## [165] "C\_ROI\_dollars" "AAP\_ROI\_dollars"   
## [167] "JBLU\_ROI\_dollars" "MSFT\_ROI\_dollars"   
## [169] "KGJI\_ROI\_dollars" "EPD\_ROI\_dollars"   
## [171] "TJX\_ROI\_dollars" "HOFT\_ROI\_dollars"   
## [173] "LUV\_ROI\_dollars" "NKE\_ROI\_dollars"   
## [175] "TM\_ROI\_dollars" "VZ\_ROI\_dollars"   
## [177] "SIG\_ROI\_dollars" "TGT\_dailyChange"   
## [179] "FTR\_dailyChange" "UBSI\_dailyChange"   
## [181] "HD\_dailyChange" "JPM\_dailyChange"   
## [183] "XOM\_dailyChange" "CVX\_dailyChange"   
## [185] "NSANY\_dailyChange" "MGM\_dailyChange"   
## [187] "TEVA\_dailyChange" "HST\_dailyChange"   
## [189] "WFC\_dailyChange" "WWE\_dailyChange"   
## [191] "INO\_dailyChange" "SCE.PB\_dailyChange"   
## [193] "FFIN\_dailyChange" "GOOG\_dailyChange"   
## [195] "WM\_dailyChange" "ONCY\_dailyChange"   
## [197] "S\_dailyChange" "F\_dailyChange"   
## [199] "ARWR\_dailyChange" "COST\_dailyChange"   
## [201] "AAL\_dailyChange" "JWN\_dailyChange"   
## [203] "NUS\_dailyChange" "HMC\_dailyChange"   
## [205] "AMZN\_dailyChange" "T\_dailyChange"   
## [207] "HRB\_dailyChange" "RRGB\_dailyChange"   
## [209] "ADDYY\_dailyChange" "PCG\_dailyChange"   
## [211] "ROST\_dailyChange" "JNJ\_dailyChange"   
## [213] "NFLX\_dailyChange" "M\_dailyChange"   
## [215] "KSS\_dailyChange" "DLTR\_dailyChange"   
## [217] "WMT\_dailyChange" "C\_dailyChange"   
## [219] "AAP\_dailyChange" "JBLU\_dailyChange"   
## [221] "MSFT\_dailyChange" "KGJI\_dailyChange"   
## [223] "EPD\_dailyChange" "TJX\_dailyChange"   
## [225] "HOFT\_dailyChange" "LUV\_dailyChange"   
## [227] "NKE\_dailyChange" "TM\_dailyChange"   
## [229] "VZ\_dailyChange" "SIG\_dailyChange"

Lets us pick one stock, look at the stats we added for that stock and then pull out some googled articles of that stock as a company in the news since 2007 till today’s date of Feb. 18, 2020 to compare the sentiments on the company with words that we will count the number of times the company is in the news, the comments by readers, zoom in on the dates of those articles, and see how the company behaved. Lets choose the highest ROI in dollars out of our stocks and compare it to the lowest ROI in dollars. Lets also use the poisson formula that chose the day of the month, because some people might want to buy stocks on pay day around the 1st or 15th for most, or also some on every Friday or every other Friday which would also use the day of the week.

m <- StocksSTATS[order(StocksSTATS$Date, decreasing=FALSE)[3303], 124:176]  
t <- as.data.frame(t(m))  
colnames(t) <- row.names(m)  
t$StockROI <- row.names(t)  
  
Troi <- t[order(t$'2020-02-14', decreasing=TRUE),]  
  
mostLeast <- rbind(head(Troi,3),tail(Troi,3))  
mostLeast <- na.omit(mostLeast)  
mostLeast

## 2020-02-14 StockROI  
## AMZN\_ROI\_dollars 2094.450 AMZN\_ROI\_dollars  
## GOOG\_ROI\_dollars 1292.331 GOOG\_ROI\_dollars  
## SCE.PB\_ROI\_dollars 679.000 SCE.PB\_ROI\_dollars  
## FTR\_ROI\_dollars -225.180 FTR\_ROI\_dollars  
## C\_ROI\_dollars -431.710 C\_ROI\_dollars

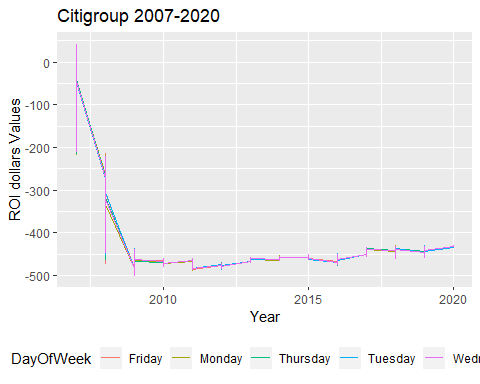
The above table shows the three highest returns on investment and the three lowest since Jan 3, 2007 to Feb 14, 2020. Lets use the lowest stock for now (C is Citigroup bank), because AMZN (Amazon) is always in the news and it would fluctuate a lot I would think, but we could look at the quartiles for each and get the news releases of each date where the stock was in that quartile range, look at the median ROI, the min and max too, and cross referencing with the other stat fields.

amzn <- grep('AMZN', colnames(StocksSTATS))  
c <- grep('^C[.|\_]', colnames(StocksSTATS))  
C\_stock <- StocksSTATS[,c(c,107,112:116,123:124)]  
amzn\_stock <- StocksSTATS[,c(amzn,107,112:116,123:124)]

Citigroup is our C\_stock table and Amazon is our amzn\_stock table. Lets look at the daily ratios of volume and ROI in dollars times the unemployment rate column and the day of the week and day of the year and poisson columns.

ggplot(data = C\_stock, aes(x=Year, y=C\_ROI\_dollars,group=DayOfWeek)) +  
 geom\_line(aes(color=DayOfWeek))+  
 #geom\_point()+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Citigroup 2007-2020')+  
 ylab('ROI dollars Values')

## Warning in pal\_name(palette, type): Unknown palette paired

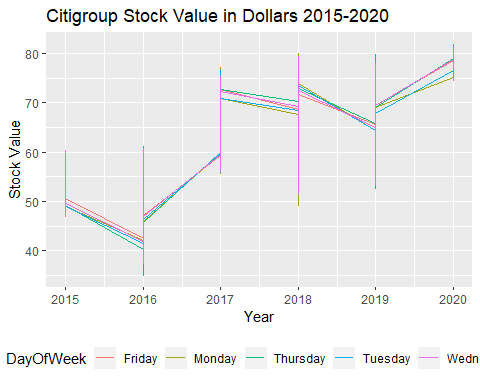


We can see from the plot above that buying Citigroup stock anywhere before 2010, was a bad idea. But we also see that the stock would have been good to buy around 2010-2016, as it overall increased its return on investment in dollars initially invested.

Lets look at the years from 2016-2020 to see this plotted Citigroup stock.

y2015plus <- subset(C\_stock, C\_stock$Year>2014)  
  
ggplot(data = y2015plus, aes(x=Year, y=C.Close,group=DayOfWeek)) +  
 geom\_line(aes(color=DayOfWeek))+  
 #geom\_point()+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Citigroup Stock Value in Dollars 2015-2020')+  
 ylab('Stock Value')

## Warning in pal\_name(palette, type): Unknown palette paired

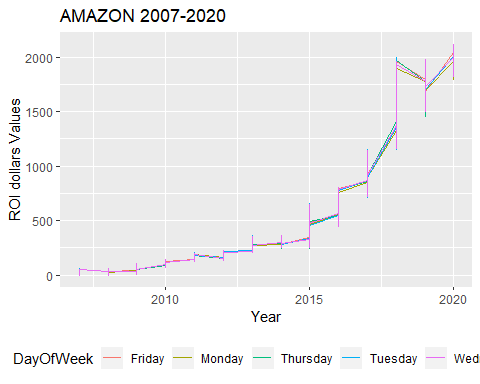


We see from the above plot that Citigroup was good to buy at the start of 2016 or 2019 if you want to see an increase all year long, but in 2017-2018 it decreased.Overall, if investing since 2016, the stock increased from the high $40 to the mid-high $70 range. This would be good to cross reference with unemployment rates and the news articles online text mined for public sentiment on Citigroup.

Lets look at amazon for the same quick plotted analysis as done with Citigroup.

ggplot(data = amzn\_stock, aes(x=Year, y=AMZN\_ROI\_dollars,group=DayOfWeek)) +  
 geom\_line(aes(color=DayOfWeek))+  
 #geom\_point()+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('AMAZON 2007-2020')+  
 ylab('ROI dollars Values')

## Warning in pal\_name(palette, type): Unknown palette paired

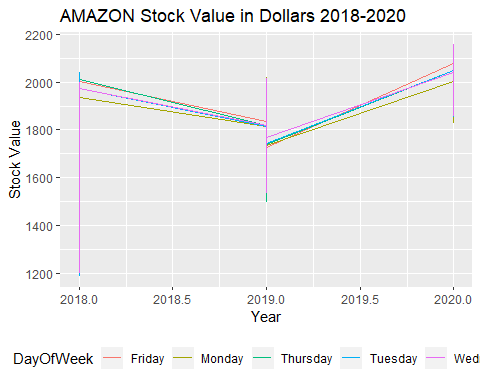


We can see from the plot above that buying AMAZON stock anywhere before 2010, was a great idea. But we also see that the stock would have been good to buy around 2010-2018 or 2019 but not in 2018, as it overall increased its return on investment in dollars initially invested.In 2018, you bought high and it decreased the entire year. This would be great to see what happened in 2018 with the value. So we will.

Lets look at the years from 2018-2020 to see this plotted Citigroup stock.

y2015plus <- subset(amzn\_stock, amzn\_stock$Year>2017)  
  
ggplot(data = y2015plus, aes(x=Year, y=AMZN.Close,group=DayOfWeek)) +  
 geom\_line(aes(color=DayOfWeek))+  
 #geom\_point()+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('AMAZON Stock Value in Dollars 2018-2020')+  
 ylab('Stock Value')

## Warning in pal\_name(palette, type): Unknown palette paired



The chart above shows how the value in dollars and day of the week from 2018-2020 decreases in 2018 and increases in 2019. If you bought in 2018, you lost money the entire year, but you gained it back in 2019 plus some additional earnings.

Lets group by the day of the month in this time series of the Citigroup stock and get the median value for the volumne of stocks traded for Citigroup by days 1-31 of the month.

v1 <- as.vector(colnames(C\_stock)[2])  
poisson\_Citi <- C\_stock %>% group\_by(dayOfMonth) %>% summarise\_at(vars(v1), median,  
 na.rm=T)  
poisson\_Citi <- as.data.frame(poisson\_Citi)  
colnames(poisson\_Citi)[2] <- 'Citi\_Median\_Volume'  
poisson\_Citi <- poisson\_Citi[order(poisson\_Citi$Citi\_Median\_Volume, decreasing=T),]  
headTail\_Citi\_volume <- rbind(head(poisson\_Citi,3), tail(poisson\_Citi,3))  
headTail\_Citi\_volume

## dayOfMonth Citi\_Median\_Volume  
## 16 16 22388100  
## 31 31 22302200  
## 3 3 21522100  
## 25 25 17960700  
## 20 20 17548500  
## 2 2 17134600

From the above table we see that the most volume of trades for Citigroup is at the middle and end of the month, and the lowest volume of trades are at the beginning of the new month and the third week of the month.

Lets look at the statistics of citigroup.

summary(C\_stock)

## C.Close C.Volume C\_ROI\_dollars C\_dailyChange   
## Min. : 10.20 Min. : 1005100 Min. :-500.3 Min. :-298.3000   
## 1st Qu.: 41.80 1st Qu.: 13043950 1st Qu.:-468.7 1st Qu.: -0.6750   
## Median : 51.59 Median : 19535800 Median :-458.9 Median : -0.0100   
## Mean : 93.33 Mean : 27022488 Mean :-417.2 Mean : 0.0209   
## 3rd Qu.: 69.58 3rd Qu.: 33314650 3rd Qu.:-440.9 3rd Qu.: 0.6500   
## Max. :552.50 Max. :377263800 Max. : 42.0 Max. : 510.5000   
##   
## MonthYear Date DayOfWeek Month   
## Aug-2007: 23 Min. :2007-01-03 Length:3303 Length:3303   
## Aug-2011: 23 1st Qu.:2010-04-14 Class :character Class :character   
## Aug-2012: 23 Median :2013-07-25 Mode :character Mode :character   
## Aug-2016: 23 Mean :2013-07-24   
## Aug-2017: 23 3rd Qu.:2016-11-01   
## Aug-2018: 23 Max. :2020-02-14   
## (Other) :3165   
## Year UE\_monthlyRate dayOfMonth poisson   
## Min. :2007 Min. : 3.500 Min. : 1.00 Min. :0.000000   
## 1st Qu.:2010 1st Qu.: 4.600 1st Qu.: 8.00 1st Qu.:0.000000   
## Median :2013 Median : 5.600 Median :16.00 Median :0.000563   
## Mean :2013 Mean : 6.282 Mean :15.72 Mean :0.032501   
## 3rd Qu.:2016 3rd Qu.: 8.200 3rd Qu.:23.00 3rd Qu.:0.047937   
## Max. :2020 Max. :10.000 Max. :31.00 Max. :0.215785   
## NA's :10 NA's :10

From the above summary statistics of Citigroup, we see the min, quantiles, median, mean, and max numeric values as well as length and class type for the non-numeric features of this data set.

Some interesting insights into the above table are that considering an initial investment of 510 USD, the return on the initial investment in dollars is almost the entire amount invested but not quite. Definitely about 80% from the quantile and statistics on the ROI column.

The daily changes fluctuated from a loss of 298 USD in one day to a profit of 510 USD on another day. These are good indicators of where to look on these days, to see if the public sentiment on these dates for Citigroup would indicate more people getting rid of their Citi stock or buying up more of it.

Also, the max and min volume of stock is much more and less respectively than the median volume of trades for this Citigroup stock. These dates for information would also be an interesting place to start to find a pattern with buying/selling stock and combining web scraped text from news articles and comments about Citigroup on those dates.

First, we should grab those points of interest in the data and create a table to compare these values.

C\_stock\_minmaxValueChanges <- subset(C\_stock,  
 C\_stock$C\_dailyChange==min(C\_stock$C\_dailyChange) |  
 C\_stock$C\_dailyChange==max(C\_stock$C\_dailyChange) |  
 C\_stock$C.Volume==min(C\_stock$C.Volume) |  
 C\_stock$C.Volume==max(C\_stock$C.Volume))  
C\_stock\_minmaxValueChanges

## C.Close C.Volume C\_ROI\_dollars C\_dailyChange MonthYear Date  
## 2007-04-02 510.50 2282100 0.00 510.500000 Apr-2007 2007-04-02  
## 2013-04-02 44.11 1005100 -466.39 0.320000 Apr-2013 2013-04-02  
## 2015-12-28 52.38 377263800 -458.12 -0.329998 Dec-2015 2015-12-28  
## 2008-06-02 214.60 15302800 -295.90 -298.300018 Jun-2008 2008-06-02  
## DayOfWeek Month Year UE\_monthlyRate dayOfMonth poisson  
## 2007-04-02 Monday Apr 2007 4.5 2 1.124786e-01  
## 2013-04-02 Tuesday Apr 2013 7.6 2 1.445304e-02  
## 2015-12-28 Monday Dec 2015 5.0 28 8.232787e-13  
## 2008-06-02 Monday Jun 2008 5.6 2 5.798250e-02

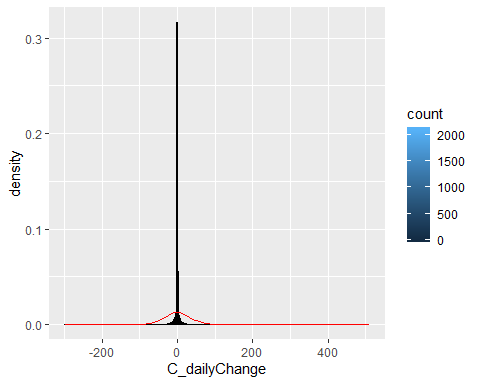
From the above information, Monday is the day of the week with the highest and lowest daily change, as well as the highest volume of trade. Tuesday is the day with the lowest volume of trade. The dates to pull an internet search of news articles about Citigroup to analyze public sentiment on Citi stock are:

* April 2, 2007
* April 2, 2013
* December 28, 2015
* June 2, 2008

This should be interesting to see what type of articles are available on line with a google search of those dates and citigroup.

Lets see if there are any other outlier dates to examine by looking at the standard deviation of the daily change on Citigroup stock. We want to see if there are any days where the stock has a daily change more than or less than this amount times three then times two. Because most values will be within the standard deviation for the Gaussian curve.

gg <- ggplot(C\_stock, aes(x=C\_dailyChange))  
gg <- gg + geom\_histogram(binwidth=2, colour="black",   
 aes(y=..density.., fill=..count..))  
#gg <- gg + scale\_fill\_gradient("Count", low="#DCDCDC", high="#7C7C7C")  
gg <- gg + stat\_function(fun=dnorm,  
 color="red",  
 args=list(mean=mean(C\_stock$C\_dailyChange),   
 sd=sd(C\_stock$C\_dailyChange)))  
  
gg



sdC <- sd(C\_stock$C\_dailyChange)  
out <- sdC\*3  
sdC;out

## [1] 32.05415

## [1] 96.16246

The standard error for the daily change in dollars is 32.05 USD and our threshold to find dates outside this normal range of daily change dollar values is 96.16 USD.

Lets add another column to this data set called threshold3 for those daily change values inside the threshold and those outside the threshold.

C\_stock$Threshold3 <- ifelse(C\_stock$C\_dailyChange < sdC, 'inside','outside')  
  
C\_outer\_SD <- subset(C\_stock, C\_stock$Threshold3=='outside')  
summary(C\_outer\_SD)

## C.Close C.Volume C\_ROI\_dollars C\_dailyChange   
## Min. :206.5 Min. : 2282100 Min. :-304.00 Min. : 40.0   
## 1st Qu.:468.5 1st Qu.:13495700 1st Qu.: -42.00 1st Qu.:397.8   
## Median :510.5 Median :16045900 Median : 0.00 Median :440.8   
## Mean :466.5 Mean :29284854 Mean : -43.99 Mean :393.6   
## 3rd Qu.:542.0 3rd Qu.:33433800 3rd Qu.: 31.50 3rd Qu.:473.7   
## Max. :552.5 Max. :81343800 Max. : 42.00 Max. :510.5   
##   
## MonthYear Date DayOfWeek Month   
## Apr-2007:1 Min. :2007-01-03 Length:13 Length:13   
## Aug-2007:1 1st Qu.:2007-04-02 Class :character Class :character   
## Dec-2007:1 Median :2007-07-02 Mode :character Mode :character   
## Feb-2007:1 Mean :2007-07-22   
## Jan-2007:1 3rd Qu.:2007-10-01   
## Jul-2007:1 Max. :2008-09-19   
## (Other) :7   
## Year UE\_monthlyRate dayOfMonth poisson   
## Min. :2007 Min. :4.400 Min. : 1.000 Min. :1.538e-05   
## 1st Qu.:2007 1st Qu.:4.500 1st Qu.: 1.000 1st Qu.:4.624e-02   
## Median :2007 Median :4.600 Median : 1.000 Median :5.402e-02   
## Mean :2007 Mean :4.731 Mean : 3.077 Mean :7.979e-02   
## 3rd Qu.:2007 3rd Qu.:4.700 3rd Qu.: 3.000 3rd Qu.:1.125e-01   
## Max. :2008 Max. :6.100 Max. :19.000 Max. :1.849e-01   
##   
## Threshold3   
## Length:13   
## Class :character   
## Mode :character   
##   
##   
##   
##

We can see from the above statistics on the subset of Citigroup stock that are outside this threshold that there are 13 dates to select in the range of Jan 2007 through Sep 2008. So we will add those dates to our data set of text scraped news articles on Citigroup.

NLP\_dates\_Citi <- rbind(C\_stock\_minmaxValueChanges, C\_outer\_SD[,-13])  
NLP\_dates\_Citi

## C.Close C.Volume C\_ROI\_dollars C\_dailyChange MonthYear Date  
## 2007-04-02 510.50 2282100 0.000000 510.500000 Apr-2007 2007-04-02  
## 2013-04-02 44.11 1005100 -466.389999 0.320000 Apr-2013 2013-04-02  
## 2015-12-28 52.38 377263800 -458.119999 -0.329998 Dec-2015 2015-12-28  
## 2008-06-02 214.60 15302800 -295.899994 -298.300018 Jun-2008 2008-06-02  
## 2007-04-021 510.50 2282100 0.000000 510.500000 Apr-2007 2007-04-02  
## 2007-08-01 468.50 13495700 -42.000000 397.800003 Aug-2007 2007-08-01  
## 2007-12-03 330.60 81343800 -179.899994 266.250008 Dec-2007 2007-12-03  
## 2007-02-01 547.30 80864600 36.799988 467.409989 Feb-2007 2007-02-01  
## 2007-01-03 552.50 43508100 42.000000 473.709999 Jan-2007 2007-01-03  
## 2007-07-02 516.40 32822200 5.900024 441.990020 Jul-2007 2007-07-02  
## 2007-06-01 545.10 23057000 34.599976 473.939972 Jun-2007 2007-06-01  
## 2007-03-01 510.80 8981300 0.299988 440.769989 Mar-2007 2007-03-01  
## 2007-05-01 542.00 13337900 31.500000 479.779999 May-2007 2007-05-01  
## 2007-11-01 385.10 33433800 -125.399994 322.950004 Nov-2007 2007-11-01  
## 2007-10-01 477.20 16045900 -33.299988 402.080009 Oct-2007 2007-10-01  
## 2007-09-04 472.10 15929600 -38.399994 400.240005 Sep-2007 2007-09-04  
## 2008-09-19 206.50 15601100 -304.000000 40.000000 Sep-2008 2008-09-19  
## DayOfWeek Month Year UE\_monthlyRate dayOfMonth poisson  
## 2007-04-02 Monday Apr 2007 4.5 2 1.124786e-01  
## 2013-04-02 Tuesday Apr 2013 7.6 2 1.445304e-02  
## 2015-12-28 Monday Dec 2015 5.0 28 8.232787e-13  
## 2008-06-02 Monday Jun 2008 5.6 2 5.798250e-02  
## 2007-04-021 Monday Apr 2007 4.5 2 1.124786e-01  
## 2007-08-01 Wednesday Aug 2007 4.6 1 4.623844e-02  
## 2007-12-03 Monday Dec 2007 5.0 3 1.403739e-01  
## 2007-02-01 Thursday Feb 2007 4.5 1 4.999048e-02  
## 2007-01-03 Wednesday Jan 2007 4.6 3 1.630676e-01  
## 2007-07-02 Monday Jul 2007 4.7 2 1.004573e-01  
## 2007-06-01 Friday Jun 2007 4.6 1 4.623844e-02  
## 2007-03-01 Thursday Mar 2007 4.4 1 5.402030e-02  
## 2007-05-01 Tuesday May 2007 4.4 1 5.402030e-02  
## 2007-11-01 Thursday Nov 2007 4.7 1 4.274780e-02  
## 2007-10-01 Monday Oct 2007 4.7 1 4.274780e-02  
## 2007-09-04 Tuesday Sep 2007 4.7 4 1.849252e-01  
## 2008-09-19 Friday Sep 2008 6.1 19 1.538069e-05

I am going to pull the data from these dates with the Google Search for the specific date on Citigroup stock, put it in a table with the date, the article title, reference, article content, and the comments if available.

Note: when searching the internet, there were limited articles and [most](https://www.nytimes.com/2008/11/23/business/23citi.html) were about Citi’s involvement in the sub-prime mortgage crisis of 2007-2008, and a [bailout](https://www.reuters.com/article/us-citigroup/citigroup-gets-massive-government-bailout-idUSTRE4AJ45G20081124) of Citigroup by the US. For the month and years of the two dates not in or around 2007-2008, there are only two for April 2013 and December 2015. Where Citi settled a [lawsuit](https://www.reuters.com/article/us-citigroup-settlement/citigroup-settles-shareholder-cdo-lawsuit-for-590-million-idUSBRE87S0UA20120829) for covering up bad mortgage loans in August 2012 and a [person reported](https://ficoforums.myfico.com/t5/Credit-Card-Approvals/Citi-Simplicity-Approved-Woohoooooo/td-p/4388074) on a forum about FICO scores how he was approved for a 4600 USD credit card with Citi. There isn’t enough data to rely on the web for NLP on Citigroup for these time frames.

We could pull based on the keywords: ‘settlement’, ‘bail-out’, ‘sub-prime loans’, but we would only get the obvious negative sentiment for these keywords. A New York Times article posted an article in Dec 2015 about the remodeling that Citigroup was doing to their offices, but the full article would have to be purchased. The fact that they spent money on remodeling could have some public sentiment of either they aren’t distributing their profits to shareholders or they are making enough profits to spend money on remodeling, which is also reported at the end of the year in 2015 to write off for that tax year. Although, I was told by an accountant that some corporations and small businesses have a different tax year and a quick search on Google returned the fiscal year is any consecutive 12-month business cycle that usually ends at the end of each quarter.

We can see that the volume of trades is highest in December 2015 from our dates, but we should compare this to which quantile this number is within for the volume of trades of Citi stock.

summary(C\_stock$C.Volume)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1005100 13043950 19535800 27022488 33314650 377263800

We already know that this is the date that the most trades in stock of Citi occured as it is the reason we added this date to our NLP data set of dates to pull information from the web for. The above will refresh the comparisons of the trade volume to this date.

It looks like public sentiment thinks Citi is going back to its old bail-out days of 2007-2008 and not a trust-worthy stock for their personal portfolios. But they are still around, and the fact that people that have a less than trust-worthy credit profile were given a credit card with a high value could indicate some people also consider that they are building a new demographic of people to invest in by earning the trust of those who have sub-par trust worthiness with credit. And, yet some other investors could also think this is a bad move to make as it depends on those same people realizing their mistakes and not making them again. Which really turns into the reason some stocks are volatile to begin with and possibly a reason to understand Game Theory, a class I dropped in my undergrad college. But nonetheless I am a data scientist with other coventional and non-conventional ways of extracting useful information, and this approach uses my math and analytic skills to fully understand the stock market and certain stocks and trends with public sentiment.

On this highest trade day, the daily change in dollars was still within the standard error by only dropping 0.33 USD. Where the standard error is 32.00 USD.

Of note is whether or not those making these trades are doing so to lower their Capital Gains at the end of the year, because there is a slight loss on it to balance out the portfolio. Also, this is the end of the year, possibly the last trading day of the year as it is. Lets look at all monthYear dates equal to Dec-2015 to see if there are any other dates past Dec 28, 2015.

dec2015 <- subset(C\_stock, C\_stock$MonthYear=='Dec-2015')  
tail(dec2015)

## C.Close C.Volume C\_ROI\_dollars C\_dailyChange MonthYear Date  
## 2015-12-23 52.63 93423000 -457.87 0.620003 Dec-2015 2015-12-23  
## 2015-12-24 52.71 119108100 -457.79 0.079998 Dec-2015 2015-12-24  
## 2015-12-28 52.38 377263800 -458.12 -0.329998 Dec-2015 2015-12-28  
## 2015-12-29 52.98 281369700 -457.52 0.599999 Dec-2015 2015-12-29  
## 2015-12-30 52.30 62625000 -458.20 -0.680001 Dec-2015 2015-12-30  
## 2015-12-31 51.75 49092600 -458.75 -0.549999 Dec-2015 2015-12-31  
## DayOfWeek Month Year UE\_monthlyRate dayOfMonth poisson  
## 2015-12-23 Wednesday Dec 2015 5 23 3.107014e-09  
## 2015-12-24 Thursday Dec 2015 5 24 6.472947e-10  
## 2015-12-28 Monday Dec 2015 5 28 8.232787e-13  
## 2015-12-29 Tuesday Dec 2015 5 29 1.419446e-13  
## 2015-12-30 Wednesday Dec 2015 5 30 2.365743e-14  
## 2015-12-31 Thursday Dec 2015 5 31 3.815715e-15  
## Threshold3  
## 2015-12-23 inside  
## 2015-12-24 inside  
## 2015-12-28 inside  
## 2015-12-29 inside  
## 2015-12-30 inside  
## 2015-12-31 inside

We now know that Dec-28-2015 is not the last trading day of the year, because the 29th through 31st for Tuesday through Thursday are also trading days. There was a fluctuation in dollars earned and lost all under a dollar. Some useful information to add in would be who or where are these trades derived. Are they financial advisors, trust fund managers, independent investors, foreign or national investors, are they hobbyists just playing the stock market on an e-trade, are they educated, experienced, and so on?

To get this information we could first find out how much it costs for a hobbyist to make a trade online from e-trade or similar and whether or not this information is shared on demographics of the stock ownership. We could also look at the American Survey on Census data from the census bureau for numer of financial workers there are and how many people graduated with a BS, MS, or Phd in Finance or Economics. If there is location data on where these stock owners live attach this information gathered to it to make a better inference on this stock and what motivates the trades. Any volunteers?

For now, we will just continue with what we have on hand for Citi. We can answer the question of whether or not, historically there are more trades in December than any other month in our data by grouping by month year and getting the median trades per month and year.

Citi\_trades\_monthYear <- C\_stock %>% group\_by(MonthYear) %>%  
 summarise\_at(vars(colnames(C\_stock[2])), mean)  
Citi\_trades\_monthYear <- Citi\_trades\_monthYear[order(Citi\_trades\_monthYear$C.Volume,decreasing=TRUE),]  
Citi\_trades\_monthYear

## # A tibble: 158 x 2  
## MonthYear C.Volume  
## <fct> <dbl>  
## 1 Dec-2011 102284343.  
## 2 Dec-2012 97253820   
## 3 Feb-2007 94010711.  
## 4 Feb-2008 80151765   
## 5 Dec-2019 79458262.  
## 6 Aug-2019 72849682.  
## 7 Feb-2015 70393405.  
## 8 Dec-2015 67380332.  
## 9 Jan-2010 64943774.  
## 10 Jan-2012 63211745   
## # … with 148 more rows

From the above table ordered from most trades to least trades per month and year by mean number of trades per month, we see that December is in the top 10 month years of high trades in 2011,2012, 2015, and 2019. February has the next highest trades but the years are the same years of the sub-prime mortgage crisis that Citigroup was involved in, but also in 2015. looking at the next top ten months we see that Dec, Jan, and Feb are in the highest mean of the trades per day grouped by month and year. What do we know about Jan and Feb outside of the assumption about December being the last day of the tax year to offset capital gains with capital losses?

Well, I know that being a student, some people get their student loans around winter quarter in January and that many people expecting tax refunds get their refunds in February. We would have to see if there are any other assumptions about these months. But we would be able to ascertain if students receiving an education are investing, and if consumers with tax refunds are using some of that money to invest.There are certainly other assumptions that could be made for why the last month of the year and the first two months of the first quarter are high trade volume days. But for now lets stick with these assumptions.

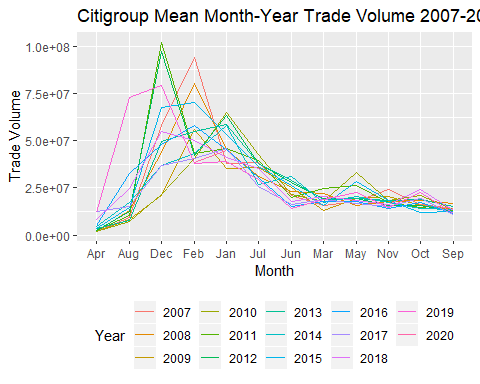
July starts to show up in the following set of ten top month years from 21-30, as the 30th highest trade month year. Jan and Feb are still in the top 40 high volume trade month years, while June shows up three times in the 30-40 top high volume trade month and years. July could also be the start of the third quarter and the remaining balance on student loans made. Lets see where September/October show up in these top ordered volumes. They are near the end of the top trade months.

So, possibly this indicates no ties to student loan payments, but tax refunds could be likely for February being a high trade month. We definitely know December is a top trade day.

Lets plot this data.

Citi\_trades\_monthYear$Month <- gsub('-[0-9]{4}','',Citi\_trades\_monthYear$MonthYear)  
Citi\_trades\_monthYear$Year <- gsub('[a-zA-z]{3}-','',Citi\_trades\_monthYear$MonthYear)  
  
ggplot(data = Citi\_trades\_monthYear, aes(x=Month, y=C.Volume,group=Year)) +  
 geom\_line(aes(color=Year))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Citigroup Mean Month-Year Trade Volume 2007-2020')+  
 ylab('Trade Volume')

## Warning in pal\_name(palette, type): Unknown palette paired



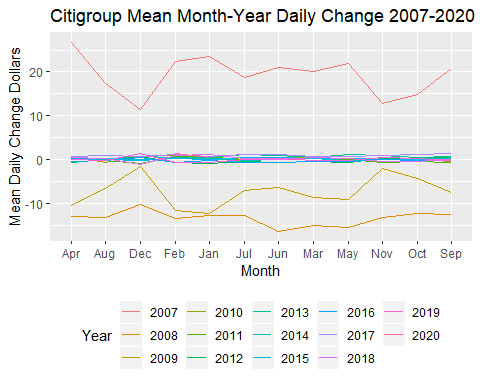
We can see that December is definitely the highest trading month, then February as the next highest, and January as the third highest trading month.

Lets look at the daily change mean values per month, by grouping by MonthYear and taking the mean value of the daily change, order by highest to smallest, and plot.

Citi\_meanMonthly\_dailyChange <- C\_stock %>% group\_by(MonthYear) %>%   
 summarise\_at(vars(as.vector(colnames(C\_stock))[4]), mean)

Citi\_meanMonthly\_dailyChange$Year <-  
 gsub('[a-zA-Z]{3}-','',Citi\_meanMonthly\_dailyChange$MonthYear)  
Citi\_meanMonthly\_dailyChange$Month <-  
 gsub('-[0-9]{4}','',Citi\_meanMonthly\_dailyChange$MonthYear)  
  
  
ggplot(data = Citi\_meanMonthly\_dailyChange, aes(x=Month, y=C\_dailyChange,group=Year)) +  
 geom\_line(aes(color=Year))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Citigroup Mean Month-Year Daily Change 2007-2020')+  
 ylab('Mean Daily Change Dollars')

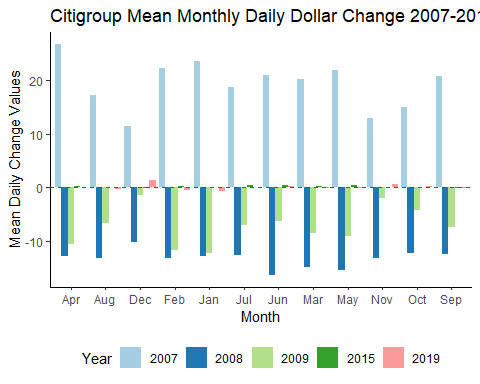
## Warning in pal\_name(palette, type): Unknown palette paired



From the above line chart, it is not obvious what years those years having almost no change are.The year 2007 is at the top with the highest positive mean daily change values fluctuating to around 20 USD per day. While the years 2008 and 2009 have the highest negative mean of daily change values per month with average daily decreases around a daily loss of 5-15 USD.

Lets make a bar chart of 2007, 2008, 2009, 2015, and 2019 of this data on mean daily value changes per month.

y4 <- subset(Citi\_meanMonthly\_dailyChange,  
 Citi\_meanMonthly\_dailyChange$Year==2008 |   
 Citi\_meanMonthly\_dailyChange$Year==2009 |   
 Citi\_meanMonthly\_dailyChange$Year==2007 |  
 Citi\_meanMonthly\_dailyChange$Year==2015 |  
 Citi\_meanMonthly\_dailyChange$Year==2019)  
ggplot(data = y4, aes(x=Month, y=C\_dailyChange,fill=Year)) +  
 geom\_bar(stat='identity', position=position\_dodge())+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette='Paired') +  
 geom\_hline(yintercept=0, linetype="dashed", color = "red")+  
 theme\_classic()+  
 theme(legend.position="bottom")+  
 ggtitle('Citigroup Mean Monthly Daily Dollar Change 2007-2019')+  
 ylab('Mean Daily Change Values')

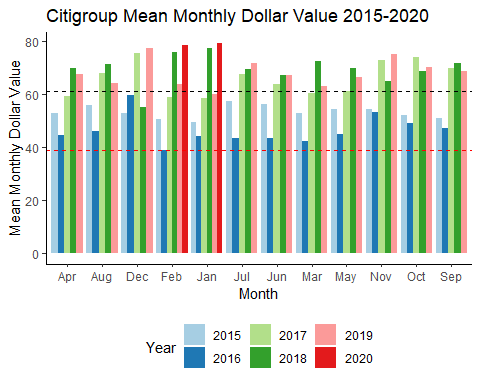


From the above, we can see the Citigroup stock had increases per day in value from the previous day in 2007, but that in 2008 and 2009 those daily increases turned to daily decreases from day to day as the sub-prime loans collapsed that Citigroup held. And in 2015 and 2019 years after Citigroup’s bailout there was a mean monthly daily change value next to nothing as the daily change from day to day fluctuated around zero dollars for the month.

This could mean it is gaining strength and remains as is safe to buy as it increases. But lets look at the years 2015-2019 to see how the value of the Citigroup stock has faired by month year to confirm this assertion just made.

y4value <- subset(C\_stock, C\_stock$Year>2014)  
y4valMY <- y4value %>% group\_by(MonthYear) %>%  
 summarise\_at(vars(as.vector(colnames(y4value)[1])), mean)

y4valMY$Year <- gsub('[a-zA-Z]{3}-','', y4valMY$MonthYear)  
y4valMY$Month <- gsub('-[0-9]{4}','', y4valMY$MonthYear)  
  
ggplot(data = y4valMY, aes(x=Month, y=C.Close,fill=Year)) +  
 geom\_bar(stat='identity', position=position\_dodge())+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette='Paired') +  
 geom\_hline(yintercept=min(y4valMY$C.Close), linetype="dashed", color = "red")+  
 geom\_hline(yintercept=mean(y4valMY$C.Close), linetype="dashed", color = "black")+  
 theme\_classic()+  
 theme(legend.position="bottom")+  
 ggtitle('Citigroup Mean Monthly Dollar Value 2015-2020')+  
 ylab('Mean Monthly Dollar Value')

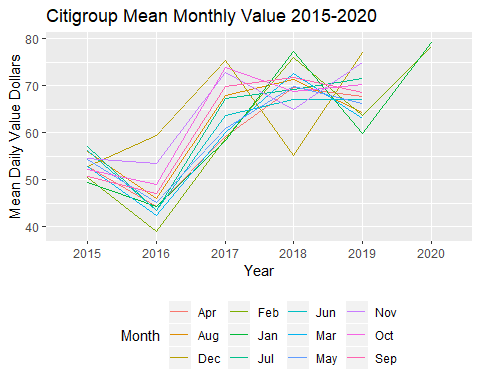


From the above bar chart, we can see that the minimum value is the dashed red line which occured in February 2016. And that every month since 2016 has been above this minimum value. It has almost double from it’s minimum value in January and February 2020.The mean value from 2015-2020 (Jan-Feb) is just above 60 USD which is 1 1/2 times its minimum value.

Lets look at the line chart of this by years 2015-2020.

ggplot(data = y4valMY, aes(x=Year, y=C.Close,group=Month)) +  
 geom\_line(aes(color=Month))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Citigroup Mean Monthly Value 2015-2020')+  
 ylab('Mean Daily Value Dollars')

## Warning in pal\_name(palette, type): Unknown palette paired



The above line chart of the mean monthly dollar value of the Citigroup stock show that all months move the same direction of decreasing in 2015, increasing in 2016, except for in 2017 and 2018 where 3-6 months decreased and 6-9 months increased monthly mean values. The span of 2019 through 2020 can’t be analyzed yet, but both months available for Jan and Feb increased since the year prior. Overall, since 2015 the value has increased from 50-60 USD to between 75-80 USD. This could make it a good stock to have in your portfolio as it has steadily been increasing since it’s historical rough patches of the sub-prime mortgage loan accounts, the public bailout, and the lawsuit settlement payout. But nothing has been in the news about them to discourage investors from dropping this stock from their stock folder.