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$prevDay <- prevDay[1:length(prevDay$previousDayValue)-1,1]  
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  
colnames(Close2)[55:58] <- paste('portfolio',colnames(Close2)[55:58], sep='\_')

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$portfolio\_DailyVolume <- rowSums(Volume1, na.rm=TRUE)  
  
dayVol <- as.data.frame(Volume1$portfolio\_DailyVolume)  
colnames(dayVol) <- 'portfolio\_previousDayVolume'  
zero <- as.data.frame(as.numeric(dayVol$portfolio\_previousDayVolume[1]))  
colnames(zero) <- 'portfolio\_previousDayVolume'  
prevDay1 <- rbind(zero,dayVol)  
Volume1$portfolio\_prevDayVolume <-  
 prevDay1[1:(length(prevDay1$portfolio\_previousDayVolume)-1),1]  
  
dailyVolumeChange <- as.data.frame(Volume1$portfolio\_DailyVolume-Volume1$portfolio\_prevDayVolume)  
colnames(dailyVolumeChange) <- 'portfolio\_dailyVolumeChange'  
  
Volume2 <- cbind(Volume1,dailyVolumeChange)  
Volume2$portfolio\_VolumeRatioDaily2Initial <- Volume2$portfolio\_DailyVolume/Volume2$portfolio\_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" "portfolio\_DailyValue"   
## [109] "portfolio\_prevDay" "portfolio\_dailyValueChange"   
## [111] "portfolio\_ROI\_dollars" "Date"   
## [113] "DayOfWeek" "Month"   
## [115] "Year" "UE\_monthlyRate"   
## [117] "portfolio\_DailyVolume" "portfolio\_prevDayVolume"   
## [119] "portfolio\_dailyVolumeChange" "portfolio\_VolumeRatioDaily2Initial"

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

Stocks$portfolio\_ValueRatioDaily2Initial <-  
 Stocks$portfolio\_DailyValue/Stocks$portfolio\_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$portfolio\_DailyRatios\_X\_UE <-  
 Stocks$portfolio\_ValueRatioDaily2Initial\*Stocks$portfolio\_VolumeRatioDaily2Initial\*Stocks$UE\_monthlyRate

Add an exponential calculation field based on the unemployment rate for rate, and using t=1/12 for 12 months, and a binary value of 1 or 2 where the daily change is positive is assigned a 1 and a negative is a 2. This will make those values decreasing daily have a lower poisson and those values increasing a higher poisson value. This is a modified poisson used for probability of an outcome occuring with a constant rate. Added to rank daily changes based on unemployment rate of each month.

Stocks <- Stocks[complete.cases(Stocks$UE\_monthlyRate),]  
Stocks$dayOfMonth <- day(Stocks$Date)  
dayOfMonth <- day(Stocks$Date)  
ue1 <- Stocks$UE\_monthlyRate  
  
incrDecr <- ifelse(Stocks$portfolio\_dailyValueChange>0,1,2)  
  
Stocks$portfolio\_poisson <- round((exp(-(ue1\*1/12))\*(ue1\*1/12)^incrDecr)/(factorial(incrDecr)),5)  
  
summary(Stocks$portfolio\_poisson)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.03177 0.07392 0.22652 0.19506 0.29808 0.36217

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

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)

All the columns we now have are:

StocksSTATS <- StocksSTATS[,c(1:106,125:230,107:124)]  
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] "TGT\_ROI\_dollars" "FTR\_ROI\_dollars"   
## [109] "UBSI\_ROI\_dollars" "HD\_ROI\_dollars"   
## [111] "JPM\_ROI\_dollars" "XOM\_ROI\_dollars"   
## [113] "CVX\_ROI\_dollars" "NSANY\_ROI\_dollars"   
## [115] "MGM\_ROI\_dollars" "TEVA\_ROI\_dollars"   
## [117] "HST\_ROI\_dollars" "WFC\_ROI\_dollars"   
## [119] "WWE\_ROI\_dollars" "INO\_ROI\_dollars"   
## [121] "SCE.PB\_ROI\_dollars" "FFIN\_ROI\_dollars"   
## [123] "GOOG\_ROI\_dollars" "WM\_ROI\_dollars"   
## [125] "ONCY\_ROI\_dollars" "S\_ROI\_dollars"   
## [127] "F\_ROI\_dollars" "ARWR\_ROI\_dollars"   
## [129] "COST\_ROI\_dollars" "AAL\_ROI\_dollars"   
## [131] "JWN\_ROI\_dollars" "NUS\_ROI\_dollars"   
## [133] "HMC\_ROI\_dollars" "AMZN\_ROI\_dollars"   
## [135] "T\_ROI\_dollars" "HRB\_ROI\_dollars"   
## [137] "RRGB\_ROI\_dollars" "ADDYY\_ROI\_dollars"   
## [139] "PCG\_ROI\_dollars" "ROST\_ROI\_dollars"   
## [141] "JNJ\_ROI\_dollars" "NFLX\_ROI\_dollars"   
## [143] "M\_ROI\_dollars" "KSS\_ROI\_dollars"   
## [145] "DLTR\_ROI\_dollars" "WMT\_ROI\_dollars"   
## [147] "C\_ROI\_dollars" "AAP\_ROI\_dollars"   
## [149] "JBLU\_ROI\_dollars" "MSFT\_ROI\_dollars"   
## [151] "KGJI\_ROI\_dollars" "EPD\_ROI\_dollars"   
## [153] "TJX\_ROI\_dollars" "HOFT\_ROI\_dollars"   
## [155] "LUV\_ROI\_dollars" "NKE\_ROI\_dollars"   
## [157] "TM\_ROI\_dollars" "VZ\_ROI\_dollars"   
## [159] "SIG\_ROI\_dollars" "TGT\_dailyChange"   
## [161] "FTR\_dailyChange" "UBSI\_dailyChange"   
## [163] "HD\_dailyChange" "JPM\_dailyChange"   
## [165] "XOM\_dailyChange" "CVX\_dailyChange"   
## [167] "NSANY\_dailyChange" "MGM\_dailyChange"   
## [169] "TEVA\_dailyChange" "HST\_dailyChange"   
## [171] "WFC\_dailyChange" "WWE\_dailyChange"   
## [173] "INO\_dailyChange" "SCE.PB\_dailyChange"   
## [175] "FFIN\_dailyChange" "GOOG\_dailyChange"   
## [177] "WM\_dailyChange" "ONCY\_dailyChange"   
## [179] "S\_dailyChange" "F\_dailyChange"   
## [181] "ARWR\_dailyChange" "COST\_dailyChange"   
## [183] "AAL\_dailyChange" "JWN\_dailyChange"   
## [185] "NUS\_dailyChange" "HMC\_dailyChange"   
## [187] "AMZN\_dailyChange" "T\_dailyChange"   
## [189] "HRB\_dailyChange" "RRGB\_dailyChange"   
## [191] "ADDYY\_dailyChange" "PCG\_dailyChange"   
## [193] "ROST\_dailyChange" "JNJ\_dailyChange"   
## [195] "NFLX\_dailyChange" "M\_dailyChange"   
## [197] "KSS\_dailyChange" "DLTR\_dailyChange"   
## [199] "WMT\_dailyChange" "C\_dailyChange"   
## [201] "AAP\_dailyChange" "JBLU\_dailyChange"   
## [203] "MSFT\_dailyChange" "KGJI\_dailyChange"   
## [205] "EPD\_dailyChange" "TJX\_dailyChange"   
## [207] "HOFT\_dailyChange" "LUV\_dailyChange"   
## [209] "NKE\_dailyChange" "TM\_dailyChange"   
## [211] "VZ\_dailyChange" "SIG\_dailyChange"   
## [213] "MonthYear" "portfolio\_DailyValue"   
## [215] "portfolio\_prevDay" "portfolio\_dailyValueChange"   
## [217] "portfolio\_ROI\_dollars" "Date"   
## [219] "DayOfWeek" "Month"   
## [221] "Year" "UE\_monthlyRate"   
## [223] "portfolio\_DailyVolume" "portfolio\_prevDayVolume"   
## [225] "portfolio\_dailyVolumeChange" "portfolio\_VolumeRatioDaily2Initial"  
## [227] "portfolio\_ValueRatioDaily2Initial" "portfolio\_DailyRatios\_X\_UE"   
## [229] "dayOfMonth" "portfolio\_poisson"

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

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.

m <- StocksSTATS[order(StocksSTATS$Date, decreasing=FALSE)[length(StocksSTATS$Date)], 107:159]  
t <- as.data.frame(t(m))  
colnames(t) <- row.names(m)  
t$StockROI <- row.names(t)  
  
Troi <- t[order(t$'2020-01-31', decreasing=TRUE),]  
  
mostLeast <- rbind(head(Troi,3),tail(Troi,3))  
mostLeast <- na.omit(mostLeast)  
mostLeast

## 2020-01-31 StockROI  
## AMZN\_ROI\_dollars 1968.300 AMZN\_ROI\_dollars  
## GOOG\_ROI\_dollars 1205.821 GOOG\_ROI\_dollars  
## SCE.PB\_ROI\_dollars 679.000 SCE.PB\_ROI\_dollars  
## MGM\_ROI\_dollars -40.520 MGM\_ROI\_dollars  
## FTR\_ROI\_dollars -225.200 FTR\_ROI\_dollars  
## C\_ROI\_dollars -436.090 C\_ROI\_dollars

The above table shows the three highest returns on investment and the three lowest since Jan 3, 2007 to Jan 31, 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,213:230)]  
amzn\_stock <- StocksSTATS[,c(amzn,213:230)]

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

## dayOfMonth Citi\_Median\_Volume  
## 16 16 22388100  
## 31 31 22302200  
## 3 3 21221500  
## 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. :-298.300   
## 1st Qu.: 41.80 1st Qu.: 13019600 1st Qu.:-468.7 1st Qu.: -0.680   
## Median : 51.49 Median : 19493900 Median :-459.0 Median : -0.010   
## Mean : 93.38 Mean : 26987469 Mean :-417.1 Mean : 0.021   
## 3rd Qu.: 69.46 3rd Qu.: 33280800 3rd Qu.:-441.0 3rd Qu.: 0.650   
## Max. :552.50 Max. :377263800 Max. : 42.0 Max. : 510.500   
##   
## MonthYear portfolio\_DailyValue portfolio\_prevDay  
## Aug-2007: 23 Min. :1229 Min. :1229   
## Aug-2011: 23 1st Qu.:2821 1st Qu.:2821   
## Aug-2012: 23 Median :3542 Median :3541   
## Aug-2016: 23 Mean :3988 Mean :3986   
## Aug-2017: 23 3rd Qu.:5104 3rd Qu.:5104   
## Aug-2018: 23 Max. :7910 Max. :7910   
## (Other) :3155   
## portfolio\_dailyValueChange portfolio\_ROI\_dollars Date   
## Min. :-1014.322 Min. :-1748.9 Min. :2007-01-03   
## 1st Qu.: -39.065 1st Qu.: -157.4 1st Qu.:2010-04-12   
## Median : 2.276 Median : 563.9 Median :2013-07-18   
## Mean : 1.475 Mean : 1009.6 Mean :2013-07-16   
## 3rd Qu.: 43.517 3rd Qu.: 2126.4 3rd Qu.:2016-10-21   
## Max. : 1025.453 Max. : 4931.7 Max. :2020-01-31   
##   
## DayOfWeek Month Year UE\_monthlyRate   
## Length:3293 Length:3293 Min. :2007 Min. : 3.500   
## Class :character Class :character 1st Qu.:2010 1st Qu.: 4.600   
## Mode :character Mode :character Median :2013 Median : 5.600   
## Mean :2013 Mean : 6.282   
## 3rd Qu.:2016 3rd Qu.: 8.200   
## Max. :2020 Max. :10.000   
##   
## portfolio\_DailyVolume portfolio\_prevDayVolume portfolio\_dailyVolumeChange  
## Min. :1.133e+08 Min. :1.133e+08 Min. :-714176400   
## 1st Qu.:3.370e+08 1st Qu.:3.370e+08 1st Qu.: -50722061   
## Median :4.194e+08 Median :4.196e+08 Median : 250560   
## Mean :4.752e+08 Mean :4.753e+08 Mean : -55791   
## 3rd Qu.:5.716e+08 3rd Qu.:5.716e+08 3rd Qu.: 50561500   
## Max. :1.611e+09 Max. :1.611e+09 Max. : 620907605   
##   
## portfolio\_VolumeRatioDaily2Initial portfolio\_ValueRatioDaily2Initial  
## Min. :0.1981 Min. :0.4236   
## 1st Qu.:0.5891 1st Qu.:0.9720   
## Median :0.7333 Median :1.2206   
## Mean :0.8307 Mean :1.3742   
## 3rd Qu.:0.9992 3rd Qu.:1.7591   
## Max. :2.8163 Max. :2.7259   
##   
## portfolio\_DailyRatios\_X\_UE dayOfMonth portfolio\_poisson  
## Min. : 0.9658 Min. : 1.00 Min. :0.03177   
## 1st Qu.: 4.4923 1st Qu.: 8.00 1st Qu.:0.07392   
## Median : 5.6528 Median :16.00 Median :0.22652   
## Mean : 6.4285 Mean :15.74 Mean :0.19506   
## 3rd Qu.: 7.8497 3rd Qu.:23.00 3rd Qu.:0.29808   
## Max. :24.2627 Max. :31.00 Max. :0.36217   
##

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  
## 2007-04-02 510.50 2282100 0.00 510.500000 Apr-2007  
## 2013-04-02 44.11 1005100 -466.39 0.320000 Apr-2013  
## 2015-12-28 52.38 377263800 -458.12 -0.329998 Dec-2015  
## 2008-06-02 214.60 15302800 -295.90 -298.300018 Jun-2008  
## portfolio\_DailyValue portfolio\_prevDay portfolio\_dailyValueChange  
## 2007-04-02 2901.650 2891.963 9.686608  
## 2013-04-02 3433.938 3354.901 79.037872  
## 2015-12-28 5005.455 4984.970 20.485009  
## 2008-06-02 3120.541 3144.698 -24.157199  
## portfolio\_ROI\_dollars Date DayOfWeek Month Year UE\_monthlyRate  
## 2007-04-02 -76.28907 2007-04-02 Monday Apr 2007 4.5  
## 2013-04-02 455.99978 2013-04-02 Tuesday Apr 2013 7.6  
## 2015-12-28 2027.51641 2015-12-28 Monday Dec 2015 5.0  
## 2008-06-02 142.60220 2008-06-02 Monday Jun 2008 5.6  
## portfolio\_DailyVolume portfolio\_prevDayVolume  
## 2007-04-02 572035712 572035712  
## 2013-04-02 258084601 330998801  
## 2015-12-28 975152259 752607802  
## 2008-06-02 464823559 265152951  
## portfolio\_dailyVolumeChange portfolio\_VolumeRatioDaily2Initial  
## 2007-04-02 0 1.0000000  
## 2013-04-02 -72914200 0.4511687  
## 2015-12-28 222544457 1.7047052  
## 2008-06-02 199670608 0.8125779  
## portfolio\_ValueRatioDaily2Initial portfolio\_DailyRatios\_X\_UE  
## 2007-04-02 1.000000 4.500000  
## 2013-04-02 1.183444 4.057888  
## 2015-12-28 1.725038 14.703404  
## 2008-06-02 1.075437 4.893707  
## dayOfMonth portfolio\_poisson  
## 2007-04-02 2 0.25773  
## 2013-04-02 2 0.33619  
## 2015-12-28 28 0.27468  
## 2008-06-02 2 0.06828

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] 32.16953

## [1] 96.50858

The standard error for the daily change in dollars is 32.17 USD and our threshold to find dates outside this normal range of daily change dollar values is 96.51 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 < out, '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. :330.6 Min. : 2282100 Min. :-179.90 Min. :266.2   
## 1st Qu.:471.2 1st Qu.:13456250 1st Qu.: -39.30 1st Qu.:399.6   
## Median :510.6 Median :19551450 Median : 0.15 Median :441.4   
## Mean :488.2 Mean :30425167 Mean : -22.32 Mean :424.4   
## 3rd Qu.:542.8 3rd Qu.:35952375 3rd Qu.: 32.27 3rd Qu.:475.4   
## Max. :552.5 Max. :81343800 Max. : 42.00 Max. :510.5   
##   
## MonthYear portfolio\_DailyValue portfolio\_prevDay portfolio\_dailyValueChange  
## Apr-2007:1 Min. :2724 Min. :2744 Min. :-85.034   
## Aug-2007:1 1st Qu.:2899 1st Qu.:2878 1st Qu.: -4.048   
## Dec-2007:1 Median :2974 Median :2942 Median : -1.393   
## Feb-2007:1 Mean :3104 Mean :3044 Mean : 59.150   
## Jan-2007:1 3rd Qu.:3343 3rd Qu.:3076 3rd Qu.: 20.755   
## Jul-2007:1 Max. :3656 Max. :3619 Max. :734.207   
## (Other) :6   
## portfolio\_ROI\_dollars Date DayOfWeek   
## Min. :-253.961 Min. :2007-01-03 Length:12   
## 1st Qu.: -79.356 1st Qu.:2007-03-25 Class :character   
## Median : -4.371 Median :2007-06-16 Mode :character   
## Mean : 125.597 Mean :2007-06-17   
## 3rd Qu.: 364.923 3rd Qu.:2007-09-10   
## Max. : 677.926 Max. :2007-12-03   
##   
## Month Year UE\_monthlyRate portfolio\_DailyVolume  
## Length:12 Min. :2007 Min. :4.400 Min. :2.160e+08   
## Class :character 1st Qu.:2007 1st Qu.:4.500 1st Qu.:3.962e+08   
## Mode :character Median :2007 Median :4.600 Median :4.644e+08   
## Mean :2007 Mean :4.617 Mean :5.398e+08   
## 3rd Qu.:2007 3rd Qu.:4.700 3rd Qu.:6.314e+08   
## Max. :2007 Max. :5.000 Max. :1.005e+09   
##   
## portfolio\_prevDayVolume portfolio\_dailyVolumeChange  
## Min. :198190500 Min. :-197842207   
## 1st Qu.:387785669 1st Qu.: -23781530   
## Median :564614969 Median : 26069930   
## Mean :528884214 Mean : 10878309   
## 3rd Qu.:594041737 3rd Qu.: 70618878   
## Max. :971072459 Max. : 124348468   
##   
## portfolio\_VolumeRatioDaily2Initial portfolio\_ValueRatioDaily2Initial  
## Min. :0.3776 Min. :0.9388   
## 1st Qu.:0.6926 1st Qu.:0.9989   
## Median :0.8118 Median :1.0248   
## Mean :0.9436 Mean :1.0696   
## 3rd Qu.:1.1038 3rd Qu.:1.1521   
## Max. :1.7576 Max. :1.2599   
##   
## portfolio\_DailyRatios\_X\_UE dayOfMonth portfolio\_poisson Threshold3   
## Min. :1.654 Min. :1.00 Min. :0.04659 Length:12   
## 1st Qu.:3.696 1st Qu.:1.00 1st Qu.:0.05008 Class :character   
## Median :4.400 Median :1.00 Median :0.05454 Mode :character   
## Mean :4.641 Mean :1.75 Mean :0.13836   
## 3rd Qu.:5.116 3rd Qu.:2.25 3rd Qu.:0.25948   
## Max. :8.297 Max. :4.00 Max. :0.26474   
##

We can see from the above statistics on the subset of Citigroup stock that are outside this threshold that there are 12 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[,-23])  
NLP\_dates\_Citi

## C.Close C.Volume C\_ROI\_dollars C\_dailyChange MonthYear  
## 2007-04-02 510.50 2282100 0.000000 510.500000 Apr-2007  
## 2013-04-02 44.11 1005100 -466.389999 0.320000 Apr-2013  
## 2015-12-28 52.38 377263800 -458.119999 -0.329998 Dec-2015  
## 2008-06-02 214.60 15302800 -295.899994 -298.300018 Jun-2008  
## 2007-04-021 510.50 2282100 0.000000 510.500000 Apr-2007  
## 2007-08-01 468.50 13495700 -42.000000 397.800003 Aug-2007  
## 2007-12-03 330.60 81343800 -179.899994 266.250008 Dec-2007  
## 2007-02-01 547.30 80864600 36.799988 467.409989 Feb-2007  
## 2007-01-03 552.50 43508100 42.000000 488.520000 Jan-2007  
## 2007-07-02 516.40 32822200 5.900024 441.990020 Jul-2007  
## 2007-06-01 545.10 23057000 34.599976 473.939972 Jun-2007  
## 2007-03-01 510.80 8981300 0.299988 440.769989 Mar-2007  
## 2007-05-01 542.00 13337900 31.500000 479.779999 May-2007  
## 2007-11-01 385.10 33433800 -125.399994 322.950004 Nov-2007  
## 2007-10-01 477.20 16045900 -33.299988 402.080009 Oct-2007  
## 2007-09-04 472.10 15929600 -38.399994 400.240005 Sep-2007  
## portfolio\_DailyValue portfolio\_prevDay portfolio\_dailyValueChange  
## 2007-04-02 2901.650 2891.963 9.686608  
## 2013-04-02 3433.938 3354.901 79.037872  
## 2015-12-28 5005.455 4984.970 20.485009  
## 2008-06-02 3120.541 3144.698 -24.157199  
## 2007-04-021 2901.650 2891.963 9.686608  
## 2007-08-01 2778.299 2781.133 -2.834138  
## 2007-12-03 2723.978 2743.972 -19.993872  
## 2007-02-01 3279.015 3281.965 -2.949476  
## 2007-01-03 2977.939 2977.939 0.000000  
## 2007-07-02 2969.196 2946.619 22.576765  
## 2007-06-01 3003.989 3006.774 -2.785581  
## 2007-03-01 2889.381 2896.725 -7.344424  
## 2007-05-01 2957.539 2937.392 20.147648  
## 2007-11-01 3534.398 3619.433 -85.034241  
## 2007-10-01 3655.864 3611.738 44.126353  
## 2007-09-04 3571.178 2836.972 734.206543  
## portfolio\_ROI\_dollars Date DayOfWeek Month Year  
## 2007-04-02 -76.289072 2007-04-02 Monday Apr 2007  
## 2013-04-02 455.999776 2013-04-02 Tuesday Apr 2013  
## 2015-12-28 2027.516411 2015-12-28 Monday Dec 2015  
## 2008-06-02 142.602196 2008-06-02 Monday Jun 2008  
## 2007-04-021 -76.289072 2007-04-02 Monday Apr 2007  
## 2007-08-01 -199.639490 2007-08-01 Wednesday Aug 2007  
## 2007-12-03 -253.960930 2007-12-03 Monday Dec 2007  
## 2007-02-01 301.076786 2007-02-01 Thursday Feb 2007  
## 2007-01-03 0.000000 2007-01-03 Wednesday Jan 2007  
## 2007-07-02 -8.742779 2007-07-02 Monday Jul 2007  
## 2007-06-01 26.049900 2007-06-01 Friday Jun 2007  
## 2007-03-01 -88.557542 2007-03-01 Thursday Mar 2007  
## 2007-05-01 -20.399119 2007-05-01 Tuesday May 2007  
## 2007-11-01 556.459753 2007-11-01 Thursday Nov 2007  
## 2007-10-01 677.925528 2007-10-01 Monday Oct 2007  
## 2007-09-04 593.239860 2007-09-04 Tuesday Sep 2007  
## UE\_monthlyRate portfolio\_DailyVolume portfolio\_prevDayVolume  
## 2007-04-02 4.5 572035712 572035712  
## 2013-04-02 7.6 258084601 330998801  
## 2015-12-28 5.0 975152259 752607802  
## 2008-06-02 5.6 464823559 265152951  
## 2007-04-021 4.5 572035712 572035712  
## 2007-08-01 4.6 686001371 572681959  
## 2007-12-03 5.0 1005429691 971072459  
## 2007-02-01 4.5 933350159 809001691  
## 2007-01-03 4.6 613250413 565411759  
## 2007-07-02 4.7 460278863 658121070  
## 2007-06-01 4.6 381151267 397701502  
## 2007-03-01 4.4 215973129 198190500  
## 2007-05-01 4.4 314742689 233827359  
## 2007-11-01 4.7 468477291 563818179  
## 2007-10-01 4.7 401234791 446710205  
## 2007-09-04 4.7 425224899 358038171  
## portfolio\_dailyVolumeChange portfolio\_VolumeRatioDaily2Initial  
## 2007-04-02 0 1.0000000  
## 2013-04-02 -72914200 0.4511687  
## 2015-12-28 222544457 1.7047052  
## 2008-06-02 199670608 0.8125779  
## 2007-04-021 0 1.0000000  
## 2007-08-01 113319412 1.1992282  
## 2007-12-03 34357232 1.7576345  
## 2007-02-01 124348468 1.6316292  
## 2007-01-03 47838654 1.0720492  
## 2007-07-02 -197842207 0.8046331  
## 2007-06-01 -16550235 0.6663068  
## 2007-03-01 17782629 0.3775518  
## 2007-05-01 80915330 0.5502151  
## 2007-11-01 -95340888 0.8189651  
## 2007-10-01 -45475414 0.7014156  
## 2007-09-04 67186728 0.7433538  
## portfolio\_ValueRatioDaily2Initial portfolio\_DailyRatios\_X\_UE  
## 2007-04-02 1.0000000 4.500000  
## 2013-04-02 1.1834435 4.057888  
## 2015-12-28 1.7250378 14.703404  
## 2008-06-02 1.0754368 4.893707  
## 2007-04-021 1.0000000 4.500000  
## 2007-08-01 0.9574896 5.281943  
## 2007-12-03 0.9387687 8.250061  
## 2007-02-01 1.1300522 8.297218  
## 2007-01-03 1.0262916 5.061081  
## 2007-07-02 1.0232786 3.869810  
## 2007-06-01 1.0352692 3.173112  
## 2007-03-01 0.9957719 1.654204  
## 2007-05-01 1.0192614 2.467577  
## 2007-11-01 1.2180652 4.688499  
## 2007-10-01 1.2599262 4.153540  
## 2007-09-04 1.2307408 4.299916  
## dayOfMonth portfolio\_poisson  
## 2007-04-02 2 0.25773  
## 2013-04-02 2 0.33619  
## 2015-12-28 28 0.27468  
## 2008-06-02 2 0.06828  
## 2007-04-021 2 0.25773  
## 2007-08-01 1 0.05008  
## 2007-12-03 3 0.05723  
## 2007-02-01 1 0.04833  
## 2007-01-03 3 0.05008  
## 2007-07-02 2 0.26474  
## 2007-06-01 1 0.05008  
## 2007-03-01 1 0.04659  
## 2007-05-01 1 0.25411  
## 2007-11-01 1 0.05184  
## 2007-10-01 1 0.26474  
## 2007-09-04 4 0.26474

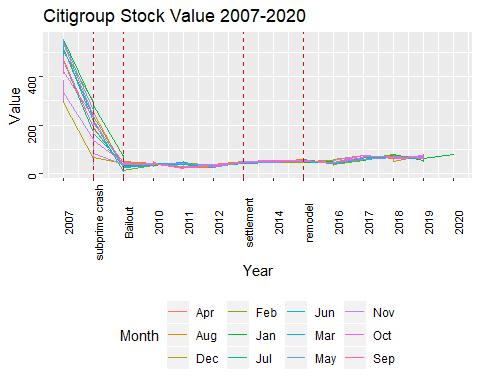
I am going to pull the data from these dates with the Google Search for the specific date on Citigroup stock, put it in a table with the date, the article title, reference, article content, and the comments if available.

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.

Lets plot this as a simple line chart of the value of the stock over the years.

ggplot(data = C\_stock, 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")+  
 scale\_x\_continuous(breaks=c(2007,2008,2009,2010,2011,2012,2013,2014,2015,2016,2017,2018,2019,2020),  
 labels=c(2007,'subprime crash','Bailout',2010,2011,2012,'settlement',2014,'remodel',2016,2017,2018,2019,2020))+  
 theme(axis.text = element\_text(colour = "black", angle=90, size = rel(.75)))+  
 geom\_vline(xintercept=c(2008,2009,2013,2015), linetype='dashed', color='red')+  
 ggtitle('Citigroup Stock Value 2007-2020')+  
 ylab('Value')

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



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 13019600 19493900 26987469 33280800 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  
## 2015-12-23 52.63 93423000 -457.87 0.620003 Dec-2015  
## 2015-12-24 52.71 119108100 -457.79 0.079998 Dec-2015  
## 2015-12-28 52.38 377263800 -458.12 -0.329998 Dec-2015  
## 2015-12-29 52.98 281369700 -457.52 0.599999 Dec-2015  
## 2015-12-30 52.30 62625000 -458.20 -0.680001 Dec-2015  
## 2015-12-31 51.75 49092600 -458.75 -0.549999 Dec-2015  
## portfolio\_DailyValue portfolio\_prevDay portfolio\_dailyValueChange  
## 2015-12-23 4998.690 4968.045 30.64500  
## 2015-12-24 4984.970 4998.690 -13.72002  
## 2015-12-28 5005.455 4984.970 20.48501  
## 2015-12-29 4738.190 5005.455 -267.26507  
## 2015-12-30 4800.285 4738.190 62.09506  
## 2015-12-31 4707.685 4800.285 -92.59999  
## portfolio\_ROI\_dollars Date DayOfWeek Month Year UE\_monthlyRate  
## 2015-12-23 2020.751 2015-12-23 Wednesday Dec 2015 5  
## 2015-12-24 2007.031 2015-12-24 Thursday Dec 2015 5  
## 2015-12-28 2027.516 2015-12-28 Monday Dec 2015 5  
## 2015-12-29 1760.251 2015-12-29 Tuesday Dec 2015 5  
## 2015-12-30 1822.346 2015-12-30 Wednesday Dec 2015 5  
## 2015-12-31 1729.746 2015-12-31 Thursday Dec 2015 5  
## portfolio\_DailyVolume portfolio\_prevDayVolume  
## 2015-12-23 903674159 619024059  
## 2015-12-24 752607802 903674159  
## 2015-12-28 975152259 752607802  
## 2015-12-29 1248436459 975152259  
## 2015-12-30 534260059 1248436459  
## 2015-12-31 504630159 534260059  
## portfolio\_dailyVolumeChange portfolio\_VolumeRatioDaily2Initial  
## 2015-12-23 284650100 1.5797513  
## 2015-12-24 -151066357 1.3156658  
## 2015-12-28 222544457 1.7047052  
## 2015-12-29 273284200 2.1824450  
## 2015-12-30 -714176400 0.9339628  
## 2015-12-31 -29629900 0.8821655  
## portfolio\_ValueRatioDaily2Initial portfolio\_DailyRatios\_X\_UE  
## 2015-12-23 1.722706 13.607238  
## 2015-12-24 1.717978 11.301424  
## 2015-12-28 1.725038 14.703404  
## 2015-12-29 1.632930 17.818897  
## 2015-12-30 1.654330 7.725412  
## 2015-12-31 1.622417 7.156201  
## dayOfMonth portfolio\_poisson Threshold3  
## 2015-12-23 23 0.27468 inside  
## 2015-12-24 24 0.05723 inside  
## 2015-12-28 28 0.27468 inside  
## 2015-12-29 29 0.05723 inside  
## 2015-12-30 30 0.27468 inside  
## 2015-12-31 31 0.05723 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: 157 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 147 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



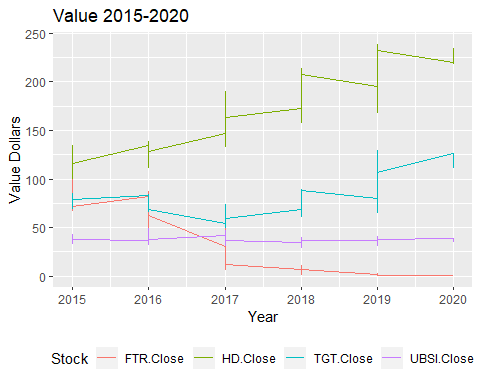
The above line chart of the mean monthly dollar value of the Citigroup stock show that all months move the same direction of decreasing in 2015, increasing in 2016, except for in 2017 and 2018 where 3-6 months decreased and 6-9 months increased monthly mean values. The span of 2019 through 2020 can’t be analyzed yet, but January 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.

We saw that Citigroup is maintaining its current value and slightly increasing over the last four years. Lets start subset sampling stocks and look at the changes they have made in value over the last four years. And see if we notice anything we want to further exploit.

Value1 <- StocksSTATS[,c(1:53,160:230)]  
Value2 <- subset(Value1, Year>2014)

sub1 <- Value2[,c(1:4,115)]  
sub1tidy <- gather(sub1, 'Stock','Value',1:4)  
  
ggplot(data = sub1tidy, aes(x=Year, y=Value,group=Stock)) +  
 geom\_line(aes(color=Stock))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Value 2015-2020')+  
 ylab('Value Dollars')

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



The first four stocks in our set of 53 is shown in the line chart above from 2015-2020.

From the above line chart, it is obvious that over the last five years, the pink line for FTR is a terrible stock as it has been on the decline, but we would have to look at it further to see why it has been decreasing in value since 2015.

The olive color line for HD indicates it has been on a steady increase from the 120-125 USD range in 2015 to the 220-225 USD range in 2020.

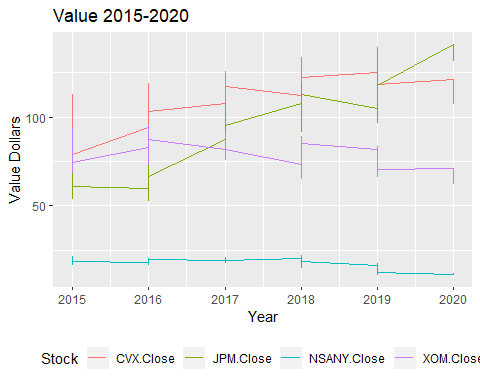
Also, increasing steadily is the blue line for TGT, which started at 75-80 in 2015 and is at 125 in 2020 in value.

The purple line for UBSI has been maintaining steadily from 45 range to 45 range over five years. \*\*\*

Lets look at the next four stocks.

sub1 <- Value2[,c(5:8,115)]  
  
sub1tidy <- gather(sub1, 'Stock','Value',1:4)  
  
ggplot(data = sub1tidy, aes(x=Year, y=Value,group=Stock)) +  
 geom\_line(aes(color=Stock))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Value 2015-2020')+  
 ylab('Value Dollars')

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

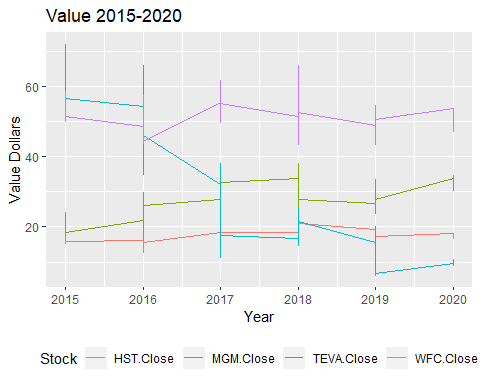


From the above subset of the next four stock in our 53 stocks, we can see that there are two stocks increasing significantly for JPM and CVX. We also note that the XOM and NSANY stocks have decreased over the last five years. \*\*\*

Now for the next four stocks.

sub1 <- Value2[,c(9:12,115)]  
sub1tidy <- gather(sub1, 'Stock','Value',1:4)  
  
ggplot(data = sub1tidy, aes(x=Year, y=Value,group=Stock)) +  
 geom\_line(aes(color=Stock))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Value 2015-2020')+  
 ylab('Value Dollars')

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



The above line chart shows the third subset of four stocks of our 53 stocks.

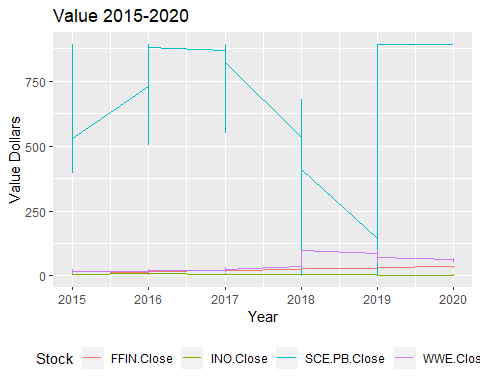
The MGM stock has increased significantly since 2005, and slight increases are shown for WFC and HST though not significantly. There is some cyclical movements in the WFC with 2016 giving a steady increase all year, then declining 2017-2019, and ending with a steady increase in 2019.

The TEVA stock has had a huge loss over the last five years, with the last year showing an an increase slightly. It started at the 55 range in 2015 and is at the 10 range in 2020. This could indicate that it is a good time to buy TEVA, since it is priced low and shows an increase in the last year, where the last four years it has been decreasing annually for each year. This would require further analysis for why it has been decreasing over the last five years. \*\*\*

Now for the next four stocks in our subset four.

sub1 <- Value2[,c(13:16,115)]  
sub1tidy <- gather(sub1, 'Stock','Value',1:4)  
  
ggplot(data = sub1tidy, aes(x=Year, y=Value,group=Stock)) +  
 geom\_line(aes(color=Stock))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Value 2015-2020')+  
 ylab('Value Dollars')

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



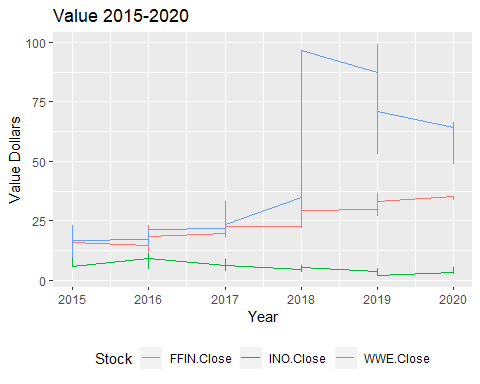
The above line chart shows that SCE.PB is on its own scale that outweighs the scale of the other smaller valued stocks, there is also volatility and cyclical movements in SCE.PB which makes it a good choice to further analyze with timelines of web article events that could have triggered these changes in value of a steady increase in 2015, a high jump increase in 2016, then a steep decline throughout 2017 and 2018, then a huge jump of an increase to the same level at 2016. This is a utility company so government contracts could be involved with all that entails, and possible fires causing damage and settlements in the declining years. But for now it is just speculation and assumptions.

The other stocks are getting limited spotlight above, and they need their own scale as SCE.PB pushed down their scaled visual line charts.

Now for the next four stocks in our subset four.

sub1 <- Value2[,c(13,14,16,115)]  
sub1tidy <- gather(sub1, 'Stock','Value',1:3)  
  
ggplot(data = sub1tidy, aes(x=Year, y=Value,group=Stock)) +  
 geom\_line(aes(color=Stock))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Value 2015-2020')+  
 ylab('Value Dollars')

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



From the above line chart, we see that WWE had a huge jump in 2018 of an increase from the 40 range to the 90 range but then decreased during 2018 and 2019 to a price still much higher at the 60 range than its starting value in 2015 of the 20 range.

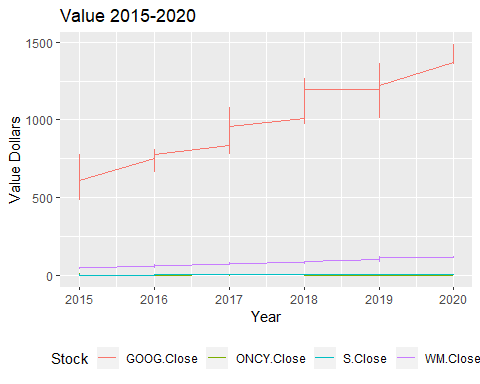
The FFIN stock has been steadily increasing over the last five years with a flat line on the value in 2017 and 2018.

The INO stock has declined since 2016 after an increasing year in 2015, but lost only slightly in value over a five year span returning no profits over that time span.

Now for the next stocks in our subset.

sub1 <- Value2[,c(17:20,115)]  
sub1tidy <- gather(sub1, 'Stock','Value',1:4)  
  
ggplot(data = sub1tidy, aes(x=Year, y=Value,group=Stock)) +  
 geom\_line(aes(color=Stock))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Value 2015-2020')+  
 ylab('Value Dollars')

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

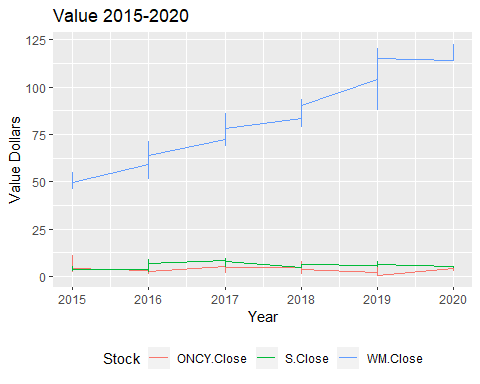


In the above subset of stocks, Google out scales the other three stocks and shows that it has been increasing steadily every year, except 2018 where it is almost the same price all year.

Lets look at the other three stocks that our on a lower scaled value to analyze them.

sub1 <- Value2[,c(18:20,115)]  
sub1tidy <- gather(sub1, 'Stock','Value',1:3)  
  
#min2015 <- subset(StocksSTATS, StocksSTATS$MonthYear=='Jan-2015')  
#m15 <- min2015[1,"ONCY.Close"]  
  
ggplot(data = sub1tidy, aes(x=Year, y=Value,group=Stock)) +  
 geom\_line(aes(color=Stock))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 #geom\_hline(yintercept=m15, color='red')  
 ggtitle('Value 2015-2020')+  
 ylab('Value Dollars')

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



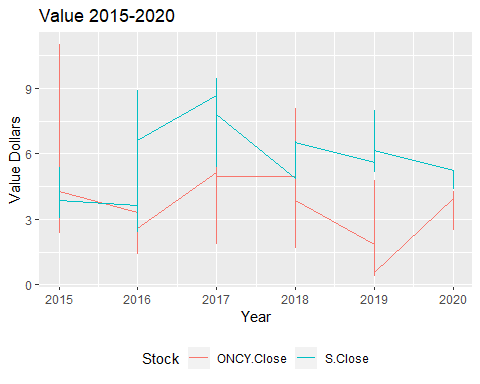
The line chart above shows that WM has increased significantly every year since 2015, with a slight decrease in 2019, but overall has increased from the 50 range in 2015 to the 113 range in 2020.

The ONCY and S stocks have had slight increases and decreases in the last five years but look like they have increased slightly overall from 2015-2020.

Lets look at S and ONCY stocks more closely.

sub1 <- Value2[,c(19:20,115)]  
sub1tidy <- gather(sub1, 'Stock','Value',1:2)  
  
ggplot(data = sub1tidy, aes(x=Year, y=Value,group=Stock)) +  
 geom\_line(aes(color=Stock))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Value 2015-2020')+  
 ylab('Value Dollars')

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

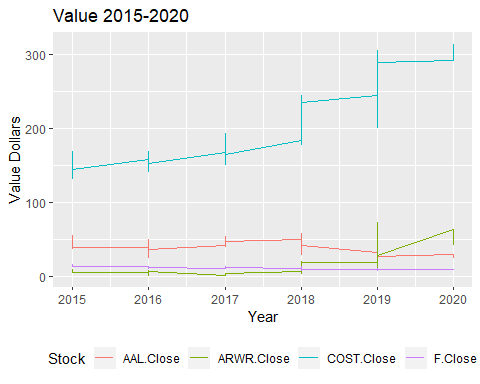


It looks like these two stocks, ONCY and S, have had cyclical patterns in the last five years, and if that is true, then S stock hasn’t reached its cyclical minimum and ONCY stock hasn’t reached it cyclical maximum. And if this is not the case then there are some triggers in the value of this stock in 2016, where they both increased, then steadily decreased in 2017. A global minimum in the last five years is seen in 2019 for ONCY stock, while the global maximums for both stock is in 2017. The start of 2016 showed both stocks had a local minima while S stock had its global minima this year, but only for this last five year period.

Now for the next stocks in our subset.

sub1 <- Value2[,c(21:24,115)]  
sub1tidy <- gather(sub1, 'Stock','Value',1:4)  
  
ggplot(data = sub1tidy, aes(x=Year, y=Value,group=Stock)) +  
 geom\_line(aes(color=Stock))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Value 2015-2020')+  
 ylab('Value Dollars')

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

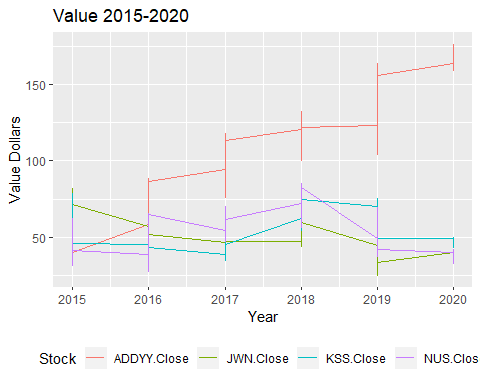


The above subset shows that ARWR and COST stock have been increasing the last two years, but ARWR stock had some near flat changes in value for years 2015, 2016, and 2017. The purple line for Ford is relatively maintaining value, but no increases or decreases of note for Ford in the last five years. The AAL stock had a global maxima in 2018 but overall decreased in value slightly in the last five years. \*\*\*

Now for the next stocks in our subset.

sub1 <- Value2[,c(25:28,115)]  
sub1tidy <- gather(sub1, 'Stock','Value',1:4)  
  
ggplot(data = sub1tidy, aes(x=Year, y=Value,group=Stock)) +  
 geom\_line(aes(color=Stock))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Value 2015-2020')+  
 ylab('Value Dollars')

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

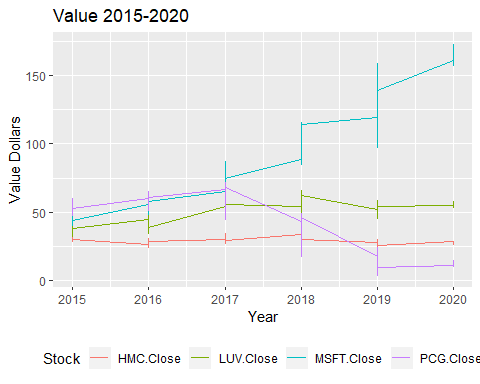


The above line chart shows that ADDYY has been significantly increasing over the last five years it jumped from the 40 USD range to the 165 USD range in 2020. The other three stocks all moved together with slightly different rates of increase and decrease. But the JWN stock lost value over the last five years, while KSS and NUS stocks both increased only marginally after some cyclical rise and falls in value. \*\*\*

Now for the next stocks in our subset.

sub1 <- Value2[,c(29:32,115)]  
sub1tidy <- gather(sub1, 'Stock','Value',1:4)  
  
ggplot(data = sub1tidy, aes(x=Year, y=Value,group=Stock)) +  
 geom\_line(aes(color=Stock))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Value 2015-2020')+  
 ylab('Value Dollars')

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

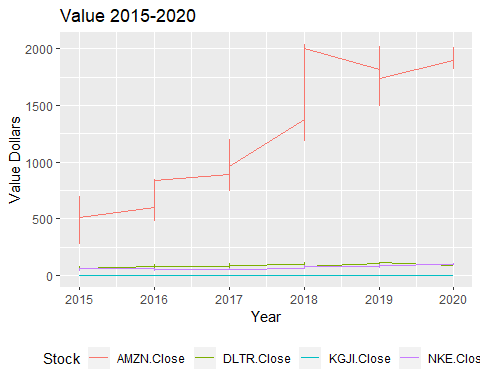


The above line chart shows that MSFT increased steadily the last five years with none of the years having declining values in stock. PCG stock had a local maxima in 2017 but a local minima in 2019 which led to an overall loss in value from 2015-2020. The LUV stock is the olive colored stock that had an increase overall in value by about 10 USD. And the HMC stock slightly stayed the same and may have decreased marginally in the last five years. \*\*\*

Now for the next stocks in our subset.

sub1 <- Value2[,c(33:36,115)]  
sub1tidy <- gather(sub1, 'Stock','Value',1:4)  
  
ggplot(data = sub1tidy, aes(x=Year, y=Value,group=Stock)) +  
 geom\_line(aes(color=Stock))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Value 2015-2020')+  
 ylab('Value Dollars')

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

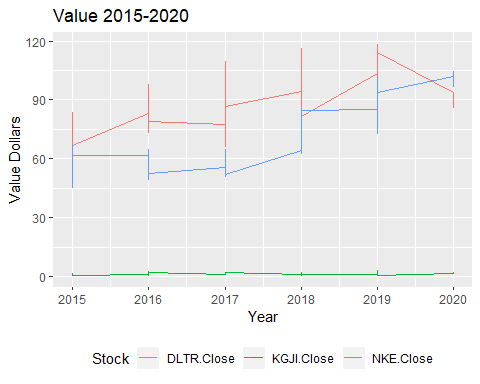


The above line chart shows that AMZN stock is on its own scale and has saw an overall huge jump in value in the last five years, with every year increasing, except in 2018 where it decreased from its local maxima at the start of 2018. Its value in 2015 was in the 500 USD range and at the start of 2020 was in the 1700-1800 USD range.

Lets look at the scale more appropriate for the other three stocks of DLTR, KGJI, and NKE.

sub1 <- Value2[,c(33:35,115)]  
sub1tidy <- gather(sub1, 'Stock','Value',1:3)  
  
ggplot(data = sub1tidy, aes(x=Year, y=Value,group=Stock)) +  
 geom\_line(aes(color=Stock))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Value 2015-2020')+  
 ylab('Value Dollars')

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

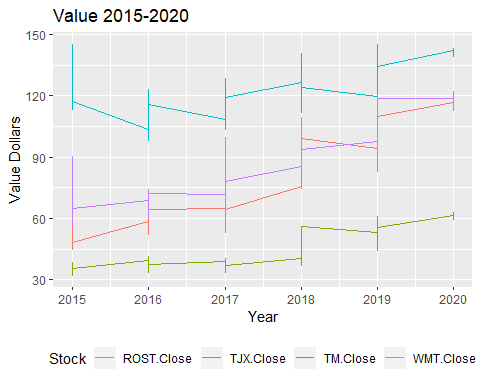


The above line chart shows the smaller scale value changes by year for DLTR, KGJI, and NKE. Both NKE and DLTR stocks have increased in value over the last five years, while DLTR did see a decreasing value throughout the last year of 2019. The KGJI stock showed marginal changes in value over the last five years, with no significant local minimas or local maximas.It does look like a slight increase overall from 2015-2020 for KGJI stock. \*\*\*

Now for the next stocks in our subset.

sub1 <- Value2[,c(37:40,115)]  
sub1tidy <- gather(sub1, 'Stock','Value',1:4)  
  
ggplot(data = sub1tidy, aes(x=Year, y=Value,group=Stock)) +  
 geom\_line(aes(color=Stock))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Value 2015-2020')+  
 ylab('Value Dollars')

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

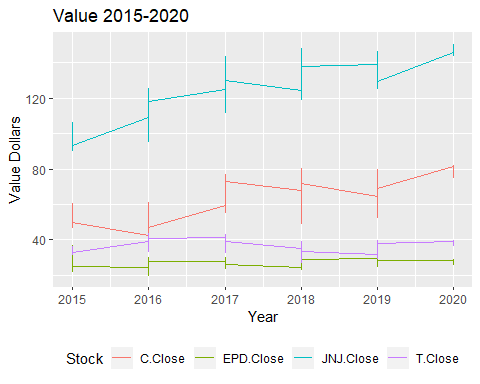


In the above line chart we see that all of the stocks increased noticeably in the last five years. The TM stock had some years that decreased in 2015, 2016, and 2018, but always starts the new year at a higher value than the year before. In 2018 WMT increased, while the other three stocks of TJX, TM, and ROST saw slight decreases. \*\*\*

Now for the next stocks in our subset.

sub1 <- Value2[,c(41:44,115)]  
sub1tidy <- gather(sub1, 'Stock','Value',1:4)  
  
ggplot(data = sub1tidy, aes(x=Year, y=Value,group=Stock)) +  
 geom\_line(aes(color=Stock))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Value 2015-2020')+  
 ylab('Value Dollars')

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

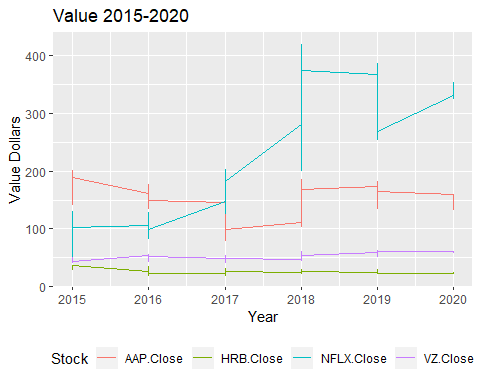


The above line chart also shows an overall increase in value over the last five years with significant jumps in value for C and JNJ stocks. In 2017, there were some decreases in value throughout the year for all these stocks of C, EPD, JNJ, and T stocks, but in two years they all started 2019 at the same values of 2017 and saw increasing values throughout 2019. \*\*\*

Now for the next stocks in our subset.

sub1 <- Value2[,c(45:48,115)]  
sub1tidy <- gather(sub1, 'Stock','Value',1:4)  
  
ggplot(data = sub1tidy, aes(x=Year, y=Value,group=Stock)) +  
 geom\_line(aes(color=Stock))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Value 2015-2020')+  
 ylab('Value Dollars')

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



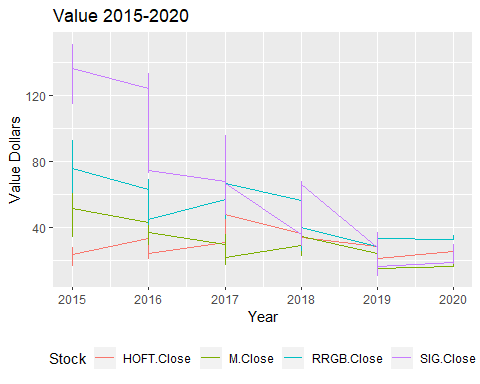
The above line chart shows that NFLX increased significantly while HRB and AAP saw losses over the last five years. VZ stock saw a slight increase in value over the last five years. In 2017 Netflix saw a huge inrease, while in 2018 it stayed somewhat stagnant with a sharp drop in value at the start of 2019 that saw an increasing year throughout 2019.

In 2017, there was a sharp drop in value for AAP, but by the start of 2018 the value increased to a value above the start of 2017.

Now for the last five stocks in our subset.

sub1 <- Value2[,c(49:53,115)]  
sub1tidy <- gather(sub1, 'Stock','Value',1:4)  
  
ggplot(data = sub1tidy, aes(x=Year, y=Value,group=Stock)) +  
 geom\_line(aes(color=Stock))+  
 scale\_y\_continuous()+  
 scale\_fill\_brewer(palette="paired") +  
 theme(legend.position="bottom")+  
 ggtitle('Value 2015-2020')+  
 ylab('Value Dollars')

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



Our last set of stock show that RRGB and SIG saw significant losses over the last five years, while M stock showed a smaller loss. HOFT stock saw an increase over the last five years, but only marginally or slightly. In 2017 M stock saw an increasing year for its value after having two years from 2015-2016 see decreasing values throughout those years. M stock and HOFT stock seemed to be negatively correlated for years 2015-2018, with both stocks having different rates of decrease in 2018 and an increase in value of similar rates of increase in 2019. All of these stocks decreased at different rates in 2018, and increased at different rates in 2019. \*\*\*

Lets group by the year and get the mean values over the last five years for each stock value.

Value3 <- Value2[,c(1:53,112,115)]  
  
yearMeans <- Value3 %>% group\_by(Year) %>%  
 summarise\_at(vars(as.vector(colnames(Value3)[1:53])), mean)  
  
yearMeansTidy <- gather(yearMeans,'Stock','YearMeanValue',2:54)  
  
stock5yrMeans <- yearMeansTidy %>% group\_by(Stock) %>%  
 summarise\_at(vars(as.vector(colnames(yearMeansTidy)[3])), mean)  
colnames(stock5yrMeans)[2] <- 'stock5yrMeans'

Stock5year <- merge(stock5yrMeans,yearMeansTidy, by.x='Stock', by.y='Stock')

stock5yrOrdered <- Stock5year[with(Stock5year, order(Stock, Year)),]

Lets add a field that shows if the stock had an increase of 10% during the year and a field that shows if it decreased

ymn <- stock5yrOrdered$YearMeanValue  
YMN <- c(ymn[1],ymn[1:length(ymn)-1])  
  
stc2 <- stock5yrOrdered$Stock  
STC2 <- c('xyz',stc2[1:length(stc2)-1])  
  
STC3 <- ifelse(stc2==STC2, 1,0)  
  
stock5yrOrdered$Direction5yr10PercentChange <- ifelse(STC3==1 & stock5yrOrdered$YearMeanValue-YMN > .10\*YMN,'up10',  
 ifelse(STC3==1 & stock5yrOrdered$YearMeanValue-YMN <= -0.10\*YMN, 'down10',  
 ifelse(STC3==1 & stock5yrOrdered$YearMeanValue-YMN <= 0, 'down', ifelse(STC3==1 & stock5yrOrdered$YearMeanValue-YMN > 0, 'up', ''))))  
  
show1 <- cbind(head(stock5yrOrdered), tail(stock5yrOrdered))  
show1

## Stock stock5yrMeans Year YearMeanValue Direction5yr10PercentChange  
## 2 AAL.Close 38.67371 2015 45.12210   
## 3 AAL.Close 38.67371 2016 38.18385 down10  
## 4 AAL.Close 38.67371 2017 47.49072 up10  
## 1 AAL.Close 38.67371 2018 42.80195 down  
## 5 AAL.Close 38.67371 2019 30.87933 down10  
## 6 AAL.Close 38.67371 2020 27.56429 down10  
## Stock stock5yrMeans Year YearMeanValue Direction5yr10PercentChange  
## 2 XOM.Close 78.737 2015 82.82845   
## 3 XOM.Close 78.737 2016 86.21968 up  
## 4 XOM.Close 78.737 2017 81.86159 down  
## 1 XOM.Close 78.737 2018 79.95570 down  
## 5 XOM.Close 78.737 2019 73.73464 down  
## 6 XOM.Close 78.737 2020 67.82191 down

Lets get these subsets of stocks that within the time span of 2015-2020 increased by more than 10% annually, decreased by 10% or more annually, decreased, or increased.

Stocks10PercentAnnualDecrease2015\_2020 <- subset(stock5yrOrdered, stock5yrOrdered$Direction5yr10PercentChange=='down10')  
  
stocks10Decr <- Stocks10PercentAnnualDecrease2015\_2020 %>% group\_by(Stock) %>% count(n=n())  
colnames(stocks10Decr)[2] <- 'nTimesDecr10\_5yr'  
stocks10Decr <- stocks10Decr[,-3]  
  
Stocks10PercentAnnualIncrease2015\_2020 <- subset(stock5yrOrdered, stock5yrOrdered$Direction5yr10PercentChange=='up10')  
  
stocks10Incr <- Stocks10PercentAnnualIncrease2015\_2020 %>% group\_by(Stock) %>% count(n=n())  
colnames(stocks10Incr)[2] <- 'nTimesIncr10\_5yr'  
stocks10Incr <- stocks10Incr[,-3]  
  
StocksAnnualIncrease2015\_2020 <- subset(stock5yrOrdered, stock5yrOrdered$Direction5yr10PercentChange=='up')  
  
StocksIncrZerobase <- StocksAnnualIncrease2015\_2020 %>% group\_by(Stock) %>% count(n=n())  
colnames(StocksIncrZerobase)[2] <- 'nTimesIncrFromZero\_5yrs'  
StocksIncrZerobase <- StocksIncrZerobase[,-3]  
   
StocksAnnualDecrease2015\_2020 <- subset(stock5yrOrdered, stock5yrOrdered$Direction5yr10PercentChange=='down')  
  
StocksDecrZerobase <- StocksAnnualDecrease2015\_2020 %>% group\_by(Stock) %>% count(n=n())  
colnames(StocksDecrZerobase)[2] <- 'nTimesDecrFromZero\_5yrs'  
StocksDecrZerobase <- StocksDecrZerobase[,-3]

Lets merge these sets together with outer joins.

Stocks5yrChanges\_outerJoin <- merge(stocks10Decr,stocks10Incr, by.x='Stock', by.y='Stock', all=TRUE)  
  
Stocks5yrChanges\_outerJoin1 <- merge(Stocks5yrChanges\_outerJoin,StocksDecrZerobase, by.x='Stock', by.y='Stock', all=TRUE)  
  
Stocks5yrChanges\_outerJoin2 <- merge(Stocks5yrChanges\_outerJoin1,StocksIncrZerobase, by.x='Stock', by.y='Stock', all=TRUE)  
  
stock\_5yr\_stats\_2015\_2020 <- merge(stock5yrOrdered,Stocks5yrChanges\_outerJoin2, by.x='Stock', by.y='Stock', all=TRUE)

Write this file out to analyze those stocks having decreased and increased the most in the last 5 years.

write.csv(stock\_5yr\_stats\_2015\_2020,'stocks\_STATS\_N\_Changes.csv', row.names=FALSE)

Lets attach the stock name to this data set above by reading in the file with the names on it when hand picking these stocks by searching manually in finance.yahoo.com.

stockNames <- read.csv('yahooStockBasket.csv', header=T, sep=',', na.strings=c('',' '))  
stock\_5yr\_stats\_2015\_2020$Stock <- gsub('[.]Close','', stock\_5yr\_stats\_2015\_2020$Stock)  
stock\_5yr\_stats\_2015\_2020$Stock <- as.factor(stock\_5yr\_stats\_2015\_2020$Stock)  
StockNames\_STATS\_2015\_2020 <- merge(stockNames,stock\_5yr\_stats\_2015\_2020,  
 by.x='stock', by.y='Stock')  
  
StockNames\_STATS\_2015\_2020$nTimesDecr10\_5yr <-  
 ifelse(is.na(StockNames\_STATS\_2015\_2020$nTimesDecr10\_5yr==TRUE),  
 0,StockNames\_STATS\_2015\_2020$nTimesDecr10\_5yr)  
  
StockNames\_STATS\_2015\_2020$nTimesIncr10\_5yr <-  
 ifelse(is.na(StockNames\_STATS\_2015\_2020$nTimesIncr10\_5yr==TRUE),  
 0,StockNames\_STATS\_2015\_2020$nTimesIncr10\_5yr)  
  
StockNames\_STATS\_2015\_2020$nTimesDecrFromZero\_5yrs <-  
 ifelse(is.na(StockNames\_STATS\_2015\_2020$nTimesDecrFromZero\_5yrs==TRUE),  
 0,StockNames\_STATS\_2015\_2020$nTimesDecrFromZero\_5yrs)  
  
StockNames\_STATS\_2015\_2020$nTimesIncrFromZero\_5yrs <-  
 ifelse(is.na(StockNames\_STATS\_2015\_2020$nTimesIncrFromZero\_5yrs==TRUE),  
 0,StockNames\_STATS\_2015\_2020$nTimesIncrFromZero\_5yrs)  
  
StockNames\_STATS\_2015\_2020$Direction5yr10PercentChange <-  
 ifelse(StockNames\_STATS\_2015\_2020$Direction5yr10PercentChange=='',0,StockNames\_STATS\_2015\_2020$Direction5yr10PercentChange)  
  
write.csv(StockNames\_STATS\_2015\_2020, 'StockNames\_STATS\_2015\_2020.csv', row.names=FALSE)  
  
show2 <- rbind(head(StockNames\_STATS\_2015\_2020,3),tail(StockNames\_STATS\_2015\_2020,3))  
show2

## stock  
## 1 AAL  
## 2 AAL  
## 3 AAL  
## 310 XOM  
## 311 XOM  
## 312 XOM  
## stockInfo  
## 1 American Airlines Group Inc. (AAL)\nNasdaqGS - NasdaqGS Real Time Price. Currency in USD  
## 2 American Airlines Group Inc. (AAL)\nNasdaqGS - NasdaqGS Real Time Price. Currency in USD  
## 3 American Airlines Group Inc. (AAL)\nNasdaqGS - NasdaqGS Real Time Price. Currency in USD  
## 310 Exxon Mobil Corporation (XOM)\nNYSE - NYSE Delayed Price. Currency in USD  
## 311 Exxon Mobil Corporation (XOM)\nNYSE - NYSE Delayed Price. Currency in USD  
## 312 Exxon Mobil Corporation (XOM)\nNYSE - NYSE Delayed Price. Currency in USD  
## stockExchange stock5yrMeans Year YearMeanValue Direction5yr10PercentChange  
## 1 Nasdaq 38.67371 2018 42.80195 down  
## 2 Nasdaq 38.67371 2017 47.49072 up10  
## 3 Nasdaq 38.67371 2020 27.56429 down10  
## 310 NYSE 78.73700 2016 86.21968 up  
## 311 NYSE 78.73700 2020 67.82191 down  
## 312 NYSE 78.73700 2019 73.73464 down  
## nTimesDecr10\_5yr nTimesIncr10\_5yr nTimesDecrFromZero\_5yrs  
## 1 3 1 1  
## 2 3 1 1  
## 3 3 1 1  
## 310 0 0 4  
## 311 0 0 4  
## 312 0 0 4  
## nTimesIncrFromZero\_5yrs  
## 1 0  
## 2 0  
## 3 0  
## 310 1  
## 311 1  
## 312 1

Lets the mean annual unemployment rates using the original table to combine with this table of the n times a stock increases/decreases per year in the last five years.

ue$Annual <- round(rowMeans(ue[,2:13], na.rm=T),2)  
ue\_15\_20 <- ue[9:14,c(1,14)]  
colnames(ue\_15\_20)[2] <- 'Annual\_UE'

Now, combine the unemployment and the newest stats with counts table.

stock\_5yrs\_ue <- merge(ue\_15\_20,StockNames\_STATS\_2015\_2020, by.x='Year', by.y='Year')

Add in a boolean field to show if the YearMeanValue is greater than the Stock5yrMeans column as a 1 if true and a 0 if not.

stock\_5yrs\_ue$YearMeanGreaterThan5yrMean <- ifelse(stock\_5yrs\_ue$YearMeanValue >  
 stock\_5yrs\_ue$stock5yrMeans,1,0)

write.csv(stock\_5yrs\_ue,'stock\_2015-2020\_ue.csv',row.names=FALSE)

Make separate portfolios for each of the stocks that increased by more than 10% annually more than at least 1 time, decreased more than 10% annually more than at least 1 time, then get the mean value of the YearMeanValue column. Compare this to the portfolio of the stocks that never decreased more than 10% annually.

sub\_D10 <- subset(StockNames\_STATS\_2015\_2020, StockNames\_STATS\_2015\_2020$nTimesDecr10\_5yr > 0)  
  
sub\_nvr\_D10 <- subset(StockNames\_STATS\_2015\_2020, StockNames\_STATS\_2015\_2020$nTimesDecr10\_5yr == 0)  
  
pMean\_sub\_D10 <- mean(sub\_D10$stock5yrMeans)  
pMean\_sub\_D10

## [1] 41.81591

pMean\_sub\_nvr\_D10 <- mean(sub\_nvr\_D10$stock5yrMeans)  
pMean\_sub\_nvr\_D10

## [1] 191.595

pMean\_sub\_nvr\_D10\*length(unique(sub\_nvr\_D10$stock))

## [1] 4023.495

Lets now do the reverse and look at those stocks that increased more than 10% at least one time in the last five years of 2015-2020 and compare the means.

sub\_I10 <- subset(StockNames\_STATS\_2015\_2020, StockNames\_STATS\_2015\_2020$nTimesIncr10\_5yr > 3)  
  
sub\_nvr\_I10 <- subset(StockNames\_STATS\_2015\_2020, StockNames\_STATS\_2015\_2020$nTimesIncr10\_5yr == 0)  
  
pMean\_sub\_I10 <- mean(sub\_I10$stock5yrMeans)  
pMean\_sub\_I10

## [1] 204.6173

pMean\_sub\_nvr\_I10 <- mean(sub\_nvr\_I10$stock5yrMeans)  
pMean\_sub\_nvr\_I10

## [1] 32.77691

pMean\_sub\_I10\*length(unique(sub\_I10$stock))

## [1] 1432.321

Now lets look at those stocks that increased at least one time in the last five years but never by more than 10% to compare the portfolio mean to those above that decreased over the last five years but never by more than 10%.

sub\_I10 <- subset(StockNames\_STATS\_2015\_2020, StockNames\_STATS\_2015\_2020$nTimesIncrFromZero\_5yr > 0)  
  
sub\_nvr\_I10 <- subset(StockNames\_STATS\_2015\_2020, StockNames\_STATS\_2015\_2020$nTimesIncrFromZero\_5yr == 0)  
  
pMean\_sub\_I0 <- mean(sub\_I10$YearMeanValue)  
pMean\_sub\_I0

## [1] 129.7607

pMean\_sub\_nvr\_I0 <- mean(sub\_nvr\_I10$YearMeanValue)  
pMean\_sub\_nvr\_I0

## [1] 50.44023

Lets get the portfolio mean value in 2015 and compare to the portfolio mean value in 2020 of all stocks in this set to compare the above values to.

p2015 <- subset(StockNames\_STATS\_2015\_2020, StockNames\_STATS\_2015\_2020$Year==2015)  
p2020 <- subset(StockNames\_STATS\_2015\_2020, StockNames\_STATS\_2015\_2020$Year==2020)  
pm2015 <- mean(p2015$YearMeanValue)  
pm2020 <- mean(p2020$YearMeanValue)  
  
pm2015

## [1] 73.33239

pm2020

## [1] 133.2523

pm2015\*length(unique(StockNames\_STATS\_2015\_2020$stock))

## [1] 3813.284

pm2020\*length(unique(StockNames\_STATS\_2015\_2020$stock))

## [1] 6929.117

The portfolio mean was 73 USD in 2015 and 133 USD in 2020. Knowing this combined with the other last three clues to stock volatility (by analyzing the volatility by year using the number of times the stock has an increase or decrease by 10% during each of the last five years) we can say that all of these stocks without carefully selecting them did increase the whole portfolio in the last five years from 73 USD to 133 USD, but the best set of stocks are the ones that increased at least three times during the last five years more than 10% annually, because the portfolio of that set of stocks was 204 USD average value of the stock with a small portfolio subset for only those stocks at 1432 USD and the subset that never decreased by more than 10% annually over the last five years had an average value of 192 USD and a price for that subset of stocks as a portfolio of 4023 USD.

What would the ROI be for all stocks compared to only those stocks that increased at least three times during the last five years by more than 10 per cent?

incr3\_10 <- subset(sub\_I10, sub\_I10$Year==2015)  
mean\_incr3\_10 <- mean(incr3\_10$YearMeanValue)  
ROI\_all <- pm2020/pm2015  
ROI\_Incr3\_10 <- pMean\_sub\_I10/mean\_incr3\_10  
ROI\_all

## [1] 1.8171

ROI\_Incr3\_10

## [1] 2.420963

The **return on investment** is almost doubled over five years selecting all the stocks in this portfolio of stocks that had values from 2007-2020 (done in the beginning is 53 out of 65) with a value of 1.817 or **182 per cent**. While the portfolio with only the stocks that never decreased more than 10% over the last five years had an ROI of 1.74 or **174 per cent**.

Either way, these stocks returned a profit over the span of the last five years. We could see what the [annual ROI rate](https://www.investopedia.com/terms/a/apr.asp) is knowing the amount of 156 USD for best portfolio and 133 USD for all stocks, compared to the initial value of 73 USD.

n =365\*5  
principalPlusFees <- pMean\_sub\_nvr\_D10  
initial <- pm2015  
apr <- ((principalPlusFees/initial)/n)\*365\*100  
apr

## [1] 52.25385

Using the formula from the web, the apr on this portfolio is 52.25%, which seems really high. This is the average value over the last five years of all the stocks in the subset of the portfolio having never decreased by 10%.

There are four sets of counts for those that increased more than 10%, decreased more than 10%, increased more than zero but less than 10%, and decreased more than zero but less than 10% within the five year span from 2015-2020. Lets see if there is a better subset of choices for a better market portfolio.

Lets add a five year poisson column using lambda=(unemployment rate), time=(1), and k=(YearMeanGreaterThan5yrMean).

ue2 <- stock\_5yrs\_ue$Annual\_UE  
t <- 1  
k <- stock\_5yrs\_ue$YearMeanGreaterThan5yrMean  
stock\_5yrs\_ue$poisson5yrUE <- round((exp(-ue2\*t)\*(ue2\*t)^k)/(factorial(k)),5)

Lets get a subset of those stocks that have cyclical patterns within five years, so that we have three years the stock increases more than 10% exactly 3 times, and two years where the stock decreases less than 10% exactly 2 times. Separately, get the stocks it increases greater than 10% exactly 3 times, and decreases more than 10% exactly 2 times. Also get the reverse of these values

cyclical <- subset(stock\_5yrs\_ue, stock\_5yrs\_ue$nTimesIncr10\_5yr==3 & (stock\_5yrs\_ue$nTimesDecr10\_5yr==2 | stock\_5yrs\_ue$nTimesDecrFromZero\_5yrs==2))  
  
cyclical2 <- subset(stock\_5yrs\_ue, stock\_5yrs\_ue$nTimesIncrFromZero\_5yrs >=2 & (stock\_5yrs\_ue$nTimesDecr10\_5yr >= 2 | stock\_5yrs\_ue$nTimesDecrFromZero\_5yrs >= 2))  
  
c1 <- as.character(unique(cyclical$stock))  
c2 <- as.character(unique(cyclical2$stock))  
cycle <- c(c1,c2)  
cycle1 <- as.data.frame(cycle)  
colnames(cycle1) <- 'Stock'  
  
portCycle <- merge(cycle1,stock\_5yrs\_ue, by.x='Stock', by.y='stock')  
portCycle\_2015 <- subset(portCycle, Year==2015)  
portCycle\_2020 <- subset(portCycle, Year==2020)  
  
mean(portCycle\_2015$YearMeanValue)

## [1] 36.228

mean(portCycle\_2020$YearMeanValue)

## [1] 37.91673

mean(portCycle$YearMeanValue)

## [1] 36.02808

The above shows that the **cyclical stocks that have highs and lows the time span of the loan aren’t great investments**, as these ones only profited $1.70 over five years. We could see if we dropped these stocks if they would make the portfolio better profit wise.

m1 <- grep('MGM', stock\_5yrs\_ue$stock)  
m2 <- grep('ONCY', stock\_5yrs\_ue$stock)  
m3 <- grep('ARWR', stock\_5yrs\_ue$stock)  
m4 <- grep('S', stock\_5yrs\_ue$stock)  
m5 <- grep('JBLU', stock\_5yrs\_ue$stock)  
m6 <- grep('CVX', stock\_5yrs\_ue$stock)  
m7 <- grep('WFC', stock\_5yrs\_ue$stock)  
m8 <- grep('HRB', stock\_5yrs\_ue$stock)  
m9 <- grep('HMC', stock\_5yrs\_ue$stock)  
m10 <- grep('RRGB', stock\_5yrs\_ue$stock)  
m11 <- grep('LUV', stock\_5yrs\_ue$stock)  
m12 <- grep('UBSI', stock\_5yrs\_ue$stock)  
c3 <- c(m1,m2,m3,m4,m5,m6,m7,m8,m9,m10,m11,m12)  
stock\_5yrs\_ue\_dropCycle <- stock\_5yrs\_ue[-c3,]  
mean(stock\_5yrs\_ue\_dropCycle$YearMeanValue)

## [1] 134.3824

The mean of the portfolio without the cyclical stocks is 134 USD. This is not much better than all of the stock in the portfolio of 133 USD.

Lets get their poisson values for the cyclical stock, and drop any stock with this median value from our set of stocks, and compare the value of the portfolio after dropping stock with this poisson value.

summary(portCycle$poisson5yrUE)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00509 0.02045 0.02689 0.03908 0.05658 0.09837

The median poisson is 0.02689. Lets drop any stock within the range of 0.025 and 0.03.

stock\_5yrs\_ue\_drop <- subset(stock\_5yrs\_ue, stock\_5yrs\_ue$poisson5yrUE < 0.025 |   
 stock\_5yrs\_ue$poisson5yrUE > 0.03)  
unique(stock\_5yrs\_ue\_drop$stock)

## [1] COST CVX NUS HD MSFT DLTR MGM KGJI JPM ADDYY HOFT C   
## [13] NFLX AMZN TGT ARWR S GOOG NKE TJX WMT JNJ VZ WM   
## [25] LUV ROST T FFIN WWE TEVA XOM AAL ONCY AAP NSANY HRB   
## [37] PCG HMC HST SIG INO RRGB WFC EPD JWN UBSI TM JBLU   
## [49] FTR M KSS F   
## 65 Levels: AAL AAP ADDYY AMC AMZN ARWR ASCCY C COST CSSEP CVX DLTR EPD ... YELP

mean(stock\_5yrs\_ue\_drop$YearMeanValue)

## [1] 121.3385

Removing the stock with poisson values in the range of the cyclical stock median value is worse than removing just the cyclical stock at 121 USD portfolio value compared to 134 USD.

There is more to do and work with after realizing I rushed through it. It will be updated with more accurate information as it develops.