Subset Large Stock Data Counts ROI Group Counts

Janis Corona

3/22/2020

%%%%%%%%%%%%%%%

This file relies on all others made, but can be ran using the kaggle large data set. To understand the file, it would be helpful to look at all files in the github folder. Both links are below.

library(lubridate)

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:lubridate':  
##   
## intersect, setdiff, union

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

%%%%%%%%%%%%%%%

Retrieve the large (approximately 442 mb file size) from [Kaggle](https://www.kaggle.com/janiscorona/stock-day-counts-increasing-and-decreasing-0720) called allStocksGathered1.csv for the large data processed in the last part of this script and to get the individual stock stats and counts of increasing and decreasing days.This script is in [github](https://github.com/JanJanJan2018/Quantative-Finance) as newStocksLagsCountsGroups.Rmd that made the csv large data file just mentioned.

allStocksGathered1 <- read.csv('allStocksGathered1.csv',sep=',', header=TRUE,  
 na.strings=c('',' ','NA'))

This data set is also needed and can be retrieved at the github link above.

newStocks <- read.csv('StockSwingTradeBotCom.csv', header=TRUE, na.strings=c('',' ','NA'),  
 sep=',')  
  
head(newStocks)

## stockSymbol stockName dailyClose volatility avgVolume  
## 1 A Agilent Technolog... 65.76 34.988 2,515,221  
## 2 AA Alcoa Inc. 5.83 121.188 7,431,897  
## 3 AAC AAC Holdings, Inc. 0.48 99.106 279,165  
## 4 AACG ATA Creativity Gl... 0.74 45.638 25,584  
## 5 AAL American Airlines... 10.29 97.428 24,309,238  
## 6 AAMC Altisource Asset ... 11.12 91.945 7,226

Lets create the lag and count/group by counts fields to analyze by each stock, using subsets of each stock.

DF <- allStocksGathered1  
DF$Date <- as.Date(DF$Date)

What stock ticker are you interested in? And what lag are you interested in? You can keep the lag from the top of this script or these defaults or change them here. You need to enter the stock name in the chunk for Rmarkdown below. Also, pick a start date and end date of the time you want to select for running the counts and groups of counts.

stock\_1 <- toupper("ubsi")  
lag <- 3  
startDate <- '2005-01-22'  
endDate <- '2008-01-22'

Lets subset our large table to get the dates requested.

sDF <- subset(DF, DF$Date>=startDate & DF$Date<=endDate)

lagN <- paste('lag',lag,sep='')  
cat('The number of days to retrieve the stock value compared to each day value listed as an instance is ',lag,'and the stock to look up this information for is ',stock\_1)

## The number of days to retrieve the stock value compared to each day value listed as an instance is 3 and the stock to look up this information for is UBSI

stknme <- as.character(newStocks[newStocks$stockSymbol==stock\_1,2])  
cat('\nThis stock is ',stknme)

##   
## This stock is United Bankshares...

Lstock\_1 <- subset(sDF, sDF$stockName==stock\_1)  
  
cat('The number of days for trading that this time period will provide counts of increasing and decreasing days is ', length(Lstock\_1$Date), ' trading days.')

## The number of days for trading that this time period will provide counts of increasing and decreasing days is 754 trading days.

Generic automation of above stock to look up and the lag to use for generating the counts, group of counts, and lag values to get those counts.

Lstock\_1$startDayValue <-Lstock\_1$stockValue[1]  
Lstock\_1$startDayDate <- Lstock\_1$Date[1]  
Lstock\_1$finalDayValue <-Lstock\_1$stockValue[length(Lstock\_1$stockValue)]  
Lstock\_1$finalDayDate <- Lstock\_1$Date[length(Lstock\_1$Date)]  
  
stock\_1LN <- lag(Lstock\_1$stockValue, lag)  
Lstock\_1$lagN <- stock\_1LN  
Lstock\_1$today2\_lagN <- Lstock\_1$stockValue/Lstock\_1$lagN  
Lstock\_1 <- Lstock\_1[complete.cases(Lstock\_1),]

Lets look at the data we will be adding counts of increasing and decreasing days, for the time interval dates and stock values at the beginning and end of the time interval available or requested.

cat('\nThe lag for this table was for ',lag,'days.','\nThe stock to look up was ', stock\_1,'.\nThe start date of this stock and starting value was ',as.character(paste(Lstock\_1$startDayDate[1])),' and ','$',Lstock\_1$startDayValue[1],'\nThe end date and end date price of this stock analysis is ',as.character(paste(Lstock\_1$finalDayDate[1])),' and ','$',Lstock\_1$finalDayValue[1])

##   
## The lag for this table was for 3 days.   
## The stock to look up was UBSI .  
## The start date of this stock and starting value was 2005-01-24 and $ 33.97   
## The end date and end date price of this stock analysis is 2008-01-22 and $ 26.34

roi <- Lstock\_1$finalDayValue[1]/Lstock\_1$startDayValue[1]  
cat('\nThe return on investment as a percentage of the amount invested for this time period is ',roi)

##   
## The return on investment as a percentage of the amount invested for this time period is 0.7753901

cat('\n\nIn dollars initially invested your return is $',Lstock\_1$finalDayValue[1]-Lstock\_1$startDayValue[1], 'for the dates:',  
 as.character(paste(Lstock\_1$startDayDate[1])),'through ',  
 as.character(paste(Lstock\_1$finalDayDate[1])))

##   
##   
## In dollars initially invested your return is $ -7.63 for the dates: 2005-01-24 through 2008-01-22

Now, lets look at the counts and group counts of increasing and decreasing days for this stock and the time period available.

#assign a 1 to increasing values  
Lstock\_1$todayGrtrThan\_lagN <- ifelse(Lstock\_1$today2\_lagN>1, 1,0)  
  
Lstock\_1$cumulativeSumTodayGrtrThan\_lagN <- cumsum(Lstock\_1$todayGrtrThan\_lagN)  
  
# get the count of how many instances repeat,   
# those counts repeating are counts that measure the days cumulatively decreasing  
# those cumulative counts that don't repeat, are counting increasing days.  
# These are stock values for today's value to 7 days prior value.  
countstock\_10 <- Lstock\_1 %>% group\_by(cumulativeSumTodayGrtrThan\_lagN) %>% count(n=n())  
countstock\_10 <- as.data.frame(countstock\_10)  
countstock\_10 <- countstock\_10[,-3]  
colnames(countstock\_10)[2] <- 'nRepeatsTodayGrtrThan\_lagN'  
  
# Count the REPEATS of each number (minus the initial start)   
countstock\_10$decrDaysThisCycle <- countstock\_10$n-1  
  
# Count the number of times the cycle count repeats in this time span exactly that many days  
countstock\_10b <- countstock\_10 %>% group\_by(decrDaysThisCycle) %>% count(n=n())  
countstock\_10b <- as.data.frame(countstock\_10b)  
countstock\_10b <- countstock\_10b[,-3]  
colnames(countstock\_10b)[2] <- 'nTimesDecrDayCountsOccurs'  
  
#combine these two count matrices of decreasing days  
countsstock\_tableDecr <- merge(countstock\_10, countstock\_10b, by.x='decrDaysThisCycle',  
 by.y='decrDaysThisCycle')  
  
#combine the counts to the stock subset  
stock\_3 <- merge(Lstock\_1, countsstock\_tableDecr, by.x='cumulativeSumTodayGrtrThan\_lagN',  
 by.y='cumulativeSumTodayGrtrThan\_lagN')  
  
#assign a 1 to decreasing values  
stock\_3$todayLessThan\_lagN <- ifelse(stock\_3$today2\_lagN>1, 0,1)  
  
stock\_3$cumulativeSumTodayLessThan\_lagN <- cumsum(stock\_3$todayLessThan\_lagN)  
  
# get the count of how many instances repeat,   
# those counts repeating are counts that measure the days cumulatively increasing  
# those cumulative counts that don't repeat, are counting decreasing days.  
# These are stock values for today's value to 7 days prior value.  
countstock\_11 <- stock\_3 %>% group\_by(cumulativeSumTodayLessThan\_lagN) %>% count(n=n())  
countstock\_11 <- as.data.frame(countstock\_11)  
countstock\_11 <- countstock\_11[,-3]  
colnames(countstock\_11)[2] <- 'nRepeatsTodayLessThan\_lagN'  
  
# Count the REPEATS of each number (minus the initial start)   
countstock\_11$incrDaysThisCycle <- countstock\_11$n-1  
  
# Count the number of times the cycle count repeats in this time span exactly that many days  
countstock\_11b <- countstock\_11 %>% group\_by(incrDaysThisCycle) %>% count(n=n())  
countstock\_11b <- as.data.frame(countstock\_11b)  
countstock\_11b <- countstock\_11b[,-3]  
colnames(countstock\_11b)[2] <- 'nTimesIncrDayCountsOccurs'  
  
#combine these two count matrices of decreasing days  
countsstock\_tableIncr <- merge(countstock\_11, countstock\_11b,  
 by.x='incrDaysThisCycle',  
 by.y='incrDaysThisCycle')  
  
#combine the counts to the stock subset  
stock\_4 <- merge(stock\_3, countsstock\_tableIncr,  
 by.x='cumulativeSumTodayLessThan\_lagN',  
 by.y='cumulativeSumTodayLessThan\_lagN')

colnames(stock\_4)

## [1] "cumulativeSumTodayLessThan\_lagN" "cumulativeSumTodayGrtrThan\_lagN"  
## [3] "Date" "stockName"   
## [5] "stockValue" "startDayValue"   
## [7] "startDayDate" "finalDayValue"   
## [9] "finalDayDate" "lagN"   
## [11] "today2\_lagN" "todayGrtrThan\_lagN"   
## [13] "decrDaysThisCycle" "nRepeatsTodayGrtrThan\_lagN"   
## [15] "nTimesDecrDayCountsOccurs" "todayLessThan\_lagN"   
## [17] "incrDaysThisCycle" "nRepeatsTodayLessThan\_lagN"   
## [19] "nTimesIncrDayCountsOccurs"

stock\_5 <- stock\_4[,c(3:11,  
 12,2,14,13,15,  
 16,1,18,17,19)]  
colnames(stock\_5) <- gsub('lagN',lagN,colnames(stock\_5))  
colnames(stock\_5)

## [1] "Date" "stockName"   
## [3] "stockValue" "startDayValue"   
## [5] "startDayDate" "finalDayValue"   
## [7] "finalDayDate" "lag3"   
## [9] "today2\_lag3" "todayGrtrThan\_lag3"   
## [11] "cumulativeSumTodayGrtrThan\_lag3" "nRepeatsTodayGrtrThan\_lag3"   
## [13] "decrDaysThisCycle" "nTimesDecrDayCountsOccurs"   
## [15] "todayLessThan\_lag3" "cumulativeSumTodayLessThan\_lag3"  
## [17] "nRepeatsTodayLessThan\_lag3" "incrDaysThisCycle"   
## [19] "nTimesIncrDayCountsOccurs"

Using this information on one stock of the thousands available in our large csv file and table, lets return the count information and the number of times this stock has seen those exact days of counts.

cat('\nThe number of times this stock has decreased in the current cycle from the start of this time period retrieved in price comparison of the number of days in lags retrieved prior to each instance dates\' stock value is ', stock\_5$decrDaysThisCycle[length(stock\_5$decrDaysThisCycle)],'\n')

##   
## The number of times this stock has decreased in the current cycle from the start of this time period retrieved in price comparison of the number of days in lags retrieved prior to each instance dates' stock value is 0

cat('\nThe number of times this stock has increased in the current cycle from the start of this time period retrieved in price comparison of the number of days in lags retrieved prior to each instance dates\' stock value is ', stock\_5$incrDaysThisCycle[length(stock\_5$incrDaysThisCycle)],'\n')

##   
## The number of times this stock has increased in the current cycle from the start of this time period retrieved in price comparison of the number of days in lags retrieved prior to each instance dates' stock value is 1

cat('\nThe number of times this stock has decreased exactly this number of days compared to its price ', lag, ' days ago, is ', stock\_5$nTimesDecrDayCountsOccurs[length(stock\_5$nTimesDecrDayCountsOccurs)],'\n')

##   
## The number of times this stock has decreased exactly this number of days compared to its price 3 days ago, is 236

cat('\nThe number of times this stock has increased exactly this number of days compared to its price ', lag, ' days ago, is ', stock\_5$nTimesIncrDayCountsOccurs[length(stock\_5$nTimesIncrDayCountsOccurs)],'\n')

##   
## The number of times this stock has increased exactly this number of days compared to its price 3 days ago, is 35

The unique number of days this stock selected (for the time period retrieved) decreased is shown in the table below.

stock\_5[unique(stock\_5$decrDaysThisCycle),c(1:3,8,9,13,14)]

## Date stockName stockValue lag3 today2\_lag3 decrDaysThisCycle  
## 1 2005-01-27 UBSI 34.17 33.97 1.0058875 0  
## 11 2005-02-10 UBSI 35.02 35.97 0.9735891 11  
## 3 2005-01-31 UBSI 34.10 34.45 0.9898403 1  
## 7 2005-02-04 UBSI 35.75 34.83 1.0264140 0  
## 2 2005-01-28 UBSI 33.98 33.70 1.0083086 1  
## 4 2005-02-01 UBSI 34.83 34.17 1.0193152 0  
## 5 2005-02-02 UBSI 35.25 33.98 1.0373749 0  
## 8 2005-02-07 UBSI 35.97 35.25 1.0204255 0  
## 19 2005-02-23 UBSI 33.20 34.03 0.9756098 11  
## 9 2005-02-08 UBSI 36.11 35.22 1.0252697 11  
## 10 2005-02-09 UBSI 35.03 35.75 0.9798601 11  
## 6 2005-02-03 UBSI 35.22 34.10 1.0328446 0  
## 16 2005-02-17 UBSI 34.03 34.90 0.9750716 11  
## nTimesDecrDayCountsOccurs  
## 1 236  
## 11 2  
## 3 31  
## 7 236  
## 2 31  
## 4 236  
## 5 236  
## 8 236  
## 19 2  
## 9 2  
## 10 2  
## 6 236  
## 16 2

The unique number of days this stock selected (for the time period retrieved) increased is shown in the table below.

stock\_5[unique(stock\_5$incrDaysThisCycle),c(1:3,8,9,18,19)]

## Date stockName stockValue lag3 today2\_lag3 incrDaysThisCycle  
## 1 2005-01-27 UBSI 34.17 33.97 1.0058875 1  
## 6 2005-02-03 UBSI 35.22 34.10 1.0328446 6  
## 3 2005-01-31 UBSI 34.10 34.45 0.9898403 6  
## 4 2005-02-01 UBSI 34.83 34.17 1.0193152 6  
## 2 2005-01-28 UBSI 33.98 33.70 1.0083086 1  
## 5 2005-02-02 UBSI 35.25 33.98 1.0373749 6  
## 8 2005-02-07 UBSI 35.97 35.25 1.0204255 6  
## 7 2005-02-04 UBSI 35.75 34.83 1.0264140 6  
## 9 2005-02-08 UBSI 36.11 35.22 1.0252697 6  
## nTimesIncrDayCountsOccurs  
## 1 35  
## 6 12  
## 3 12  
## 4 12  
## 2 35  
## 5 12  
## 8 12  
## 7 12  
## 9 12

%%%%%%%%%%%%%%%%%