15.458 Project 1

Ira Li, Edward Liu and Ellie Liu

9/14/2018

1.Read the article from the New York Times, “Market Swings Are Becoming New Standard” (9/12/2011, link available on Canvas) and then use the Bloomberg terminal to get OHLC time series data for the S&P 500 index to answer the questions below. In the Bloomberg Excel “Spreadsheet Builder,” these data fields are found under “Market Activity” and designated as Open (PX OPEN), High (PX HIGH), Low (PX LOW) and Last (PX LAST), respectively.) Get the longest set of daily data available.

```
library(quantmod)
```

```
## Loading required package: xts
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: TTR
```

```
## Version 0.4-0 included new data defaults. See ?getSymbols.
```

```
suppressMessages(library("tidyverse"))
```

```
mydata = read.csv("/Users/EllieLiu/Desktop/2018 Fall/Financial Data/Proj 1/S&P500.csv")
```

a) What date range did you obtain? Is your data set complete?

The obtained data ranges from 12/30/1927 to today.The data set is not complete in the sense that open, high, low and close prices are all equal before 1982.

b) List data integrity checks that can be performed on the data. Apply them. What errors, if any, did you find? Which errors, if any, could be corrected without using alternative data sources?

```
#there is no NAs
any(is.na(mydata))
```

```
## [1] FALSE
```

```
#All data are positive values
table(sign(mydata[,2]))
```

```
##
##     1
## 22781

table(sign(mydata[,3]))

##
##     1
## 22781

table(sign(mydata[,4]))

##
##     1
## 22781

table(sign(mydata[,5]))

##
##     1
## 22781

notrade <- function(n){
  count = 0

  for (i in 1:n){
  if (mydata[i, 3] == mydata[i, 4]){
    count = count + 1
    }
  }
  count
}
notrade(nrow(mydata))

## [1] 13604

trade_overnight <- function(n){
  count = 0

  for (i in 1:n){
  if (mydata[i, 5] != mydata[i+1, 2]){
    count = count + 1
    }
  }
  count
}

trade_overnight(nrow(mydata)-1)

## [1] 18085
```

By performing data integrity check, we found no NAs and all data positive. There are 13604 days with Low = High, which means that no trade has been made throughout the day or all stocks are traded at the same price. We corrected the error by removing the first 13604 data from 12/30/1927 to 04/20/1982. We also found that for 18085 days, the previous closing price does not match with following opening price. However, the listed closing price is the last trading price during the business hours, and the exchange does not lock in the price until the following morning. Thus, it is reasonable for the open to slightly deviate from the previous close.

c) Based on your data set, estimate the probabilities that the market's daily high value occurs at the open or at the close, respectively. Do the same for the market's daily low value. That is, you should estimate four probabilities, Prob(phigh = popen), Prob(phigh = pclose), etc. Explain your method and assumptions. Can these results be used to test the random walk hypothesis?

```
mydata_n = mydata[13604:21013,]
max_open = c(mydata_n[,2] == mydata_n[,3])
num_max_open = sum(max_open) # number of days max occurs at open
p_max_open = num_max_open / dim(mydata_n)[1] # prob = 0.1164642

max_last = c(mydata_n[,5] == mydata_n[,3])
num_max_last = sum(max_last) # max occurs at last
p_max_last = num_max_last / dim(mydata_n)[1] # prob = 0.0734143

min_open = c(mydata_n[,2] == mydata_n[,4])
num_min_open = sum(min_open) # number of days min occurs at open
p_min_open = num_min_open / dim(mydata_n)[1] # prob = 0.142915

min_last = c(mydata_n[,5] == mydata_n[,4])
num_min_last = sum(min_last) # min occurs at last
p_min_last = num_min_last / dim(mydata_n)[1] # prob = 0.01997301
```

We estimated the probabilities under the assumption that the price of S&P 500 index on each day follows a Normal distribution and are i.i.d. across time. The true probabilities are estimated to be equal to the sample probabilities observed from the historical price given. The results are:

- Prob(phigh=popen) = 11.65%
- Prob(phigh=pclose) = 7.34%
- Prob(plow=popen) = 14.29%
- Prob(plow=pclose) = 2.00%

From the estimation, we conclude that the price does not follow a symmetric random walk. If it follows a symmetric random walk, then all of the four probabilities should equal to zero, since the probability of a symmetric random walk returning to its origin is 1.

d) From 1/1/1980 through 8/30/2011, find the top 20 intraday ranges. List them, ordered by size. How many occurred during the final three-year sub-period (i.e., 9/1/2008–8/30/2011)?

As shown in the chart below, fifteen of them occurred during the final three-year sub-period.

```
mydata_2 = mydata[13022:21013,]
intraday = c((mydata_2[, 3] - mydata_2[, 4]) / mydata_2[, 4])
result = cbind(mydata_2, intraday)
result = result[order(result[,6], decreasing = TRUE),]
head(result, 20)
```

```
##          Dates PX_OPEN PX_HIGH  PX_LOW PX_LAST   intraday
## 14995 10/19/87  282.70  282.70  224.83  224.84 0.25739448
## 14996 10/20/87  225.06  245.62  216.46  236.83 0.13471311
## 20310 11/13/08  853.13  913.01  818.69  911.29 0.11520844
## 20286 10/10/08  902.31  936.36  839.80  899.22 0.11497976
## 20298 10/28/08  848.92  940.51  845.27  940.51 0.11267406
## 20285  10/9/08  988.42 1005.25  909.19  909.92 0.10565448
## 20287 10/13/08  912.75 1006.93  912.75 1003.35 0.10318269
## 20289 10/15/08  994.60  994.60  903.99  907.84 0.10023341
## 20315 11/20/08  805.87  820.52  747.78  752.44 0.09727460
## 20680   5/6/10 1164.38 1167.58 1065.79 1128.15 0.09550662
## 14997 10/21/87  236.83  259.26  236.83  258.38 0.09470929
## 20290 10/16/08  909.53  947.71  865.83  946.43 0.09456822
## 20277  9/29/08 1209.07 1209.07 1106.39 1106.39 0.09280633
## 15000 10/26/87  248.20  248.22  227.26  227.67 0.09222916
## 20321  12/1/08  888.61  888.61  815.69  816.21 0.08939671
## 20282  10/6/08 1097.56 1097.56 1007.97 1056.89 0.08888161
## 18720  7/24/02  797.71  844.32  775.68  843.42 0.08849010
## 20294 10/22/08  951.67  951.67  875.81  896.78 0.08661696
## 20316 11/21/08  755.84  801.20  741.02  800.03 0.08121238
## 20317 11/24/08  801.20  865.60  801.20  851.81 0.08037943
```

```
k = as.Date(result[1:20, 1], "%m/%d/%y")
```

```
## Warning in strptime(x, format, tz = "GMT"): unknown timezone 'zone/tz/
## 2018e.1.0/zoneinfo/America/New_York'
```

```
num = 0
for (i in 1:20){
  if(k[i] < as.Date(c("2011-08-30")) & k[i] > as.Date(c("2008-09-01"))){
    num = num + 1
  }
}
num
```

```
## [1] 15
```

e) List the top 20 positive overnight returns during the period 1/1/1980–8/30/2011 in reverse chronological order, and separately list the top 20 negative overnight returns. Which three-year period had the largest number of each?

```
overnight <- function(n){
  overnight_ret <- c()
```

```
  for (i in 1:n-1){
    Ron <- (mydata_2[i+1, 2] - mydata_2[i,5])/mydata_2[i,5]
    overnight_ret <- c(overnight_ret, Ron)
  }
  return(overnight_ret)
}
overnight <- c(NA,overnight(nrow(mydata_2)))
result2 = cbind(mydata_2, overnight)
result2_max = result2[order(result2[, 6], decreasing = TRUE),]
```

Top 20 positive overnight returns:

```
head(result2_max, 20)
```

```
##          Dates PX_OPEN PX_HIGH PX_LOW PX_LAST  overnight
## 13100  4/22/80  103.43  103.43 103.43  103.43 0.03637275
## 13547  1/28/82  118.92  118.92 118.92  118.92 0.03282960
## 13242 11/12/80  134.59  134.59 134.59  134.59 0.02536950
## 13084  3/28/80  100.68  100.68 100.68  100.68 0.02504582
## 13324  3/12/81  133.19  133.19 133.19  133.19 0.02493267
## 13462  9/28/81  115.53  115.53 115.53  115.53 0.02447459
## 13486 10/30/81  121.89  121.89 121.89  121.89 0.02376953
## 13027   1/8/80  108.95  108.95 108.95  108.95 0.02003558
## 13583  3/22/82  112.77  112.77 112.77  112.77 0.01952807
## 13466  10/2/81  119.36  119.36 119.36  119.36 0.01947386
## 13193   9/3/80  126.12  126.12 126.12  126.12 0.01931625
## 13130   6/4/80  112.61  112.61 112.61  112.61 0.01900281
## 13487  11/2/81  124.20  124.20 124.20  124.20 0.01895151
## 13091   4/9/80  103.11  103.11 103.11  103.11 0.01887352
## 13216  10/6/80  131.73  131.73 131.73  131.73 0.01855718
## 13157  7/14/80  120.01  120.01 120.01  120.01 0.01841480
## 13333  3/25/81  137.11  137.11 137.11  137.11 0.01811836
## 13076  3/18/80  104.10  104.10 104.10  104.10 0.01799335
## 13237  11/5/80  131.33  131.33 131.33  131.33 0.01774644
## 13565  2/24/82  113.47  113.47 113.47  113.47 0.01757690
```

```
k2_max = as.Date(result2_max[1:20, 1], "%m/%d/%y")
k2_max = sort(k2_max, decreasing = TRUE)

num2 = 0
for (i in 1:20){
  if(k2_max[i] < as.Date(c("1982-12-31")) & k2_max[i] > as.Date(c("1980-01-
01"))){
    num2 = num2 + 1
  }
}
num2
```

```
## [1] 20
```

The period of 1/1/1980 to 12/31/1982 has all twenty of those, which is the most among all periods.

Top 20 negative overnight returns:

```
result2_min = result2[order(result2[, 6]),]
head(result2_min, 20)

##         Dates PX_OPEN PX_HIGH PX_LOW PX_LAST   overnight
## 13075 3/17/80  102.26  102.26 102.26  102.26 -0.03006734
## 13080 3/24/80   99.28   99.28  99.28   99.28 -0.02961587
## 13438 8/24/81  125.50  125.50 125.50  125.50 -0.02886327
## 13259 12/8/80  130.61  130.61 130.61  130.61 -0.02551668
## 13254 12/1/80  137.21  137.21 137.21  137.21 -0.02355537
## 13534 1/11/82  116.78  116.78 116.78  116.78 -0.02317022
## 13554  2/8/82  114.63  114.63 114.63  114.63 -0.02242879
## 13068  3/6/80  108.65  108.65 108.65  108.65 -0.02231621
## 13211 9/29/80  123.54  123.54 123.54  123.54 -0.02223981
## 13279  1/7/81  135.08  135.08 135.08  135.08 -0.02200985
## 13530  1/5/82  120.05  120.05 120.05  120.05 -0.02191625
## 13549  2/1/82  117.78  117.78 117.78  117.78 -0.02176080
## 13297  2/2/81  126.91  126.91 126.91  126.91 -0.02037823
## 13288 1/20/81  131.65  131.65 131.65  131.65 -0.02024261
## 13023  1/2/80  105.76  105.76 105.76  105.76 -0.02019641
## 13461 9/25/81  112.77  112.77 112.77  112.77 -0.01947657
## 13089  4/7/80  100.19  100.19 100.19  100.19 -0.01918747
## 14726 9/25/86  231.83  236.28 230.67  231.83 -0.01883359
## 13182 8/18/80  123.39  123.39 123.39  123.39 -0.01853325
## 13238 11/6/80  128.91  128.91 128.91  128.91 -0.01842686

k2_min = as.Date(result2_min[1:20, 1], "%m/%d/%y")
k2_min = sort(k2_min, decreasing = TRUE)

num3 = 0
for (i in 1:20){
  if(k2_min[i] < as.Date(c("1982-12-31")) & k2_min[i] > as.Date(c("1980-01-
01"))){
    num3 = num3 + 1
  }
}
num3

## [1] 19
```

The period of 1/1/1980 to 12/31/1982 has nineteen of those, which is the most among all periods.

f) List the top 20 jumps jt (ranked by absolute value) in the data set. How many occurred during the 3 years ending 8/30/2011?

```
df_close <- mydata_2 %>%
  select('Dates','PX_LAST') %>%
```

```
  mutate(ret = log(PX_LAST/lag(PX_LAST)))
df_close$std = rollapply(df_close$ret, width = 63, FUN =
sd,align='right',fill = NA)
df_close$jump_m <- df_close$ret/df_close$std
df_close$jump_m_abs <- abs(df_close$jump_m)
jump_max = df_close[order(df_close$jump_m_abs, decreasing = TRUE),]
```

Top 20 jumps:

```
head(jump_max,20)
```

```
##         Dates PX_LAST         ret         std    jump_m jump_m_abs
## 1974 10/19/87  224.84 -0.22899723 0.031442017 -7.283159   7.283159
## 2477 10/13/89  333.62 -0.06321316 0.010577637 -5.976113   5.976113
## 6855  2/27/07 1399.04 -0.03534959 0.006466978 -5.466169   5.466169
## 4509 10/27/97  876.99 -0.07112745 0.013767870 -5.166191   5.166191
## 7976   8/8/11 1119.46 -0.06895833 0.014045813 -4.909529   4.909529
## 4721  8/31/98  957.28 -0.07043759 0.014930823 -4.717596   4.717596
## 665   8/17/82  109.04  0.04645888 0.009894191  4.695571   4.695571
## 3568   2/4/94  469.81 -0.02293582 0.004962634 -4.621703   4.621703
## 3006 11/15/91  382.62 -0.03727171 0.008289952 -4.496011   4.496011
## 1695  9/11/86  235.18 -0.04928004 0.011208419 -4.396699   4.396699
## 3321  2/16/93  433.91 -0.02429288 0.005584977 -4.349682   4.349682
## 7974   8/4/11 1200.07 -0.04900164 0.011267408 -4.348972   4.348972
## 7256  9/29/08 1106.39 -0.09200241 0.021745412 -4.230888   4.230888
## 5485  9/17/01 1038.77 -0.05046794 0.012066904 -4.182344   4.182344
## 1973 10/16/87  282.70 -0.05297564 0.012938518 -4.094413   4.094413
## 7661  5/10/10 1159.73  0.04303473 0.010758786  3.999961   3.999961
## 417   8/24/81  125.50 -0.02928800 0.007762850 -3.772841   3.772841
## 4095   3/8/96  633.50 -0.03131204 0.008469952 -3.696838   3.696838
## 1524   1/8/86  207.97 -0.02764716 0.007527901 -3.672626   3.672626
## 3899  5/31/95  533.40  0.01858178 0.005063836  3.669506   3.669506
```

```
k3_max = as.Date(jump_max[1:20, 1], "%m/%d/%y")
k3_max = sort(k3_max, decreasing = TRUE)

num4 = 0
for (i in 1:20){
  if(k3_max[i] < as.Date(c("2011-08-30")) & k3_max[i] > as.Date(c("2008-09-
01"))){
    num4 = num4 + 1
  }
}
num4
```

```
## [1] 4
```

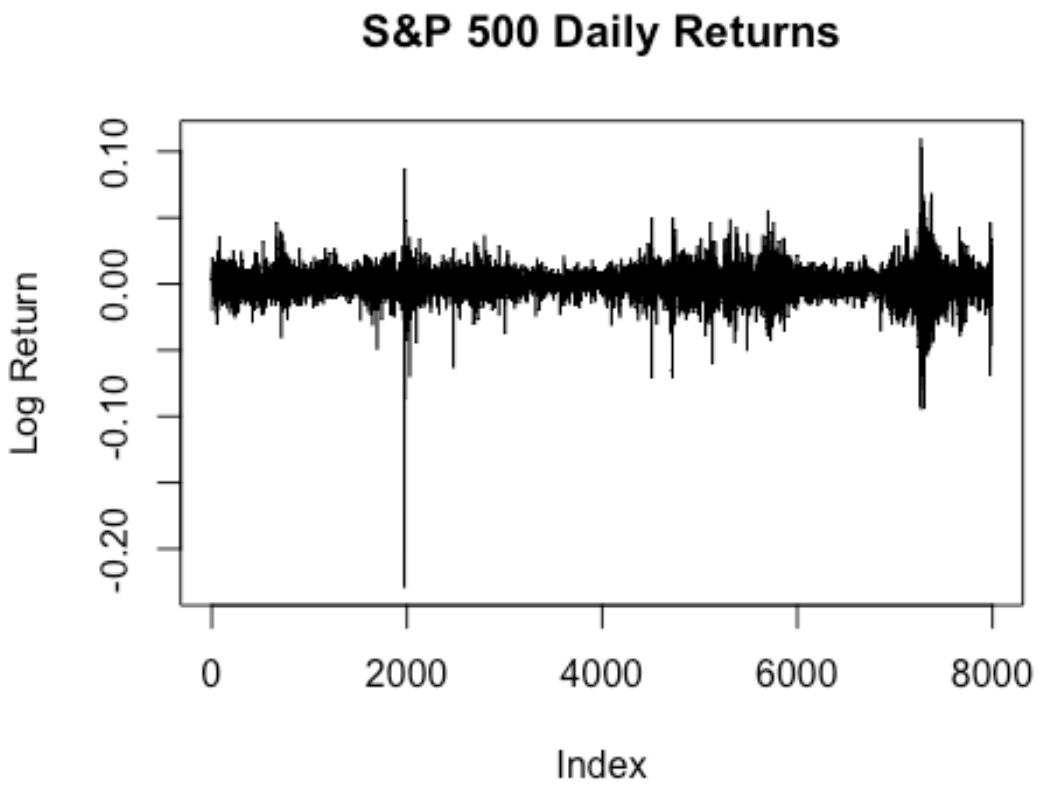
Four of them occurred during the 3 years ending 8/30/2011.

2.On October 6, 1982, equities in the U.S. soared on news of falling interest rates. Different data vendors report conflicting results for S&P 500 index OHLC data for that date. What discrepancies can you find? How economically significant are they? How could they be resolved? Explain what you think were the correct S&P 500 index values on that day and why.

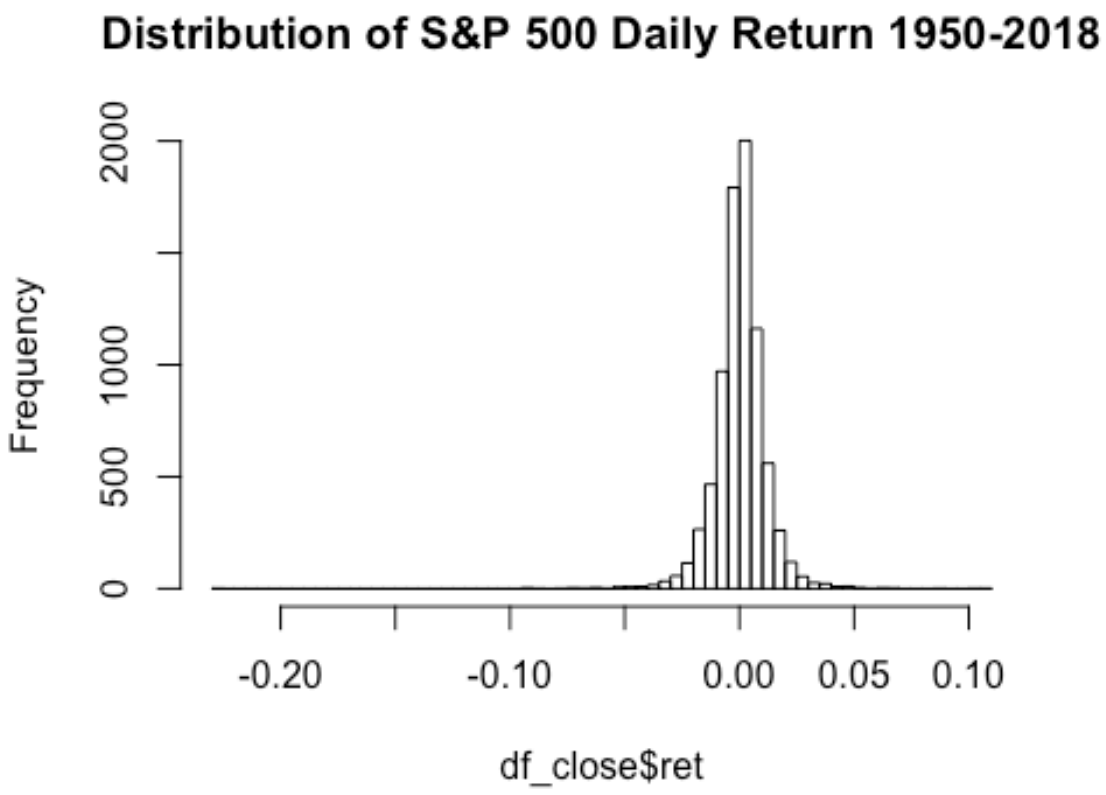
Falling interest rates affect investors' psychology and their projection on the market. Theoretically, when interest rate (the federal funds rate) decreases, the Fed is increasing the money supply and encouraging spending and investments, causing stock prices to rise.

The low provided by both vendors are 122.00, the opening price. According to Yahoo Finance, the high of October 6, 1982 is the closing price 125.97, while that reported by Bloomberg is the closing price 126.97. Both data vendors provide 125.99 as the open of October 7, 1982. We thought Yahoo Finance is more reliable, since the overnight price volatility is smaller. The discrepency is not economically significant because the value is relatively small.

```
plot(df_close$ret, type = "l", main = "S&P 500 Daily Returns", ylab = "Log
Return")
```




```
hist(df_close$ret, breaks = 50, main = "Distribution of S&P 500 Daily Return
1950-2018")
```



The log return data roughly fit normal distribution.

3. On March 6, 2015, Dow Jones announced that one of the thirty companies making up the venerable Dow Jones Industrial Average would be replaced, effective March 18, 2015.

(a) Determine the value the index divisor would have had if the change were made at the market close on the announcement date.

```
# df consists of prices of DJIA, 30 components, AMZN and Bershire Hathaway on
# announcement date 03/06/2015
df = read.csv("/Users/EllieLiu/Desktop/2018 Fall/Financial Data/Proj
1/15.458-PS1-Q3.csv", sep = ",", header = TRUE)

sum.components = sum(df)-df$INDU.Index-df$AAPL.UW.Equity-
  df$AMZN.US.Equity-df$BRK.A.US.Equity #sum of 30 components before change
sum.components.new = sum.components + df$AAPL.UW.Equity -
  df$T.US.Equity #APPL replace AT&T
index.div = sum.components/df$INDU.Index
index.div.new = sum.components.new/df$INDU.Index
cat("The index divisor changed from ", index.div, " to ", index.div.new)

## The index divisor changed from  0.1431868  to  0.1484016
```

(b) What fraction of the total index value was made up of the departing company? That is, what was the total index weight being replaced, as of the announcement date?

```
frac.att = df$T.US.Equity/sum.components
cat("The fraction of the total index value of the
    departing company (AT&T) is", frac.att)

## The fraction of the total index value of the
##     departing company (AT&T) is 0.01309421
```

(c) Did the 29 remaining companies (i.e., those not being changed) have greater or lesser total weight within the index after the change? That is, what was their total weight with and without the change, as of the announcement date?

```
frac.remaining = 1-frac.att
frac.remaining.new = 1-df$AAPL.UW.Equity/sum.components.new

cat("The fraction of the 29 remaining companies changed from",
    frac.remaining, " to ", frac.remaining.new)

## The fraction of the 29 remaining companies changed from 0.9869058  to
0.952226
```

(d) Suppose instead that the replacement member had been Amazon.com. What would be the new estimate of the divisor as of the announcement date? Suppose the replacement had been Berkshire Hathaway (Class A shares). What would be the new estimate of the divisor as of the announcement date?

```
sum.components.amzn = sum.components - df$T.US.Equity +
  df$AMZN.US.Equity
sum.components.berk = sum.components - df$T.US.Equity +
  df$BRK.A.US.Equity
index.div.amzn = sum.components.amzn/df$INDU.Index # AMZN replaced AT&T
index.div.berk = sum.components.berk/df$INDU.Index # BRK replaced AT&T
cat("If AMZN replaced AT&T, index divisor became ", index.div.amzn)

## If AMZN replaced AT&T, index divisor became  0.1625973

cat("If BRK replaced AT&T, index divisor became ", index.div.berk)

## If BRK replaced AT&T, index divisor became  12.39498
```

(e) What role did stock splits play in the timing of the replacement?

In the timing or replacement, if stocks splits occur at component companies (either the new company or the remaining 29 companies), index divisor decreases so that the index remained the same.