MA615 FINAL

Yifu Dong

Abstract

Stock return is related with a lot of things. People always want to use some factros to predict the rough trend of the stock return. For this final project, I am going to explore the relationship between stock return and book value of a company using S&P 500 data. Also, I will explore whether the stock return data follows the Benford's Law. It turns out that company value are related with stock return. Small companies are more likely to fluctuated wildly. Morever, S&P500 Index Return follows well the Benford's Law, which is amazing, whereas the distribution of the return of every individual stock doesn't follow the Benford's Law.

Introduction

In financial market, people always want to predict the stock return. There are many things influncing the ups and downs of stock price, such as economical or political emergencies, continuing growth of revenue and profit. When doing investment, our main job is to make sure the ratio of stock returns to stock price variation, where variation represents the risk. Thus, it's important to know factors influncing stock returns, as well as stock price variation. Prof. Eugene Fama told us one of those factors that are related with stock return, book value of the company. So now I'm gonna check whether there is a relationship of stock returns and book value of companies, and whether I can make prediction by looking at the book value of the company.

Definition

Now I'm going to give a detailed explanation of stock return and book value of company.

For stock return, the formula for the total stock return is the appreciation in the price plus any dividends paid, divided by the original price of the stock.

$$Stock\ Return = \frac{P_1 - P_0 + D}{P_1} \quad ,$$

where P_0 denotes initial stock price in a period, and P_1 denotes ending stock price in this period. D denotes dividends allocated in this period.

A dividend is the distribution of reward from a portion of company's earnings, and is paid to a class of its shareholders.So when we talk about stock return, we need to take dividends into consideration.

Also, what also need to be cosidered is stock split and stock dividend. Stock split is a decision to increase the number of shares that are outstanding by issuing more shares to current shareholders. For example, in a 2-for-1 stock split where the stock price is 20 USD, an additional share is given for each share held by a shareholder. And the stock price will be 10 USD after stock split. Stock dividends are very similar to stock splits in terms of influencing stock price.

However, it's easy to understand that stock split and stock dividends have no inlfuence on stock return. For example, if there is a 10% increase in stock price which is 20 USD, the stock price will be 22, then a 2-for-1 stock split will drop the price down to 11, but it still means a 10% increase, which is a growth from 10 to 11. Thus, when calculating stock return, stock split and stock dividends will not show in the stock return formula.

For book value of a company, it is a term in accounting. Theoretically, book value represents the total amount a company is worth if all its assets are sold and all the liabilities are paid back. This is the amount that the

company's creditors and investors can expect to receive if the company goes for liquidation. The formula shows below:

$$Book\ value\ of\ a\ company = Total\ assets \breve{} Total\ liabilities$$

Basically we can get the asset data and liability data from the balance sheets of a company.

For Benford's Law, it is an observation about the frequency distribution of leading digits in many real-life sets of numerical data. Here is the formula of Benford's Law :

$$\begin{aligned} P(d) =& log_{10}(d+1) - log_{10}(d) \\ =& log_{10}(\frac{d+1}{d}) \\ =& log_{10}(1+\frac{1}{d}) \end{aligned}$$

Briefly explained, Benford's Law maintains that the numeral 1 will be the leading digit in a genuine data set of numbers 30.1% of the time; the numeral 2 will be the leading digit 17.6% of the time; and each subsequent numeral, 3 through 9, will be the leading digit with decreasing frequency.

Data Source

Generally, what we need to collect is the stock prices, and book values of companies. Since there are too many companies in the stock market. So I decided to narrow our stock selecting scale, and choose individual stocks of S&P 500 as our object of study.

Our data mainly comes from https://finance.yahoo.com/. I will collect S&P performance data from Yahoo. Also, I will also use Kaggle and NYSE and find useful data.

Data collecting and cleaning

```
#Import multiple csv files, each file representing an individual stock dataset.
temp = list.files(pattern="*.csv")
myfiles = lapply(temp, read_csv)

#we get a list with 505 elements, each elements representing a dataframe of an individual stock data
#date
sandp500 <- data.frame(myfiles[[1]][,1])
#close price
for (i in 1:505) {
  temporary <- data.frame("date"=myfiles[[i]][,1],"close"=myfiles[[i]][,5])
  sandp500 <- sandp500%>%left_join(temporary,by = "date")
  colnames(sandp500)[i+1] <- myfiles[[i]][1,7]
}

#other structure: converted
price_converted <- data.frame("date"=myfiles[[1]][,1],"close"=myfiles[[1]][,5],"name"=myfiles[[1]][,7])

for (i in 2:505) {
  temporary <- data.frame("date"=myfiles[[i]][,1],"close"=myfiles[[i]][,5],"name"=myfiles[[i]][,7])
  price_converted <- price_converted%>%rbind(temporary)
}
```

```r
#Import data of S&P500 index
index <- read_excel("s&p500 .xlsx")

#Import book value per share
book_value_per_share  <- read_excel("book_to_value.xlsx")


#volume
volume_sandp500 <- data.frame(myfiles[[1]][,1])
for (i in 1:505) {
  temporary <- data.frame("date"=myfiles[[i]][,1],"close"=myfiles[[i]][,6])
  volume_sandp500 <- volume_sandp500%>%left_join(temporary,by = "date")
  colnames(volume_sandp500)[i+1] <- myfiles[[i]][1,7]
}

#weight
weight <- read_excel("weight_by_shares.xlsx")
```

First we need to collect target data from the website since there is no relevant and accessible dataset for us.

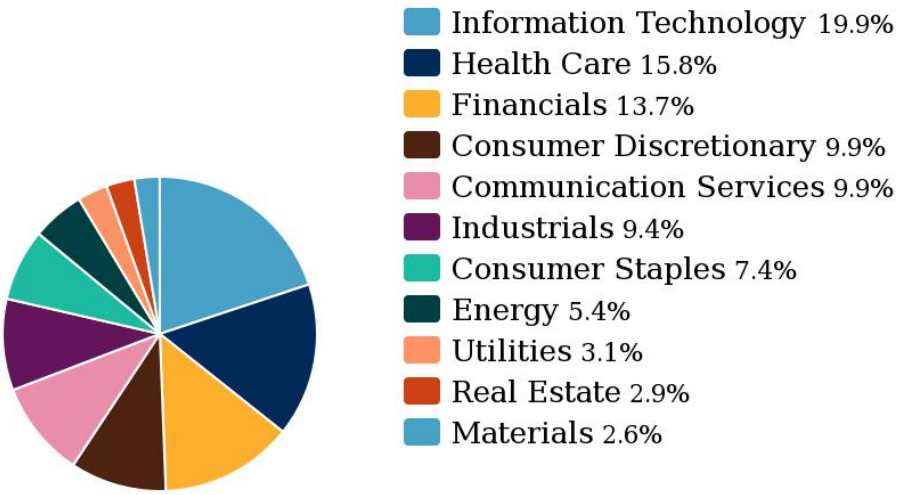
What we want to get is a dataframe, each column representing a stock close price over the past 5 years, and each row representing the close prices of all individual stocks in one day. Thus, I finally get a dataframe with 1259 rows and 506 columns, where 1259 represents that we have 1259 trading days, and 506 represents we have 506 stocks as individual stocks of S&P.

Also, we found that it's hard to collect each company's book value. They are not shown in every website we found. The alternative way is to use book value per share. Then we can use stocks outstanding and book value per share to represent book value roughly.

Why we have 506 instead of 500 individual stocks? Because the individual stocks are changing all the time. S&P will remove those bad performed stocks over the past five years and add well performed stocks in.

Data Visualization

Now we begin to draw some plots and see what will happen.



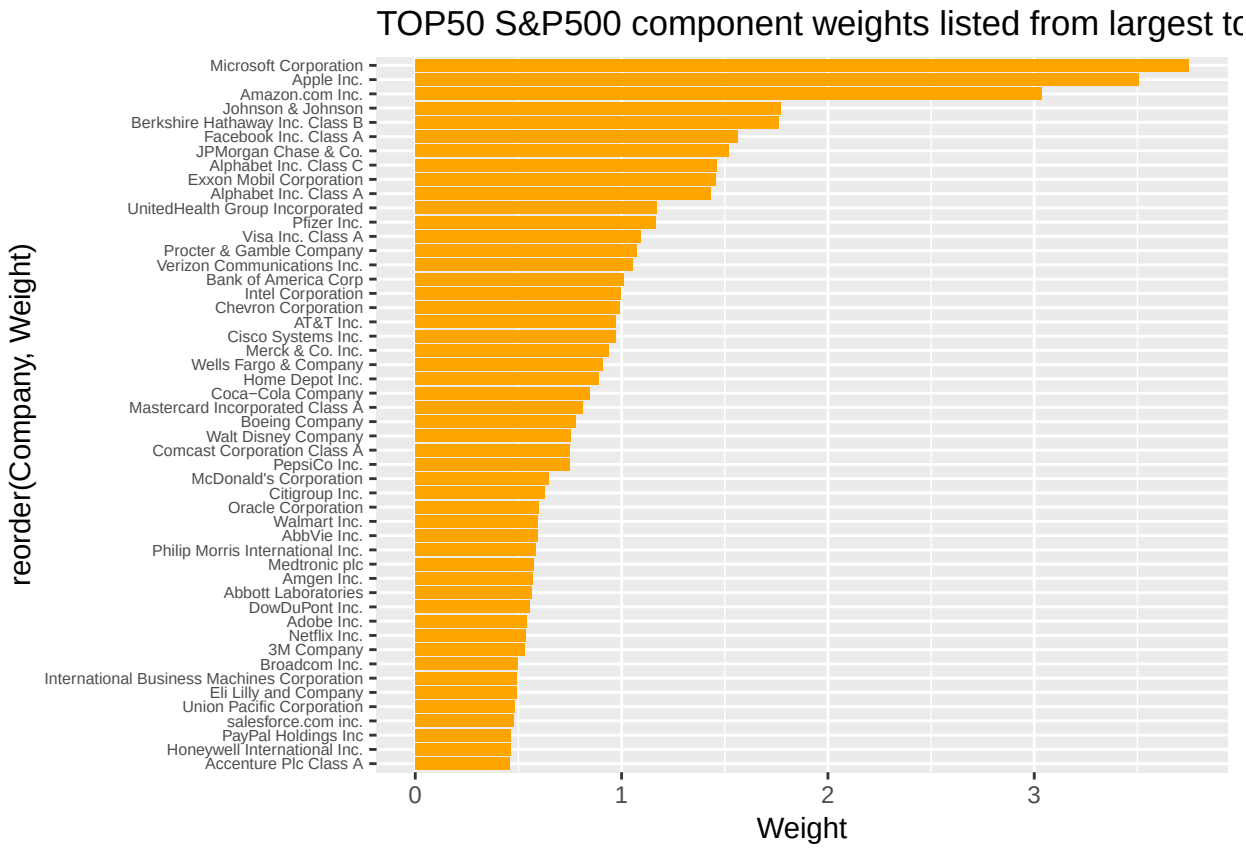
Based on GICS® sectors
The weightings for each sector of the index are rounded to the nearest tenth of a percent; therefore, the aggregate weights for the index may not equal 100%.
As of Nov 30, 2018

Resource: S&P Dow Jones Indices

Our data are about S&P500 index and its individual stocks. S&P500 is constitute of 500 individual stocks in 11 sectors which contains almost all main industries. The pie chart shows above.

Specifically, the weight of every component stock shows below:

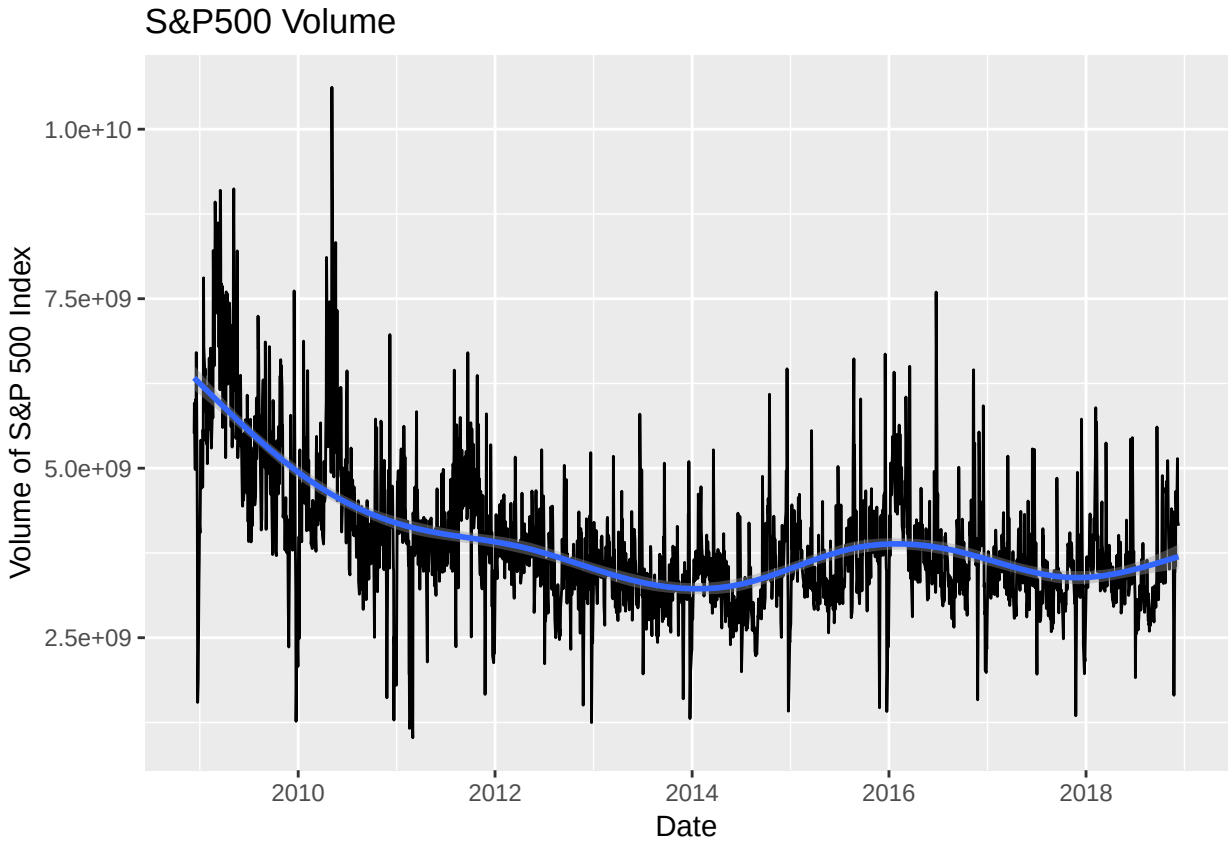
```
p <- ggplot(data=weight[1:50,],mapping = aes(x = reorder(Company,Weight),y=Weight))+
  geom_bar(stat="identity",fill="orange") +
  ggtitle("TOP50 S&P500 component weights listed from largest to smallest")+
   theme(axis.text.y = element_text(angle = 0, hjust = 1,size = 6))


p+coord_flip()
```



Now let's see the trend of trading volume:

```
ggplot(data=index,mapping = aes(x=Date , y=Volume))+
  geom_line()+ylab("Volume of S&P 500 Index")+ggtitle("S&P500 Volume")+geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



From the plot of volume, we know that the active level of investing in S&P500 is decreasing over time. In 2009, the volumn is nearly 6000,000,000 USD. But in 2018, the volumn is only 3000,000,000.

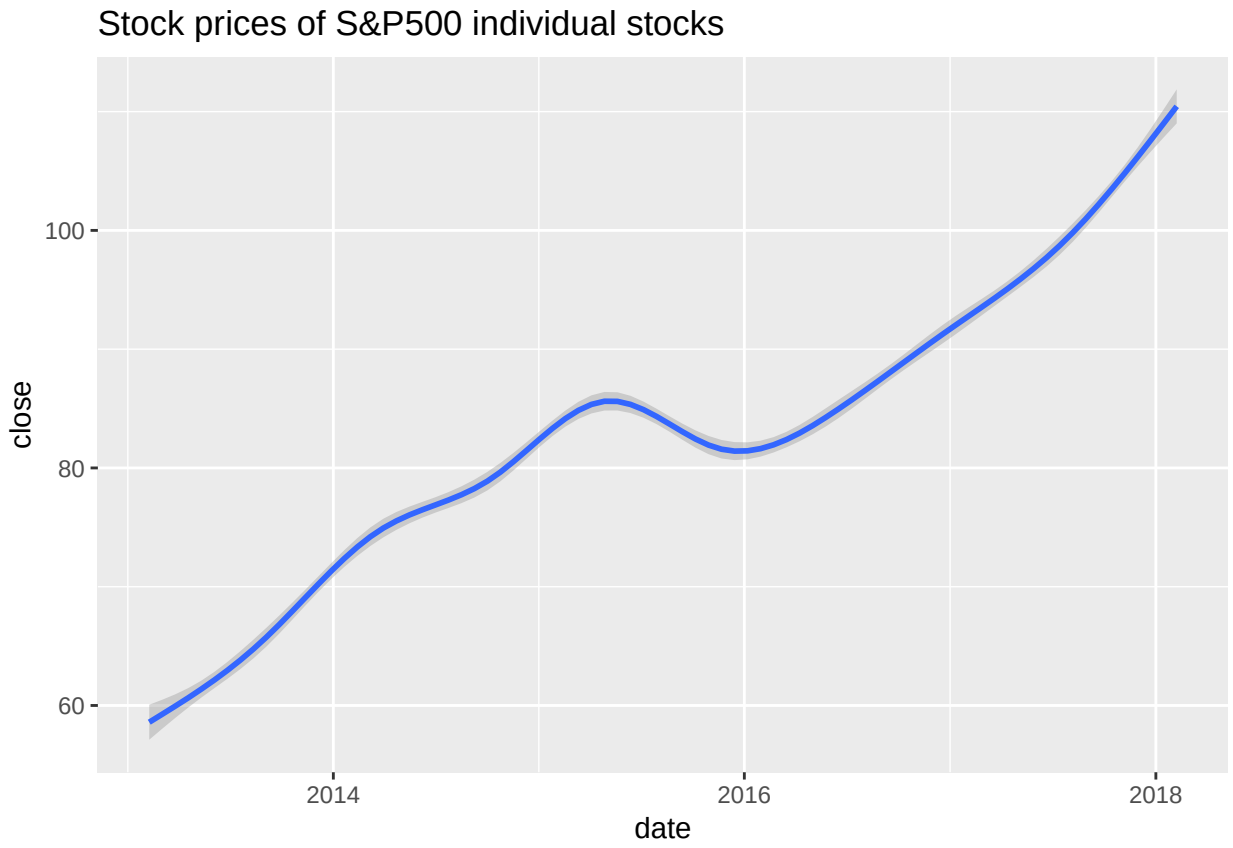
We cannot say the stock market still suffer from the financial crisis in 2008, since it may also due to that the investors in stock market are being rational over time.

Also, the stock prices of individual stocks are increasing, hence the same amount of money may not buy same amount of shares.

I show the trend of stock prices below:

```
#dataframe converting
price_converted <- data.frame("date"=myfiles[[1]][,1],"close"=myfiles[[1]][,5],"name"=myfiles[[1]][,7])

for (i in 2:505) {
  temporary <- data.frame("date"=myfiles[[i]][,1],"close"=myfiles[[i]][,5],"name"=myfiles[[i]][,7])
  price_converted <- price_converted%>%rbind(temporary)
}



ggplot(data = price_converted,aes(x=date, y=close)) +
      geom_smooth()+
      ggtitle("Stock prices of S&P500 individual stocks")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



We can see that the price of nearly every individual stock is increasing over past 5 years.

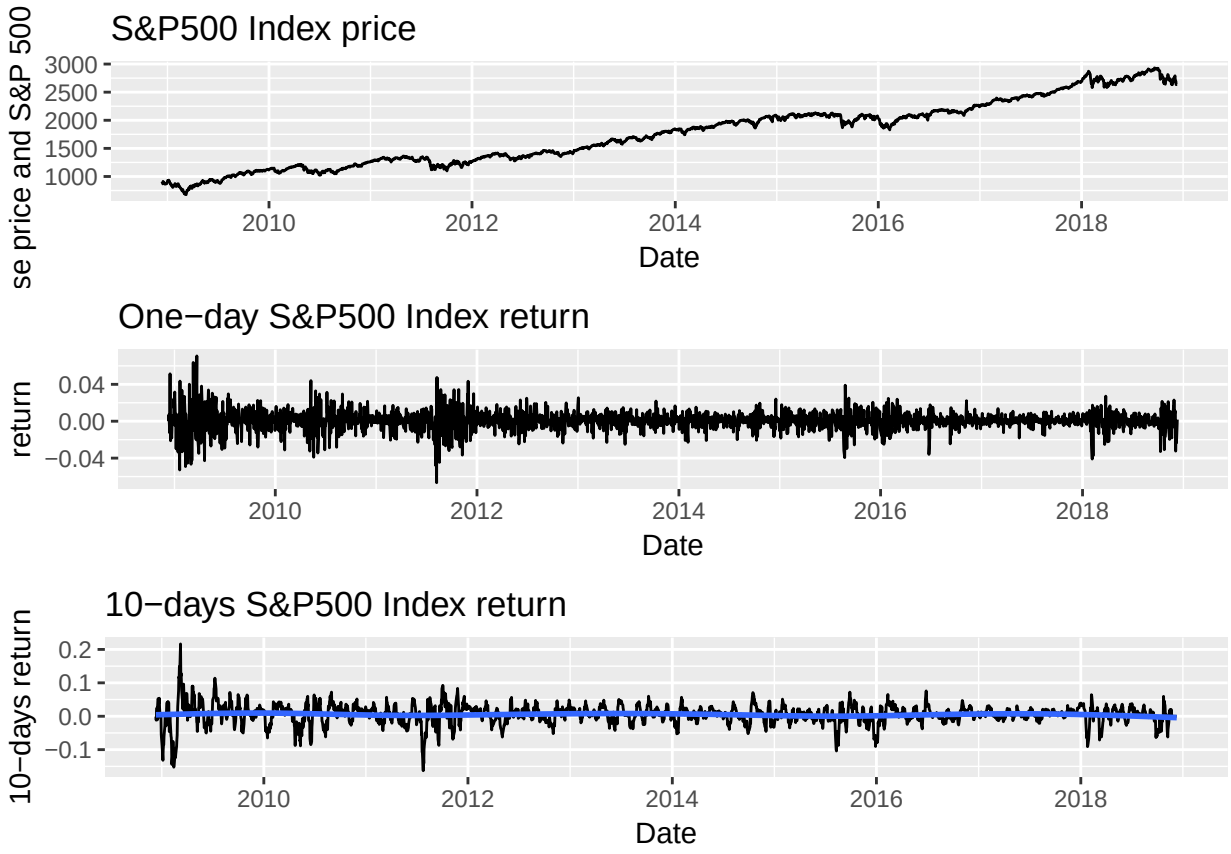
Benford Analysis

Before exploring the relationship between stock return and company value, I'd like to do benford analysis first, and figure out whether the stock return and company value obey the Benford's law. Our data comes from Yahoo Finance and other websites. Basically stock price data are real-time data and there's no way that frauds happen in these stock price data.

Index return

```
#return
index_return <- c(diff(index$Close),0)
index_return <- index_return/index$Close
index_return <- c(0,index_return[1:(length(index_return)-1)])
#mutate
index <- index%>%mutate("return"=index_return)

#10-days return
index_return_10 <- c(diff(index$Close,10),rep(0,10))
index_return_10 <- index_return_10/index$Close
index_return <- c(rep(0,10),index_return[1:(length(index_return)-10)])
index <- index%>%mutate("return10"=index_return_10)

require(gridExtra)
```

```
plot1 <- ggplot(data=index,mapping = aes(x=Date , y=Close))+
  geom_line()+ylab("Close price and S&P 500 index")+ggtitle("S&P500 Index price")
plot2 <- ggplot(data=index,mapping = aes(x=Date , y=return))+
  geom_line()+ggtitle("One-day S&P500 Index return")
plot3 <- ggplot(data=index,mapping = aes(x=Date , y=return10))+
  geom_line()+ylab("10-days return")+geom_smooth()+ggtitle("10-days S&P500 Index return")
grid.arrange(plot1,plot2,plot3,nrow=3)
```




```
mean(index_return_10)
```

```
## [1] 0.004901838
```

The plots above are the trend of S&P500 index and the reusult of Benford Analysis. For the plots of stock price and return, we can see that the return seems like random move, but the stock price is continuously increasing. Then we found that the mean of 10-days index return is 0.4%.

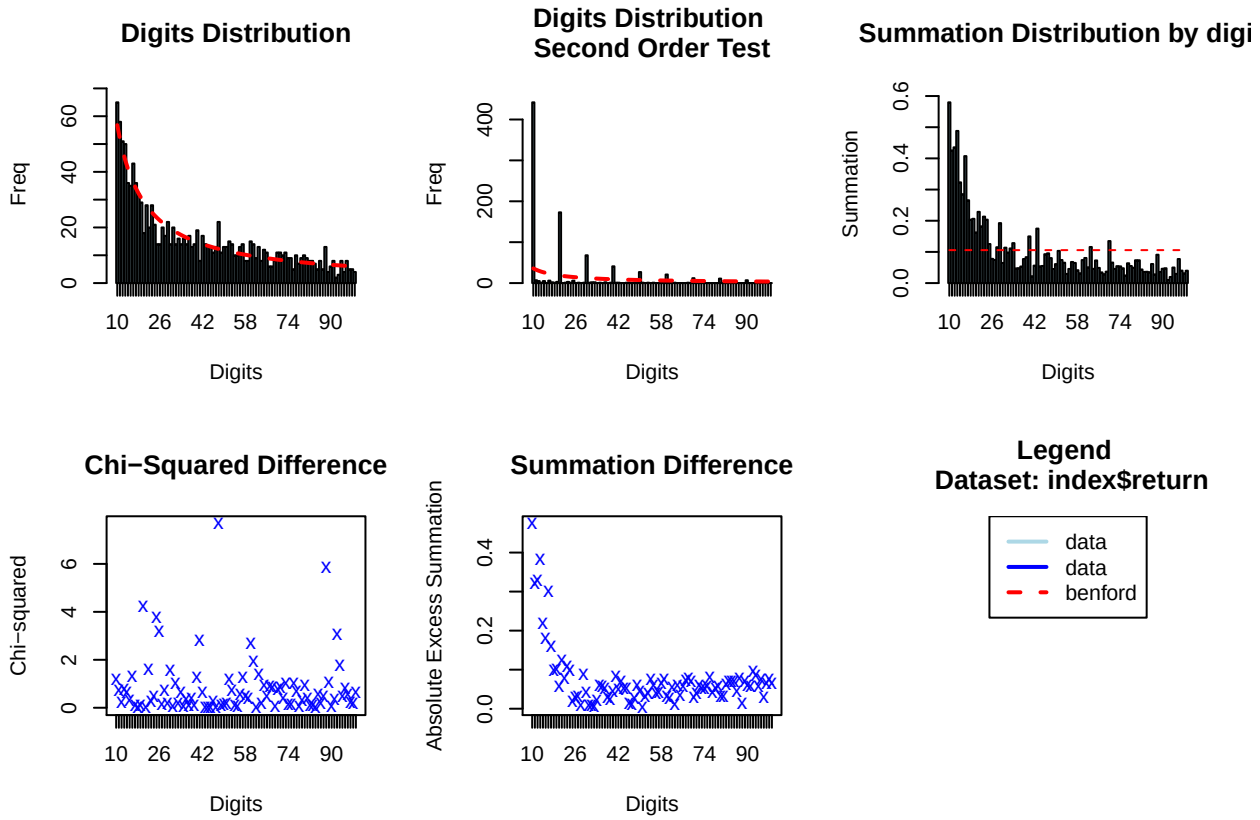
Benford Analysis:

```
#benford analysis on index return
benford_indexreturn <- benford(index$return,number.of.digits = 2)
benford_indexreturn
```

```
##
## Benford object:
##
## Data: index$return
## Number of observations used = 1372
## Number of obs. for second order = 868
## First digits analysed = 2
```

```
##
## Mantissa:
##
##    Statistic  Value
##         Mean  0.503
##          Var  0.087
##  Ex.Kurtosis -1.284
##     Skewness -0.064
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     20         11.07
## 2     48          9.71
## 3     25          9.37
## 4     26          8.49
## 5     10          8.21
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  index$return
## X-squared = 74.501, df = 89, p-value = 0.8646
##
##
##  Mantissa Arc Test
##
## data:  index$return
## L2 = 0.0036428, df = 2, p-value = 0.006752
##
## Mean Absolute Deviation: 0.00201434
## Distortion Factor: NaN
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

```
plot(benford_indexreturn)
```




```
jpsq.benftest(index$return,digits = 2,pvalmethod = "simulate",pvalsims = 10000)
```

```
##
##  JP-Square Correlation Statistic Test for Benford Distribution
##
## data:  index$return
## J_stat_squ = 0.95082, p-value = 0.8628
```

The p-value is 0.86 so we cannot reject the null hypothesis, which means that the distances between data points and benford points are not significantly different. Also the Joenssen's JP-square Test for Benford's law is used here. The result signifys that the square correlation between S&P500 index return and pbenf(2) is not zero and is high related.

Thus, **it's fairly amazing to find that S&P500 Index Return follows well the Benford's Law.**

Stock

We not only have the data of S&P500 Index, but also have the data of every individual stock. Hence, I want to know whether the distribution of the returns of all individual stocks follows the Benford's Law.

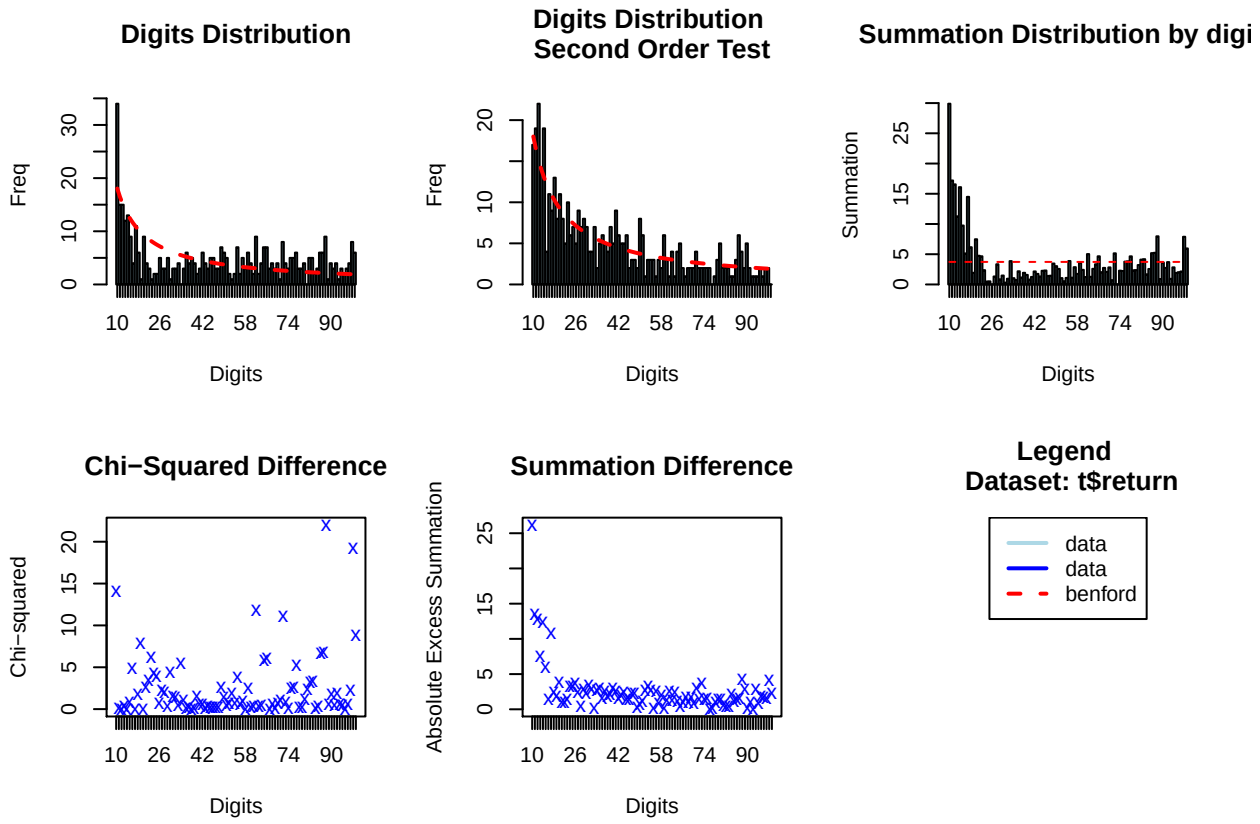
We calculate the return of all individual stocks from 2013 to 2017:

```
#calculate 5-years return of every individual stock
individual <- c(0)
for (i in 1:505) {
  stock_return <- c(diff(sandp500[,i+1]),0)
  stock_return <- stock_return/sandp500[,i+1]
  stock_return <- c(0,stock_return[1:(length(stock_return)-1)])
  individual <- c(individual, sum(stock_return[-1]))
```

```
}

#construct a dataframe make up of company names and stock returns
t <- data.frame("company"=colnames(sandp500), "return"=individual)
t <- t[-1,]

#benford analysis on stock returns
benford_individualreturn <- benford(t$return,number.of.digits = 2)
benford_individualreturn
```

```
##
## Benford object:
##
## Data: t$return
## Number of observations used = 436
## Number of obs. for second order = 435
## First digits analysed = 2
##
## Mantissa:
##
##    Statistic Value
##         Mean  0.58
##          Var  0.11
##  Ex.Kurtosis -1.27
##     Skewness -0.44
##
##
## The 5 largest deviations:
##
##   digits absolute.diff
## 1     10         15.95
## 2     19          8.71
## 3     16          7.48
## 4     23          7.06
## 5     88          6.86
##
## Stats:
##
##  Pearson's Chi-squared test
##
## data:  t$return
## X-squared = 221.23, df = 89, p-value = 3e-13
##
##
##  Mantissa Arc Test
##
## data:  t$return
## L2 = 0.11039, df = 2, p-value < 2.2e-16
##
## Mean Absolute Deviation: 0.005753439
## Distortion Factor: NaN
##
## Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!
```

```r
plot(benford_individualreturn)
```




```r
jpsq.benftest(t$return,digits = 2,pvalmethod = "simulate",pvalsims = 10000)
```

```
##
##  JP-Square Correlation Statistic Test for Benford Distribution
##
## data:  t$return
## J_stat_squ = 0.4165, p-value < 2.2e-16
```

We found from the Benford plot that the distribution doesn't fit the red line. Also, the p-value of Pearson's Chi-squared test is 2.2e-16, so we reject the null hypothesis, which means that the distances between data points and benford points are significantly different. That said, the distribution of returns of individual stocks is significantly different with Benford distribution. The JP-Square Correlation Statistic Test also proves this point. Thus, we have the conclusion that the distribution of stock return of all individual stocks.

Stock return and company value

For this section, I want to take a look at whether there is a relationship between stock return and company value.

First, how to get the data of book value without capital and book value data? I cannot find book value data on the website. But fortunately, many thanks to the website https://www.slickcharts.com/sp500, we can get Book Value per Share as well as the S&P 500 component weights. Thus, we can assume the total number of shares of all component stocks is 100,000. Then we can know the stock outstanding of every single component stock since we know the weights. Then, we will know the book value by shares of every stock and book value per share. This is absulutely not the best way to represent book value. But it's a alternative way to keep exploring.

```
weight <- read_excel("weight_by_shares.xlsx")


bookvalue <- book_value_per_share[,1]
for (i in 1:505) {
  temporary <- data.frame("weight"=weight$Weight[i]*book_value_per_share$Value)
  bookvalue <- bookvalue%>%cbind(temporary)
  colnames(bookvalue)[i+1] <- weight$Symbol[i]
}
#adjustment
bookvalue$Date <- substring(bookvalue$Date,1)




#sort dataframe by colname
bookvalue1 <- bookvalue[,2:506]
bookvalue1 <- bookvalue1[ , order(names(bookvalue1))]
bookvalue <- cbind(data.frame(bookvalue$Date),bookvalue1)


#stock return
individual <- sandp500$date
for (i in 1:505) {
  stock_return <- c(diff(sandp500[,i+1]),0)
  stock_return <- stock_return/sandp500[,i+1]
  stock_return <- c(0,stock_return[1:(length(stock_return)-1)])
  stock_return <- data_frame("return"=stock_return)
  individual <- individual%>%cbind(stock_return)
  colnames(individual)[i+1] <- colnames(sandp500)[i+1]
}

individual_seasonal <- data.frame("date"=bookvalue$bookvalue.Date)
temp <- subset(individual, individual$.>=as.Date(bookvalue$bookvalue.Date[3]) & individual$.<=as.Date(b
temp <- colSums(temp[,2:506])
for (i in 2:75) {
  t <- subset(individual, individual$.>=as.Date(bookvalue$bookvalue.Date[i+1])& individual$. <=as.Date(
  t <- colSums(t[,2:506])  #calculate seasonal return of each component
  temp <- temp%>%rbind(t)
}
individual_seasonal <- cbind(individual_seasonal,temp)
```

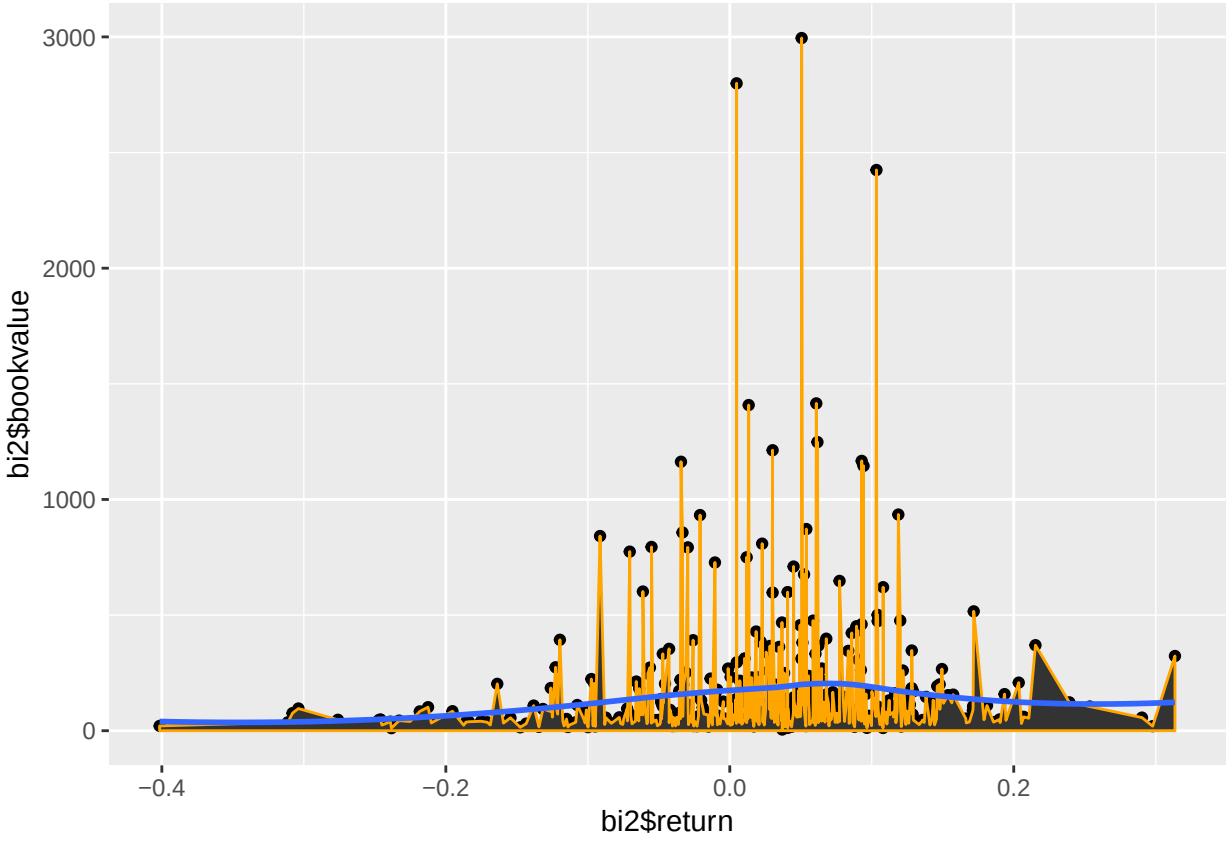
Now, we have two dataframe, "bookvalue" and "individual_seasonal", each containing 506 columns and 75 rows. Since our individual stock data are from 2013 to 2017, thus we only have about 20 rows of effective data, which is made up of 20 seasons. Now we compare the relationship season by season and see what will happen here.

It's common that S&P 500 index will add some stocks into the index and remove some other stocks. Hence, what we need to consider will be the fixed part.

```
b2 <- data.frame(t(bookvalue[5,2:506]))
i2 <- data.frame(t(individual_seasonal[5,2:506]))
b2$company <- rownames(b2)
i2$company <- rownames(i2)
bi2 <- b2%>%inner_join(i2,by="company")   #inner_join
```

```r
colnames(bi2) <- c("bookvalue","company","return")
bi2 <- bi2[order(-bi2$return),]  #sort by company value
bi2 <- na.omit(bi2)


ggplot(data = bi2,mapping = aes(x=bi2$return,y=bi2$bookvalue))+
  geom_point()+
  geom_area(col="orange")+
  geom_smooth(se=FALSE)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



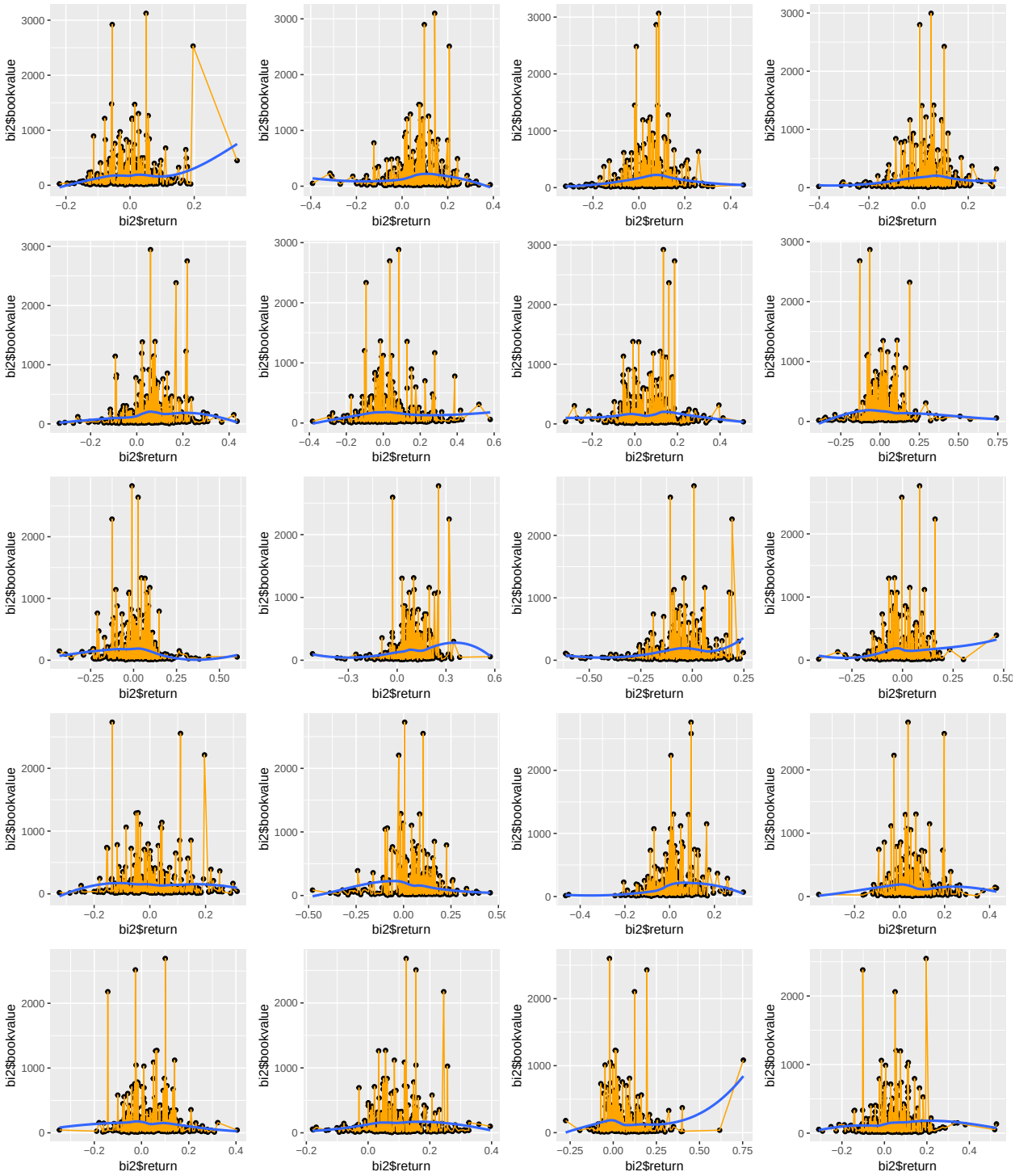
The plot above is the book value and return of one season.

From the plot above, it's clear that the higher the book value, the more possible the absolute value of return is lower. Hence, we know that for those big companies whose book value is large, the absolute value of their stock return tends to be lower than small companies. In other words, small companies are more possible to have higher return or lower return, their stock returns are more flucuated.

Then I try to draw plots for all season:

```r
myplot <- function(i){
b2 <- data.frame(t(bookvalue[i,2:506]))
i2 <- data.frame(t(individual_seasonal[i,2:506]))
b2$company <- rownames(b2)
i2$company <- rownames(i2)
bi2 <- b2%>%inner_join(i2,by="company")   #inner_join
colnames(bi2) <- c("bookvalue","company","return")
bi2 <- bi2[order(-bi2$return),]  #sort by company value
```

```r
bi2 <- na.omit(bi2)


ggplot(data = bi2,mapping = aes(x=bi2$return,y=bi2$bookvalue))+
  geom_point(stat = "identity")+
  geom_line(col="orange")+
  geom_smooth(se=FALSE)


}


require(gridExtra)
plot1 <- myplot(2)
plot2 <- myplot(3)
plot3 <- myplot(4)
plot4 <- myplot(5)
plot5 <- myplot(6)
plot6 <- myplot(7)
plot7 <- myplot(8)
plot8 <- myplot(9)
plot9 <- myplot(10)
plot10 <- myplot(11)


plot11 <- myplot(12)
plot12 <- myplot(13)
plot13 <- myplot(14)
plot14 <- myplot(15)
plot15 <- myplot(16)
plot16 <- myplot(17)
plot17 <- myplot(18)
plot18 <- myplot(19)
plot19 <- myplot(20)
plot20 <- myplot(21)
grid.arrange(plot1, plot2, plot3,plot4,plot5,plot6,plot7,plot8,plot9,plot10,
             plot11, plot12, plot13,plot14,plot15,plot16,plot17,plot18,plot19,plot20,
             nrow=5)
```



The plots above show the relationship every season. Each row represents one year from 2013 to 2017, from season 1 to season 4. We can find that we will find more big companies when return is near 0 in every season.

Thus, we can say our assumption is partly right. Companies with small book value are more likely to have higher return, but also more likely to have lower return. In other words, small companies are more likely to fluctuated wildly.

Discussion

Implication

1. From the EDA, we can roughly have the conclusion that company value are related with stock return. Small companies are more likely to fluctuated wildly. But we still need to do more than data visualization to figure out the relation.
2. From the Benford Analysis, we know that S&P500 Index Return follows well the Benford's Law, which is amazing. But the distribution of the return of every individual stock doesn't follow the Benford's Law.

Future Work

In the future, I want to figure out the reason why S&P500 Index Return follows the Benford's Law, because S&P500 index is supposed to increase or decrease following no rule. Thus, I think more data needs to be collected.

Also, I need to quantify the relationship between stock return and company value in the future.

Reference

[1]. https://www.slickcharts.com/sp500

[2]. https://www.worldbank.org/

[3]. https://finance.yahoo.com/quote/%5EGSPC?p=^GSPC

[4]. https://us.spindices.com/indices/equity/sp-500