MONASH BUSINESS SCHOOL

How will the stock market react to the PCA: Envidence From Yahoo Finance Stock Market

Kaiwen Jin
26686953

Zhiruo Zhang
28009487

Jinhao Luo
29012449

Report for
ETF5500 Assignment2

4 October 2020

Contents

1   Introduction

In the financial market, the value of stocks would be investigated by many different variables. However, a large number of stocks make investors hard to make their decision. Therefore, this report would apply linear combination (LC) by combining the variables into an index and utilise principal component analysis (PCA) to evaluate the performance of stocks.

Besides, this report will also consider the accuracy of PCA and will discuss the potential limitation of PCA in stocks performance evaluation. Based on the result, Clustering Analysis as a comparative approach will also be provided.

At last, some useful suggestions for the stocks choosing will be concluded, as well as concluding the biases generated from the analysis.

The appendix will contain some notes to improve the understanding of our reports.

2   Data Description

2.1   Description

Our data was sourced from Yahoo Finance and it contains 18 variables of 147 stocks from five major financial indices. Those 18 variables could be further classified into 3 categories. The first categories capture **Name**, **Symbol**, **Market**, **Sector**, **Industry** which are related to the background of those stocks. The second and third categories provide some measurement of the value and risk which are related to the stocks. The further description of those variables is shown in Table 1.

2.2   Limitation

This part will provide us an introduction about the limitation of the dataset, and it is shown below:

- This dataset contains a lot of missing value which would cause some bias in our final result
- This dataset does not contain enough observations. The insufficient sample space will make our final result become unreliable. Also, if we further filter out the missing values, the sample size of the data would be even smaller. And the relatively small sample would not be representative enough to clarify the overall condition.
- There is some inconsistency between the total ESG risk score and the sum of individual risk score. This inconsistency would directly increase the error in our final output.

**Table 1:** *Information of variables of the original data*

Variable	Abbreviation	Description
Name	/	The full company name of each stock
Symbol	/	The abbreviation of each stock
Market	/	Major financial indices
Sector	/	The belonging section of a stock
Industry	/	The belonging industry of a stock
Market capitalization	intra_day	How much a company is worth as determined by the stock market
Enterprise value	ent_value	A measure of a company's total value
Trailing P/E	trail_pe	Price to Earning Ratio based on the earnings per share over the previous 12 months
Forward P/E ratio	for_pe	Estimate further earnings per share in the next 12 months
PEG ratio	peg	Enhances the P/E ratio by adding the expected earnings growth into calculation
P/S ratio	ttm	Price to Sales ratio, a valuation ratio by comparing a company's stock price to its revenue
P/B ratio	mrq	Price to Book ratio is a measurement of the market's valuation of a company relative to its book value
Enterprise value-to-revenue	rev	Also refers as the EV/R, it measures the value of a stock that compares a company's enterprise value to its revenue
EV/EBITDA	ebitda	Enterprise value to earnings before interest, taxed, depreciation and amortization ratio compares the value of a company, debt included to the company's cash earnings less non-cash expenses
Total ESG risk score	tot_risk	The overall rating scores based on the Morningstar Sustainability Rating systems
Environmental Risk Score	envir_risk	Evaluation scores of the portfolios performance when they meet the environmental challenges
Social Risk Score	social_risk	Evaluation scores of the portfolios performance when they meet the social challenges
Governance Risk Score	gover_risk	Evaluation scores of the portfolios performance when they meet the governance challenges

Those limitations would be further discussed in the following sections. At last, the biases of analysis which generate from the limitations would be concluded.

3 Analysis

3.1 Preliminary Analysis

We will tidy our original dataset by removing the missing variables and further figure out other features. Figure 1 shows the general data structure and it could be classified into three types which are character, numeric and missing value.
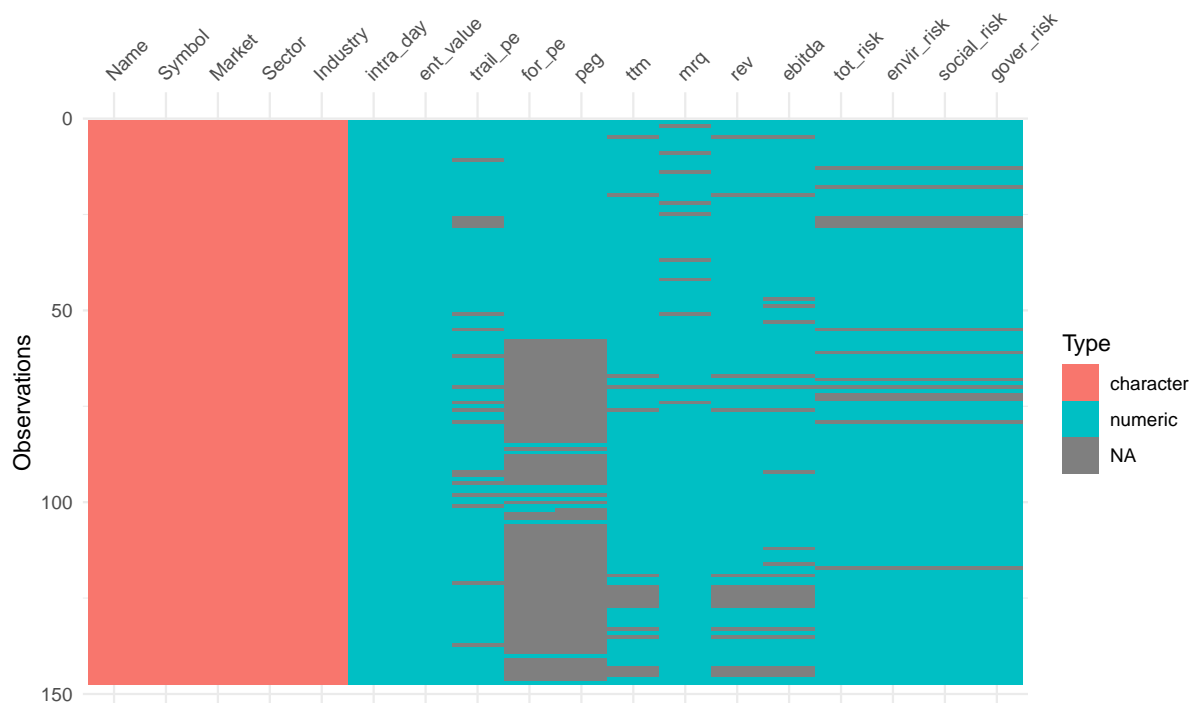
Figure 1: *The data structure of original data*

Table 2 indicates that the initial 147 observations have up to 102 missing value and also some potential outliers.

Most of the variables have a small median and mean, but an extremely high maximum value. Those extreme value would dominate our Principle Component Analysis and those outliers are shown in Table 3.

New dataset **stocks** will be generated by removing the missing value.

Table 2: *Summary table of original data*

Variable	Min	Median	Mean	Max	NA
intra_day	-2	63	95065	5110000	NA
ent_value	-264	70	85683	5130000	NA
trail_pe	0.48	20.11	43.62	1479.29	18
for_pe	3.59	19.92	43.04	1044.81	80
peg	-62.380	2.405	15.223	713.670	81
ttm	0.9	2.8	9.941	548.150	17
mrq	0.1	5.4	174.16	11765.96	10
rev	-27.720	2.875	9.827	5411.160	17
ebitda	-465.460	13.765	19.461	1117.510	23
tot_risk	11	23	25.39	75	13
envir_risk	0	4	6.731	62	13
social_risk	3	10	11.4	88	13
gover_risk	3	8	9.343	80	13

Table 3: *The summary table of outliers in each variables*

| variable	outlier
intra_day	MSFT, AAPL
ent_value	MSFT, AAPL
trail_pe	TSLA
for_pe	ILMN, TSLA
peg	DIS, VZ, KO, MMM, CVX, PCAR, CAT, XOM
ttm	ILMN, V
mrq	TSLA
rev	ILMN, V
ebitda	INTU, ILMN, TSLA, NKE

3.2 Principle Component Analysis

3.2.1 Value Analysis

Value analysis will be conducted by removing the high influential outliers. We also need to standardize the data due to the different units.

Referring to the PCA value table, we could notice that PC1 is positive correlated with the measurement of company value which is **intra_day** and **ent_value**. PC2 is positive correlated with the stock earnings ratio (**ebitda** and **trail_pe**) which means that the increase in the stock earnings ratio will increase PC2 slightly. We could notice that the variables which are related to the price-based evaluation of the stock are pretty close to the PC2.

Meanwhile, according to Figure 2 the narrow angle between **intra_day** and **ent_value** could highlight their strong association and the nearly 90 degree angle with **for_eg**,**ttm**, **mrq**, **rev** states the zero
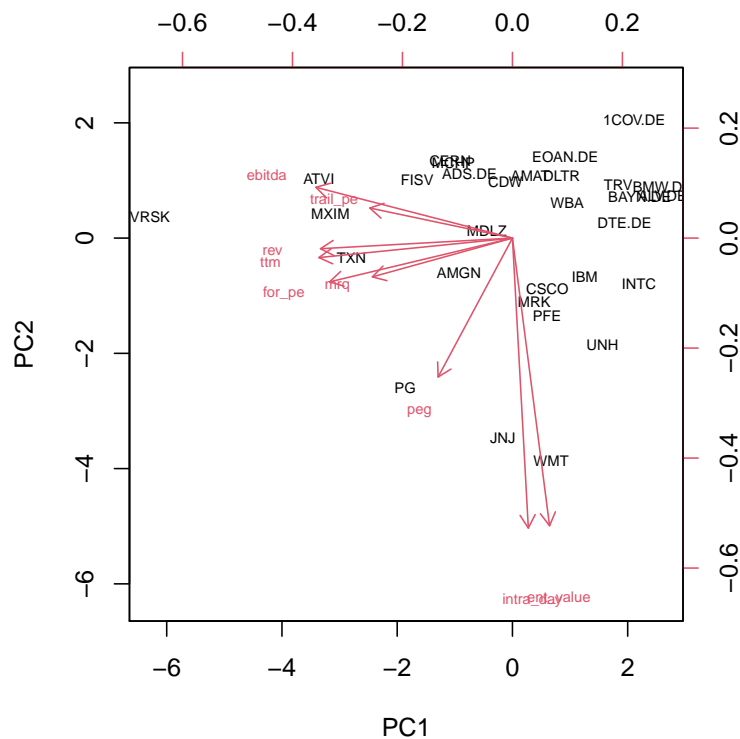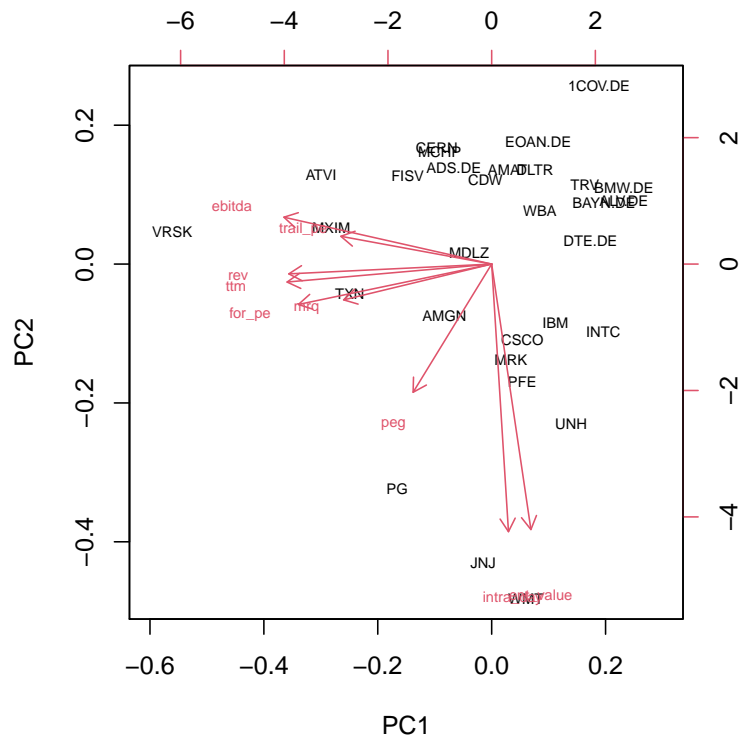
Figure 2: *Correlation Biplot of PCA of stocks' value*

relationship. Furthermore, we could say that the angle between the measurement of company value and **ebita**, **trail_pe** is close to 180 degree, their relationship should be negative. In other words, market value of a company will not influence its stock price and earn per share, but price and earnings will highly impact each other.

Distance biplot 3 indicates that **Johnson & Johnson (JNJ)** and **Walmart (WMT)** have a high value in PC1, and **Activision Blizzard (ATVI)**, **Texas Instruments Incorporated (TXN)**, **Maxim Integrated Products (MXIM)** are higher in PC2.

Furthermore, we notice that **Verisk analytics (VRSK)** is a potential outlier for the PC1, **JNJ**, and **WMT** are potential outliers for PC2. The reason is based on the characteristic of these firms. Being a data-analytics and risk-assessment firm, **VRSK** provides the consulting service instead of the goods selling. Therefore, as for the financial sector, they profit generating will not directly lead to the increasing firm size. **JNJ** and **WMT** perform oppositely since they are operating as the multinational company. The continuously increasing market share will keep their profit at a high level.

The limitation of the value analysis also exists.

**Figure 3:** *Distance Biplot of PCA of stocks' value*

Table 4: *Summary table of PCA for value analysis of stocks*

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	2.0468	1.4658	0.9319	0.9005	0.7316	0.5625	0.3081	0.1727	0.0784
Proportion of Variance	0.4655	0.2387	0.0965	0.0901	0.0595	0.0352	0.0106	0.0033	0.0007
Cumulative Proportion	0.4655	0.7042	0.8007	0.8908	0.9503	0.9855	0.9960	0.9993	1.0000

- After we filter out the outliers, the number of variables we put into use is 30 out of 147 and only 70.42% of the overall variation could be explained by the first two principles (Table 4). The really small space size is not representative and also would not accurate enough to explain the whole stock market condition.

- Screeplot (Figure 4) suggests the best PC selection is 3 which made the biplot 2 selection inadequate.

Therefore, we need to consider an alternative approach to make sure of the accuracy of our suggestion.

pca_value
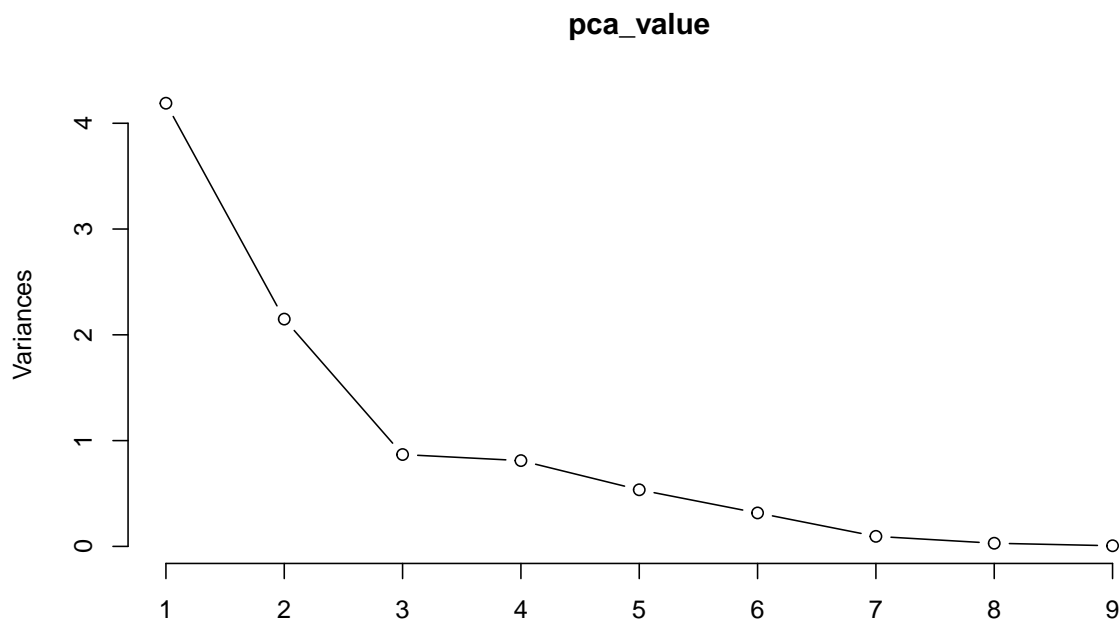
Figure 4: *Screeplot of PCs in PCA for value analysis of stocks*

Table 5: *Summary table of PCA for risks analysis of stocks*

|	PC1	PC2	PC3	PC4
Standard deviation	1.3946	1.2190	0.7544	0
Proportion of Variance	0.4862	0.3715	0.1423	0
Cumulative Proportion	0.4862	0.8577	1.0000	1

3.2.2 Risk Analysis

This part will provide a discussion about the potential risk of each stock based on the ESG risk score. We will compare the total risk score with the sum of the ESG scores to make sure the consistency of our data. Filtering out the inconsistent value would improve the accuracy of our results.

Table 5 that PC1 and PC2 have explained almost 86% of the total variation of 4 variables. Besides, Figure 5 also suggests that the principal component of one and two should be selected because they all with a variance greater than 1 according to the Kaiser's Rule.

Figure 6 shows the distance among each stock in the dataset and implies the similarity between stocks. The stocks of **VRSK** and **UnitedHealth Group Incorporated (UNH)** may be the same because they seem perfectly superimpose. Besides, **Allianz SE (ALV.DE)** and **Dollar Tree, Inc. (DLTR)**, as well as **Cerner Corporation (CERN)** and **Fiserv, Inc. (FISV)** might be similar, since they are close to each other. While the far distance between **VRSK** and **Microchip Technology Incorporated (MCHP)**, or

pca_risk
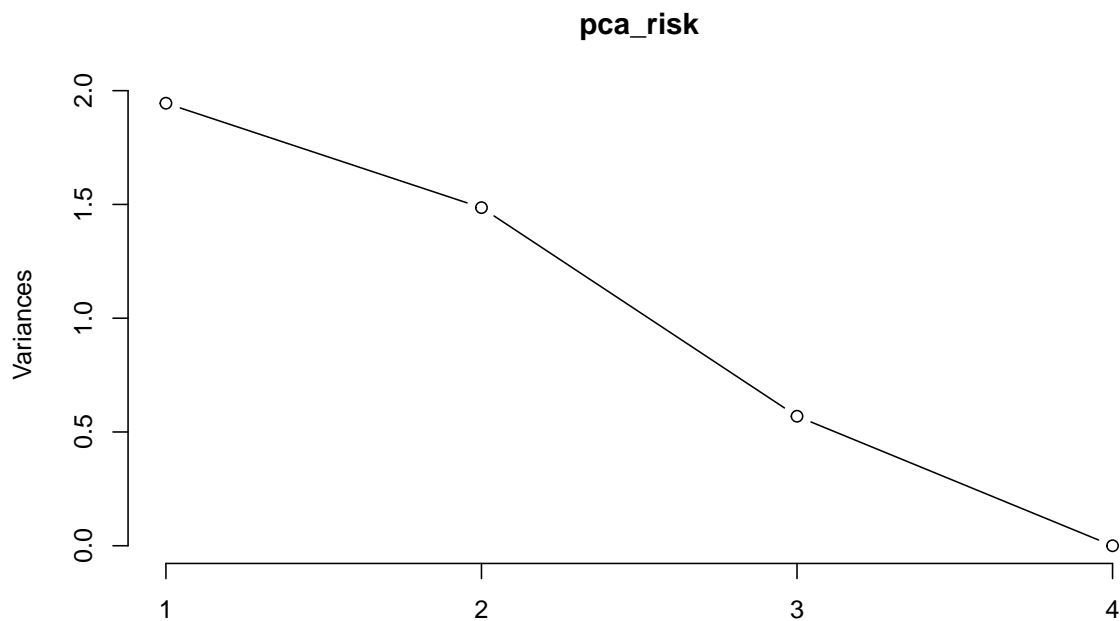
Figure 5: *Screeplot of PCs in PCA for risk analysis of stocks*

CDW Corporation (CDW) and **Pfizer Inc. (PFE)** indicate their different risk performance. To further analyzing the correlation between each stock, a correlation biplot is required.

According to Figure 7 the correlation biplot **VRSK**, **UNH**, **CERN**, and **FISV** have the high values of social risk score, which indicate that these four stocks might perform better than the others in the social challenges. While they might not good at dealing with the risk from the environment since the social risk score and environmental risk score are approximately negative correlated. In contrast, **Covestro AG (1COV.DE)** and **MCHP** have the strongest abilities to face environmental challenges. Meanwhile, **MCHP** also has the highest score of governance risk, which indicates good performance in anticipating governance challenges. **PFE**, **Bayerische Motoren Werke AG (BMW.DE)**, and **Merck & Co., Inc. (MRK)** also perform well. While the projected positions of these three stocks along the axis of governance risk score are gradually decreasing, the approximate actual values of stock performance might gradually decline.

In general, based on the total ESG risk score, **MCHP** and **BMW.DE** have the best overall performance compared with other stocks, which indicate that their strategy in risk management is quite effective. Therefore, even if some external challenges occur, they will not have any significant fluctuations. In contrast, **CDW** has a weak overall risk performance for the low value in total risk score which indicates that the **CDW** stock price will not be stable when facing the risk.
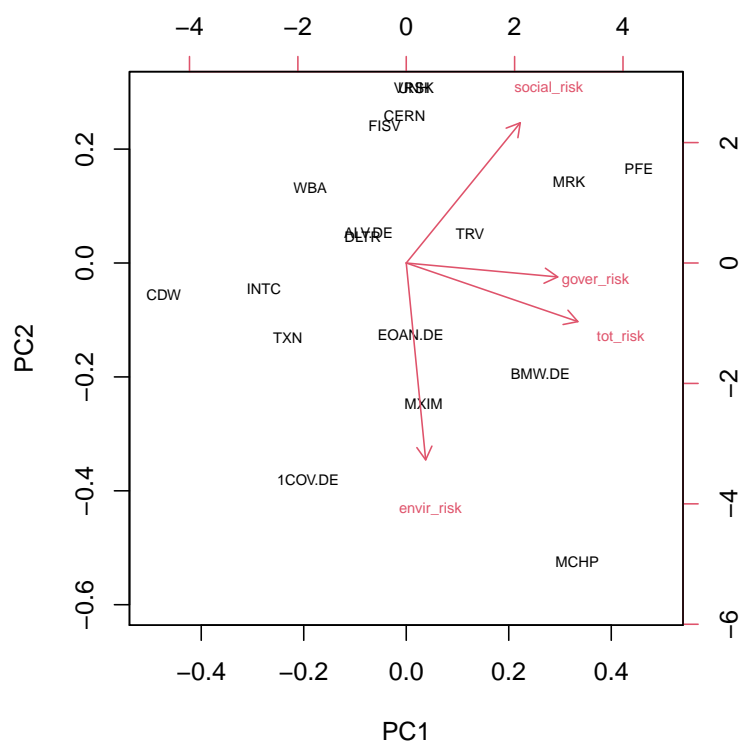
Figure 6: *Distance biplot of PCA of stocks' risk*
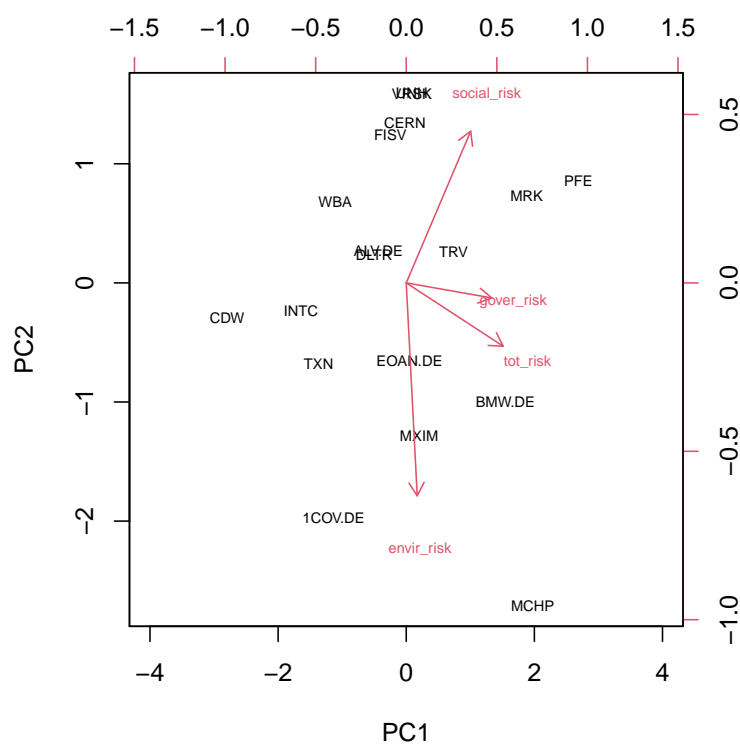
Figure 7: *Correlation Biplot of PCA of stocks' risk*

3.3 Cluster Analysis

Using the hierarchical clustering analysis with the agglomerative method, it is a bottom-up approach. We first select the data set that is consistent with the value data by equivalent stocks symbol. After standardised the data for the numeric variables, the Euclidian distance applied to find the distance between all pairs of observations. We employ Ward's methodology to sort the clusters. And the resulting clusters are shown in the dendrogram to display the sequences of merges or splits. Figure 8 claims that two and four clusters of solutions are not stable. Hence, the three cluster solution is stable which is shown in 9.
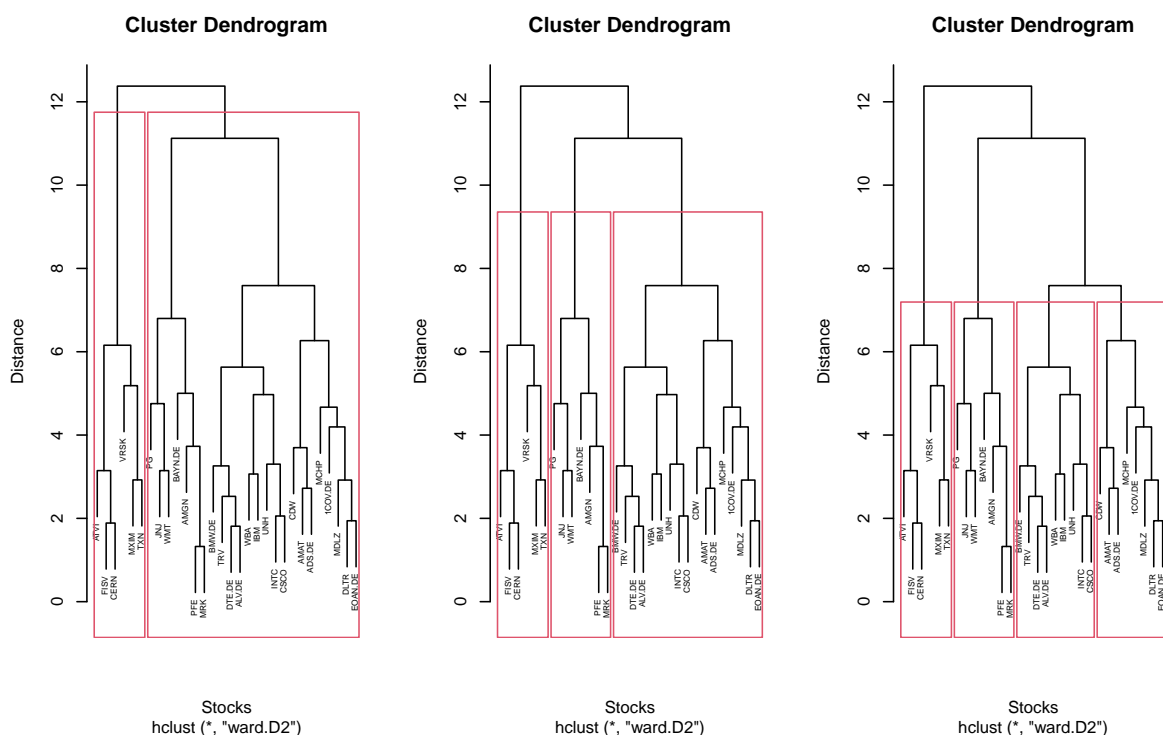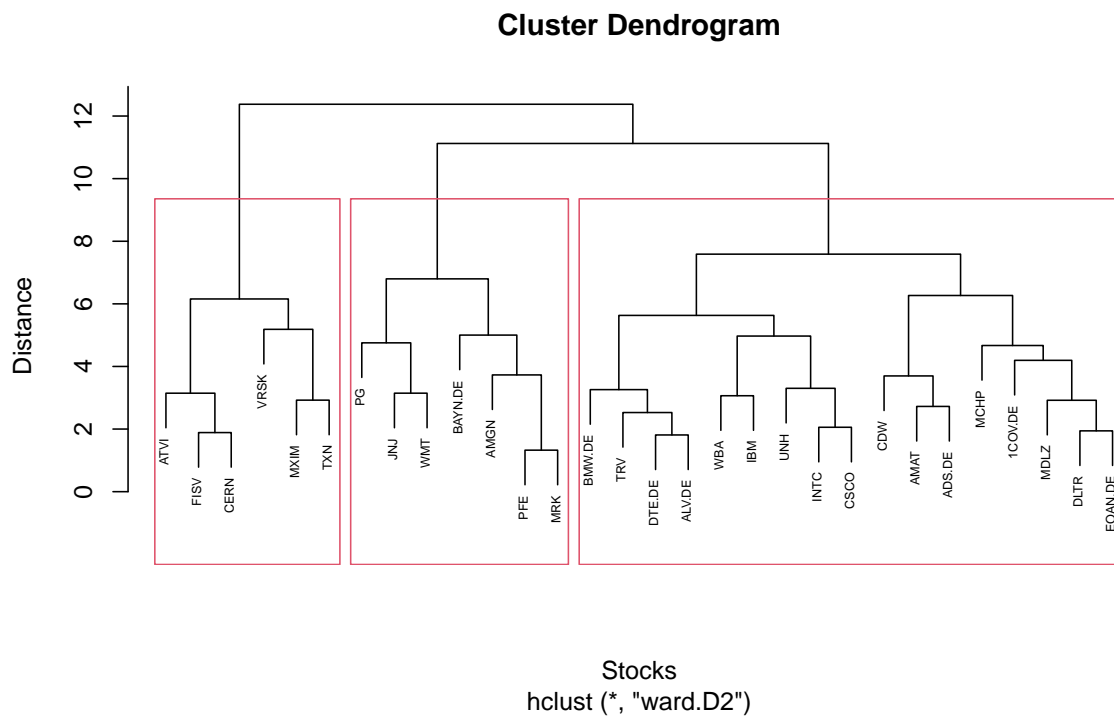

Figure 8: *Choosing clusters*

From the dendrogram, there are three different clusters. These three clusters are shown in Table 6, Table 7, and Table 8, respectively.

Cluster Dendrogram

Stocks
hclust (*, "ward.D2")

Figure 9: *Dendrogram using Ward methodology and taking Euclidian distances*

Table 6: *The stocks of the first cluster*

stock
Verisk Analytics, Inc.
Activision Blizzard, Inc.
Fiserv, Inc.
Maxim Integrated Products, Inc.
Texas Instruments Incorporated
Cerner Corporation

Table 7: *The stocks of the second cluster*

stock
Amgen Inc.
Johnson & Johnson
Pfizer Inc.
Walmart Inc.
Merck & Co., Inc.
Bayer Aktiengesellschaft
The Procter & Gamble Company

Table 8: *The stocks of the third cluster*

stock
CDW Corporation
Microchip Technology Incorporated
Dollar Tree, Inc.
Mondelez International, Inc.
Applied Materials, Inc.
Intel Corporation
UnitedHealth Group Incorporated
Cisco Systems, Inc.
The Travelers Companies, Inc.
Walgreens Boots Alliance, Inc.
International Business Machines Corporation
E.ON SE
Deutsche Telekom AG
adidas AG
Bayerische Motoren Werke AG
Allianz SE
Covestro AG

4   Conclusions

According to our general evaluation of the Yahoo Finance market, we could say that **JNJ** and **WMT** perform better in company evaluation while **ATVI**, **TXN**, and **MXIM** are better in price and earning. **VRSK**, **JNJ** and **WMT** could be considered as the special cases since they outperform in their area. What ESG risk analysis provides us is that **MCHP** and **BMW.DE** have a great performance in overall anti-risk, some other companies perform better in a specific risk score. For instance, **UNH**, **CERN**, and **FISV** perform well in anti-social risk, but they are not resilient enough when meeting the challenges from environmental risk compared with **1COV.DE** and **MCHP**. And our investment suggestions are listed below:

- Fully consider the characteristics of the firm and consider the factors which might dominate the stock value.
- Both internal and external risks would on the value of stocks and will generate the fluctuation of prices in the stock market as well.
- Investors need to make the investment decision based on their risk tolerance and well balance differences in the risk-control of each company.
- Companies need to improve the ability of self-resilience and anti-risks so than enhance the performance when facing different types of risks.

5 Acknowledgement

6 References

Auguie, B. (2017). Gridextra: Miscellaneous functions for "grid" graphics [R package version 2.3]. https://CRAN.R-project.org/package=gridExtra

Müller, K, & Wickham, H. (2020). Tibble: Simple data frames [R package version 3.0.3]. https://CRAN.R-project.org/package=tibble

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Scrucca, L, Fop, M, Murphy, TB, & Raftery, AE. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. The R Journal, 8(1), 289–317.

Tierney, N. (2017). Visdat: Visualising whole data frames. JOSS, 2(16), 355.

Wickham, H. (2016). Ggplot2: Elegant graphics for data analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H, Averick, M, Bryan, J, Chang, W, McGowan, LD, François, R, Grolemund, G, Hayes, A, Henry, L, Hester, J, Kuhn, M, Pedersen, TL, Miller, E, Bache, SM, Müller, K, Ooms, J, Robinson, D, Seidel, DP, Spinu, V, . . . Yutani, H. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686.

Wickham, H, François, R, Henry, L, & Müller, K. (2020). Dplyr: A grammar of data manipulation [R package version 1.0.1]. https://CRAN.R-project.org/package=dplyr

Zhu, H. (2019). Kableextra: Construct complex table with 'kable' and pipe syntax [R package version 1.1.0]. https://CRAN.R-project.org/package=kableExtra

A Appendix.A

A.1 Ends with Emphasis

At the end of our report, it is necessary to emphasis that due to the small sample space and the incomplete eigenvalue selection, our result might not be representative and the biplot could not fully state the overall situation. Even though, in our case, the biplot is suitable for risk analysis. We still could not deny the fact that in general, the small sample size would lead to the bias in output. Therefore, our report uses cluster analysis as an alternative approach. The agglomerative method indicates that a stable solution is three clusters. And here we would show complete linkage method (Figure 10), average linkage (Figure 11) and centroid method (Figure 12). In order to check the robustness, we compute the adjusted rand index using adjustedRandIndex function. Table 9 indicates that the complete linkage method has a relatively high level of agreement with Ward's method.
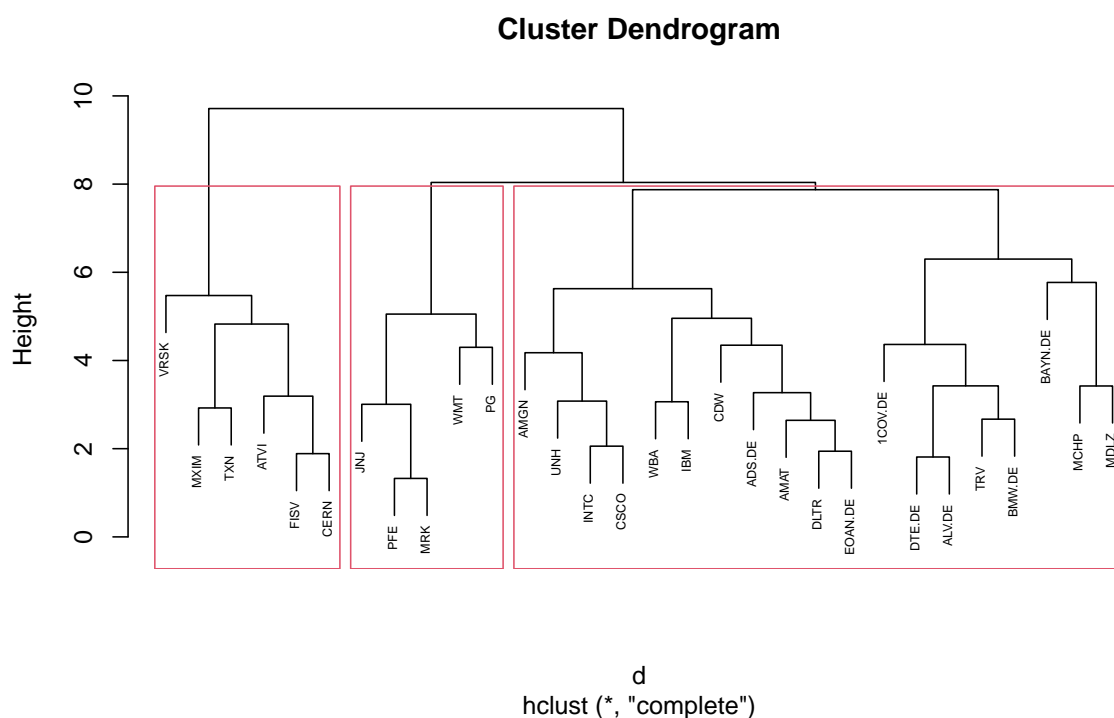
Figure 10: *Cluster dendrogram of complete linkage method*

Cluster Dendrogram

d
hclust (*, "average")
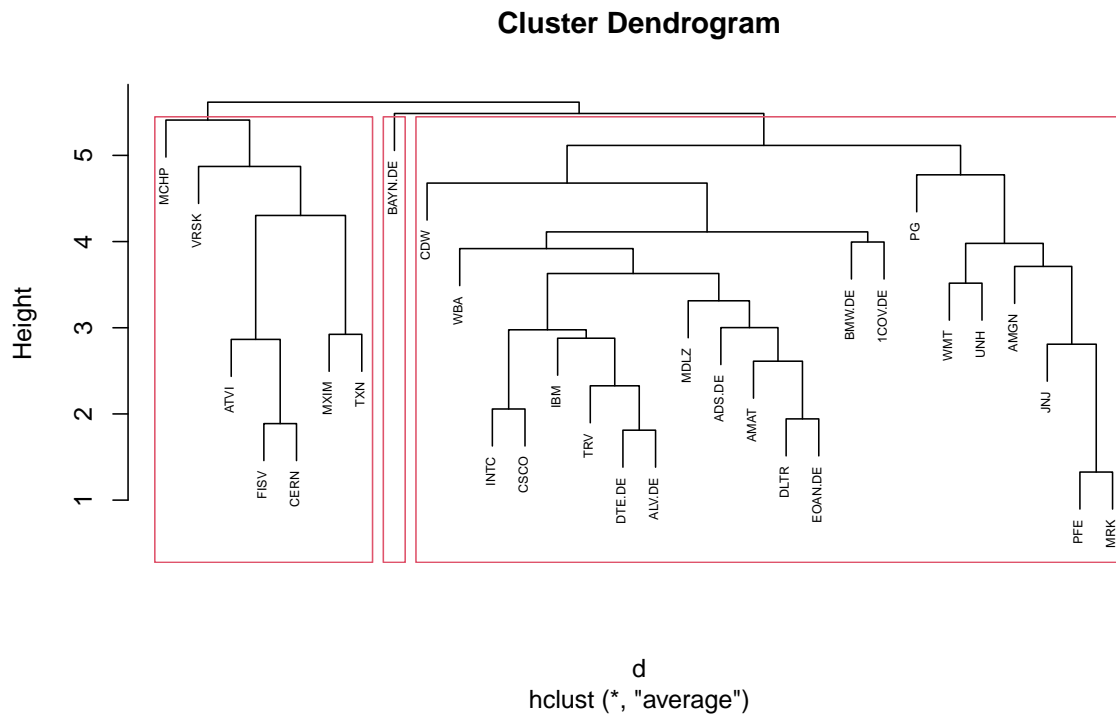
Figure 11: *Cluster dendrogram of average linkage method*

Cluster Dendrogram

d
hclust (*, "centroid")
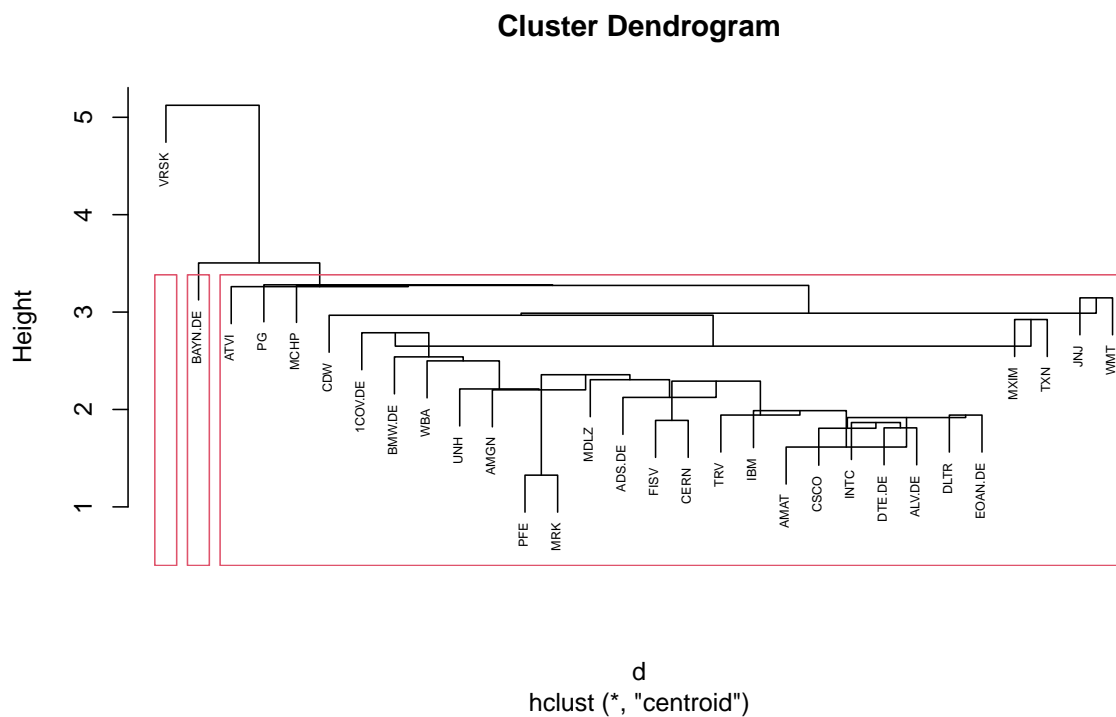
Figure 12: *Cluster dendrogram of centroid method*

Table 9: *The adjusted rand index of the three clustering methods*

	adjusted rand index
complete linkage method	0.7934295
average linkage method	0.4481954
centroid method	0.0919079

B Appendix.B

B.1 The codes used in the analysis

```r
knitr::opts_chunk$set(echo = FALSE, cache=FALSE, messages=FALSE, warning=FALSE)
# Make sure you have the latest versions of rmarkdown and bookdown installed
library(dplyr)
library(ggplot2)
library(tidyverse)
library(mclust)
library(visdat)
library(gridExtra)
library(kableExtra)
library(tibble)
```

B.1.1 Introduction

```r
# Read data

stocks_ori <- read_csv("stocks.csv")
#nrow(stocks)
#head(stocks)
stocks_ori <- stocks_ori%>%
  rename(intra_day=`Market cap (intra-day)`,
         ent_value=`Enterprise value`,
         trail_pe=`Trailing P/E`,
         for_pe=`Forward P/E`,
         peg=`PEG ratio (5-yr expected)`,
         ttm=`Price/sales (ttm)`,
         mrq=`Price/book (mrq)`,
         rev=`Enterprise value/revenue`,
         ebitda=`Enterprise value/EBITDA`,
         tot_risk = `Total ESG risk score`,
         envir_risk = `Environmental Risk Score`,
```

```
        social_risk = `Social Risk Score`,

        gover_risk = `Governance Risk Score`)
```

B.1.2   Data description

```
# Generate the table to explain the variables
tribble(~Variable, ~Abbreviation, ~Description,

        "Name", "/", "The full company name of each stock",

        "Symbol", "/", "The abbreviation of each stock",

        "Market", "/", "Major financial indices",

        "Sector", "/", "The belonging section of a stock",

        "Industry", "/", "The belonging industry of a stock",

        "Market capitalization", "intra_day", "How much a company is worth as

        determined by the stock market",

        "Enterprise value", "ent_value", "A measure of a company's total value",

        "Trailing P/E", "trail_pe",  "Price to Earning Ratio based on the earnings

        per share over the previous 12 months",

        "Forward P/E ratio", "for_pe",  "Estimate further earnings per share in

        the next 12 months",

        "PEG ratio", "peg", "Enhances the P/E ratio by adding the expected

        earnings growth into calculation",

        "P/S ratio", "ttm", "Price to Sales ratio, a valuation ratio by comparing a

        company's stock price to its revenue",

        "P/B ratio",  "mrq", "Price to Book ratio is a measurement of the market's

        valuation of a company relative to its book value",

        "Enterprise value-to-revenue",  "rev", "Also refers as the EV/R, it measures

        the value of a stock that compares a company's enterprise value to its revenue",

        "EV/EBITDA", "ebitda", "Enterprise value to earnings before interest, taxed,

        depreciation and amortization ratio compares the value of a company, debt

        included to the company's cash earnings less non-cash expenses",

        "Total ESG risk score", "tot_risk", "The overall rating scores based on

        the Morningstar Sustainability Rating systems",

        "Environmental Risk Score", "envir_risk", "Evaluation scores of the
```

```
        portfolios performance when they meet the environmental challenges",

        "Social Risk Score", "social_risk", "Evaluation scores of the portfolios

        performance when they meet the social challenges",

        "Governance Risk Score", "gover_risk","Evaluation scores of the portfolios

        performance when they meet the governance challenges") %>%

  kable(caption = "Information of variables of the original data") %>%

  kable_styling(bootstrap_options = c("bordered", "striped", "hover")) %>%

  column_spec(3, width = "170px")
```

B.1.3 Preliminary Analysis

```
# Visualise the data structure of the original data in Preliminary Analysis

vis_dat(stocks_ori)
```

```
# Summarise the data in Preliminary Analysis

sum_tab<-stocks_ori%>%

   select(-Name,-Symbol,-Market,-Sector,-Industry)%>%

   summary(stocks_ori)
```

```
# Summary table in Preliminary Analysis

tribble(~Variable, ~Min, ~Median, ~Mean, ~Max, ~`NA`,

        "intra_day","-2","63","95065","5110000","NA",

        "ent_value" ,"-264","70","85683","5130000","NA",

        "trail_pe" ,"0.48","20.11","43.62","1479.29", "18",

        "for_pe","3.59","19.92","43.04","1044.81", "80",

        "peg" ,"-62.380","2.405","15.223","713.670","81",

        "ttm" ,"0.9","2.8","9.941","548.150","17",

        "mrq" ,"0.1","5.4","174.16","11765.96","10",

        "rev" ,"-27.720","2.875","9.827","5411.160","17",

        "ebitda" ,"-465.460","13.765","19.461","1117.510","23",

        "tot_risk","11","23","25.39","75","13",

        "envir_risk","0","4","6.731","62", "13",

        "social_risk","3","10","11.4","88","13",
```

```
        "gover_risk","3","8" ,"9.343","80","13") %>%

  kable(caption = "Summary table of original data") %>%

  kable_styling(bootstrap_options = c("bordered", "striped", "hover"))
```

```
# Outliers table in Preliminary Analysis

tribble(~variable, ~outlier,

        "intra_day", "MSFT, AAPL",

        "ent_value", "MSFT, AAPL",

        "trail_pe", "TSLA",

        "for_pe", "ILMN, TSLA",

        "peg", "DIS, VZ, KO, MMM, CVX, PCAR, CAT, XOM",

        "ttm", "ILMN, V",

        "mrq", "TSLA",

        "rev", "ILMN, V",

        "ebitda", "INTU, ILMN, TSLA, NKE") %>%

  kable(caption = "The summary table of outliers in each variables") %>%

  kable_styling(bootstrap_options = c("bordered", "striped", "hover"))
```

```
# Filter out the missing values
stocks <- na.omit(stocks_ori)
```

B.1.4 Principal Component Analysis (Value Analysis)

```
# filter out the outliers
issuer<-stocks%>%

  select(Symbol,intra_day,ent_value,trail_pe,for_pe,peg,ttm,mrq,rev,ebitda)
df <- issuer[!(issuer$Symbol %in% c("MSFT", "AAPL", "TSLA", "ILMN", "DIS", "VZ",

                                    "KO", "MMM","CVX", "PCAR", "CAT", "XOM",

                                    "V", "INTU", "NKE")),]

#view(df)
```

```
# Create the principal component for value analysis
pca_value<-df%>%
```

```r
  select_if(is.numeric)%>%

  prcomp(scale.=TRUE)

rownames(pca_value$x)<-pull(df,Symbol)
```

```r
# Create the Correlation Biplot of the PCs for value analysis

cor_va<-biplot(pca_value,scale=0,cex=0.6)
```

```r
# Create the Distance Biplot of the PCs for value analysis

dis_va<-biplot(pca_value,xlim=c(-0.6,0.3),cex=0.6)
```

```r
# Create the summary table of principal components in value analysis

summary(pca_value)$importance %>%

  round(4) %>%

  kable(caption = "Summary table of PCA for value analysis of stocks") %>%

  kable_styling(bootstrap_options = c("bordered", "hover", "striped")) %>%

  column_spec(1, width = "70px")
```

```r
# Create the Screeplot to select the PCs in value analysis

screeplot(pca_value, type = "lines")
```

B.1.5 Principal Component Analysis (Risk Analysis)

```r
# filter out the outliers

risk <- subset(stocks[(stocks$Symbol %in% df$Symbol),],

            select = c("Symbol",

                    "tot_risk",

                    "envir_risk",

                    "social_risk",

                    "gover_risk"))

view(risk)
```

```r
# First check whether total esg risk score =

# environmental risk + social risk + governance risk
```

```r
d <- apply(cbind(risk$envir_risk,

                 risk$social_risk,

                 risk$gover_risk),

           1, sum)

all(risk$tot_risk == d)
```

```r
# Second check whether total esg risk score =

# environmental risk + social risk + governance risk

risk_tesg <- risk[risk$tot_risk == d,]

d2 <- apply(cbind(risk_tesg$envir_risk,

                  risk_tesg$social_risk,

                  risk_tesg$gover_risk),

            1, sum)

all(risk_tesg$tot_risk == d2)

#head(risk_tesg)

#nrow(risk_tesg)
```

```r
# Execute the PCA for risk analysis

pca_risk <- risk_tesg %>% select_if(is.numeric) %>% prcomp(scale. = TRUE)
```

```r
# Create the summary table of principal components in risk analysis

summary(pca_risk)$importance %>%

  round(4) %>%

  kable(caption = "Summary table of PCA for risks analysis of stocks") %>%

  kable_styling(bootstrap_options = c("bordered", "hover", "striped"))
```

```r
# Create the Screeplot to select the PCs in risk analysis

screeplot(pca_risk, type = "lines")
```

```r
# Create the Distance Biplot of the PCs for risk analysis

rownames(pca_risk$x) <- pull(risk_tesg, Symbol)

biplot(pca_risk, xlim=c(-0.5, 0.5), ylim = c(-0.6, 0.3),  cex = 0.6)
```

```r
# Create the Correlation Biplot of the PCs for risk analysis
biplot(pca_risk, scale = 0, xlim = c(-4, 4), cex = 0.6)
```

B.1.6 Cluster Analysis

```r
# Filter out the outliers, and generate the distance between each observation.
dat <- stocks[(stocks$Symbol %in% df$Symbol),]
d <- dat %>% select_if(is.numeric) %>% scale %>% dist
```

```r
# Create different cluster dendrogram
hc_w <- hclust(d, method = "ward.D2")
hc_w$labels <- dat$Symbol
par(mfrow = c(1,3))
plot(hc_w, cex = 0.5, xlab= "Stocks", ylab="Distance")
rect.hclust(hc_w, k = 2)
plot(hc_w, cex = 0.5, xlab= "Stocks", ylab="Distance")
rect.hclust(hc_w, k = 3)
plot(hc_w, cex = 0.5, xlab= "Stocks", ylab="Distance")
rect.hclust(hc_w, k = 4)
```

```r
# The three-cluster dendrogram
plot(hc_w, cex = 0.5, xlab= "Stocks", ylab="Distance")
rect.hclust(hc_w, k = 3)
```

```r
# The first cluster
memb_three_w <- cutree(hc_w, k = 3)
memb_one <- dat[dat$Symbol %in%
                  rownames(as.data.frame(memb_three_w) %>%
                      filter(memb_three_w == 2)),]$Name


memb_one %>%
  kable(col.names = c("stock"), caption = "The stocks of the first cluster") %>%
  kable_styling(bootstrap_options = c("bordered", "hover", "striped"))
```

```r
# The second cluster
memb_two <- dat[dat$Symbol %in%

                    rownames(as.data.frame(memb_three_w) %>%

                             filter(memb_three_w == 3)),]$Name


memb_two %>%

  kable(col.names = c("stock"),

        caption = "The stocks of the second cluster") %>%

  kable_styling(bootstrap_options = c("bordered", "hover", "striped"))
```

```r
# The third cluster
memb_three <- dat[dat$Symbol %in%

                    rownames(as.data.frame(memb_three_w) %>%

                             filter(memb_three_w == 1)),]$Name


memb_three %>%

  kable(col.names = c("stock"),

        caption = "The stocks of the third cluster") %>%

  kable_styling(bootstrap_options = c("bordered", "hover", "striped"))
```

B.1.7 Appendix.A

```r
# Cluster dendrogram of complete linkage method
hc_c <- hclust(d, method = "complete")
hc_c$labels <- dat$Symbol
plot(hc_c, cex = 0.5)
rect.hclust(hc_c, k = 3)
```

```r
# Three-cluster solution of complete linkage method
memb_three_c <- cutree(hc_c, k = 3)
```

```r
# Cluster dendrogram of average linkage method
hc_al <- hclust(d, method = "average")
```

```r
hc_al$labels <- dat$Symbol

plot(hc_al, cex = 0.5)

rect.hclust(hc_al, k = 3)


# Three-cluster solution of average linkage method

memb_three_al <- cutree(hc_al, k = 3)


# Cluster dendrogram of centroid method

hc_cm <- hclust(d, method = "centroid")

hc_cm$labels <- dat$Symbol

plot(hc_cm, cex = 0.5)

rect.hclust(hc_cm, k = 3)


# Three-cluster solution of centroid method

memb_three_cm <- cutree(hc_cm, k = 3)


# Create the table for the adjusted rand index of that three clustering methods

c <- adjustedRandIndex(memb_three_w, memb_three_c)

al <- adjustedRandIndex(memb_three_w, memb_three_al)

cm <- adjustedRandIndex(memb_three_w, memb_three_cm)



al_method <- rbind(c, al, cm)

rownames(al_method) <- c("complete linkage method",

                         "average linkage method",

                         "centroid method")

al_method %>%

  kable(col.names = c("adjusted rand index"),

        caption = "The adjusted rand index of the three clustering methods") %>%

  kable_styling(bootstrap_options = c("bordered", "hover", "striped"))
```