MONASH
BUSINESS
SCHOOL

Expert advice from experts

Kaiwen Jin
1111111

Zhiruo Zhang
2222222

Jinhao Luo
Master of Business Analytics (enrolled)

Report for
ETF5500

30 September 2020

Department of Econometrics & Business Statistics

(03) 9905 2478
BusEco-Econometrics@monash.edu

ABN: 12 377 614 012

Contents

1 Introduction

Principal component analysis (PCA) plays an important role in the evaluation of the stock price. By utilising the small number of the principle components to explain the general variation in the original data. This report will apply the PCA method to address whether it is a good measurement in the stock pricing. By investigating the measurement of stock value and stock risk to further find our the potential limitation that PCA might face. We will also compare the performance of PCA with another method which is Clustering Analysis to check the accuracy of our result.

In financial market, the value of stocks would be investigated by many different variables. However, the large number of variables of each stock might make readers hard to make the comparison. Therefore, linear combination (LC) would be applied to combine all the variables into a index and principal component analysis (PCA) would be utilised to help with evaluating stocks. Principal components (PCs) are the indexes of linear combination which explain the variance of the original variables, and the variance could help with differentiating the performing of each observations. PCA would use the principal components to evaluate the potential value and risk of each stock. In addition, cluster analysis would also be performed in this report to analyse all the stock in the dataset. At last, some useful suggestions for the stocks choosing will be concluded, as well as concluding the biases generated from the limitations in analysis.

2 Data Description

```
## Parsed with column specification:
## cols(
##   Name = col_character(),
##   Symbol = col_character(),
##   Market = col_character(),
##   Sector = col_character(),
##   Industry = col_character(),
##   `Market cap (intra-day)` = col_double(),
##   `Enterprise value` = col_double(),
##   `Trailing P/E` = col_double(),
##   `Forward P/E` = col_double(),
##   `PEG ratio (5-yr expected)` = col_double(),
##   `Price/sales (ttm)` = col_double(),
```

```
##   `Price/book (mrq)` = col_double(),
##   `Enterprise value/revenue` = col_double(),
##   `Enterprise value/EBITDA` = col_double(),
##   `Total ESG risk score` = col_double(),
##   `Environmental Risk Score` = col_double(),
##   `Social Risk Score` = col_double(),
##   `Governance Risk Score` = col_double()
## )
```

The data which used in this report was sourced from Yahoo Finance. (Table 2 shows the information of the variables from the original data, as well as the abbreviation of the variables.) The dataset contains 18 variables of 147 stocks from five major financial indices. Those 18 variables could be further classified into 3 categories. The first categories captures the 5 variables provides the basic background of those stocks which are **Name**, **Symbol**, **Market**, **Sector**, **Industry**. 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 are shown in the table below:

Names	abbreviation	Description
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 revenues
P/B ratio:	mrq	Price to Book ratio is a measurement of the market's valuation of a company relative to its book value.

Table 2: *Information of variables of the original data*

Names	Abbreviation	Description
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 prev
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
P/S ratio	ttm	Price to Sales ratio, a valuation ratio by comparing a company's stoc
P/B ratio	mrq	Price to Book ratio is a measurement of the market's valuation of a
Enterprise value-to-revenue	rev	Also refers as the EV/R, it measures the value of a stock that compa
EV/EBITDA	ebitda	Enterprise value to earnings before interest, taxed, depreciation and
Total ESG risk score	tot_risk	The overall rating scores based on the Morningstar Sustainability Ra
Environmental Risk Score	envir_risk	Evaluation scores of the portfolios performance when they meet the
Social Risk Score	social_risk	Evaluation scores of the portfolios performance when they meet the
Governance Risk Score	gover_risk	Evaluation scores of the portfolios performance when they meet the

Names	abbreviation	Description
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.

2.1 Description of the variables related to the value of stocks:

- Market capitalization refers to how much a company is worth as determined by the stock market. It is defined as the total market value of all outstanding shares. To calculate a company's market cap, multiply the number of outstanding shares by the current market value of one share. Companies are typically divided according to market capitalization: large-cap ($10 billion or more), mid-cap ($2 billion to $10 billion), and small-cap ($300 million to $2 billion). Enterprise value includes in its calculation the market capitalization of a company but also short-term and long-term debt as well as any cash on the company's balance sheet. Enterprise value is used as the basis for many financial ratios that measure the performance of a company.
- Enterprise value (EV) is a measure of a company's total value, often used as a more comprehensive alternative to equity market capitalization.
- Trailing P/E is calculated by dividing the current market value, or share price, by the earnings per share over the previous 12 months.
- The forward P/E ratio estimates a company's likely earnings per share for the next 12 months.
- The PEG ratio enhances the P/E ratio by adding in expected earnings growth into the calculation. The PEG ratio is considered to be an indicator of a stock's true value, and similar to the P/E ratio, a *lower PEG may indicate that a stock is undervalued*.
- The P/S ratio is a key analysis and valuation tool that shows *how much investors are willing to pay per dollar of sales for a stock.* The P/S ratio is typically calculated by dividing the stock price by the underlying company's sales per share.A low ratio could imply the stock is undervalued while a ratio that is higher-than-average could indicate that the stock is overvalued.
- The P/B ratio measures the market's valuation of a company relative to its book value.*The market value of equity is typically higher than the book value of a company.* P/B ratio is used by value investors to identify potential investments. P/B ratios under 1 are typically considered solid investments.
- The enterprise value-to-revenue (EV/R) multiple helps compares a company's revenues to its enterprise value. *The lower the better, in that, a lower EV/R multiple signals a company is undervalued.*
- The enterprise value to earnings before interest, taxes, depreciation, and amortization ratio (EV/EBITDA) compares the value of a company—debt included—to the company's cash earnings less non-cash expenses. The EV/EBITDA metric is a popular valuation tool that helps investors compare companies in order to make an investment decision. EV calculates a company's total value or assessed worth, while EBITDA measures a company's overall financial performance

and profitability. Typically, when evaluating a company, *an EV/EBITDA value below 10 is seen as healthy*. It's best to use the EV/EBITDA metric when comparing companies within the same industry or sector.

However, there are some limitations of the original dataset. Firstly, there are some missing values in the data, which might generate inaccuracy for analysing the data. Secondly, there are only 147 observatios contain in the data, which might insufficient enough to support the analysis and credible results. In addition, if filter out the observations with missing values, the sample size of the data might much less and would increase errors. Thirdly, some inconsistency between total risk score and sum of the other three risk score appear in the dataset, which could also increase the errors of analysis. Those limitations would be further discussed in following sections. At last, the biases in analysis which generate from the limitations would be concluded.

3 Analysis

3.1 Preliminary Analysis

Before we start our Principal Components Analysis, we will firstly observe the original data and tidy it by removing the missing variables and further figure out any other features. Figure 1 shows the data structure of the original dataset. It is clear that there are three different types of data contain in the dataset, which are character, numeric and missing value.

By further observing the missing value, figure 2 is generated. we could find that there are 11.3% missing value included in the original dataset. Excepted the five seven variables, the rest variables all contain the missing values.

Meanwhile, by summarising original variables, table 3 indicates that the initial 147 observations have up to 102 missing value.

Names	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

Names	Min	Median	Mean	Max	NA
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

In the meanwhile, table 3 shows us the information about the ourliers. Most of the variables we have here, they all have the really small median and mean, but a extremely high maximum value as well. We could say that the extremely value will definitely dominate our Principle component analysis. And here are the outliers we get:

- outliers in `intra_day` and `ent_value`: MSFT & AAPL
- outliers in `trail_pe`: TSLA
- outliers in `for_pe`: ILMN & TSLA
- outliers in `peg`: DIS, VZ, KO, MMM, CVX, PCAR, CAT, XOM
- outliers in `ttm`: ILMN, V
- outliers in `mrq`: TSLA
- outliers in `rev`: ILMN, V
- outliers in `ebitda`: INTU, ILMN, TSLA, NKE

We will generate a new dataset **stocks** by removing the missing value and conduct our following analysis based on our new dataset.

3.2 Principle Component Analysis

3.2.1 Value Analysis

Based on the preliminary analysis we have above, besides of the missing value removement, what we also need to do is the outliers removement. Since our variables are measured in the different unit, after removing the high influential values, we will standardised our data by using the **scale.** function in R. Biplot could visualise the data by selecting two principal components. The graphs and interpretation will also be provided below.

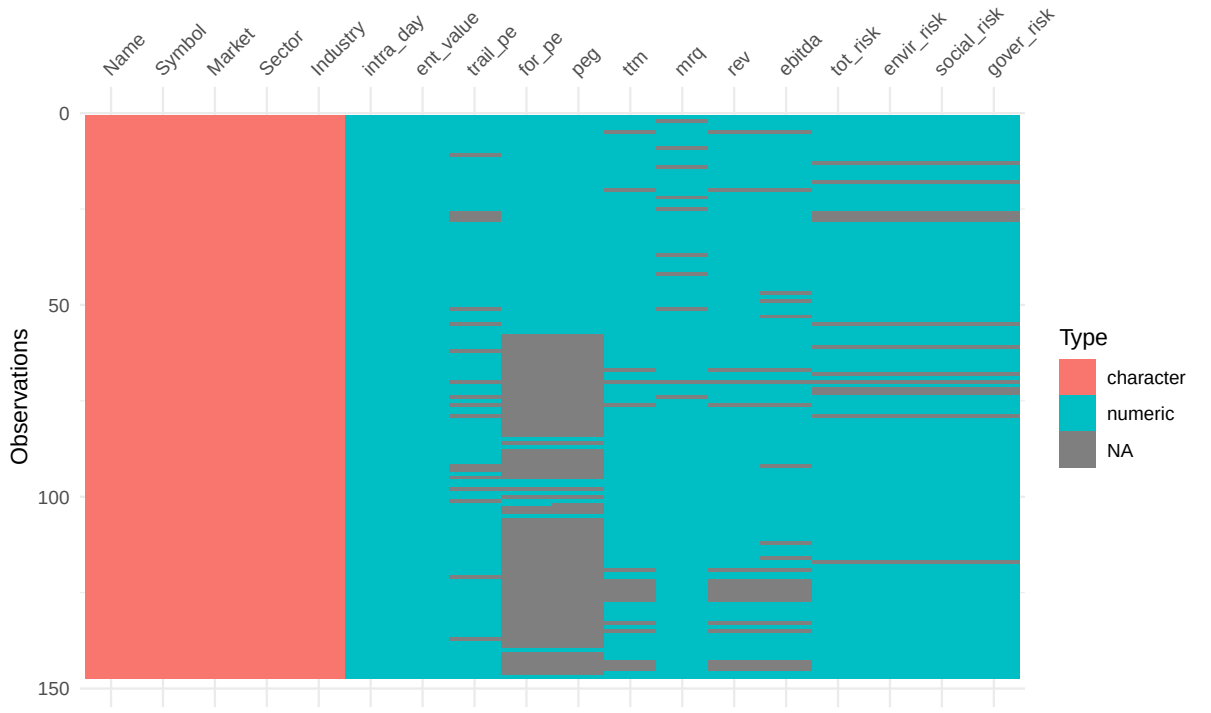


Figure 1: *The data structure of original data*

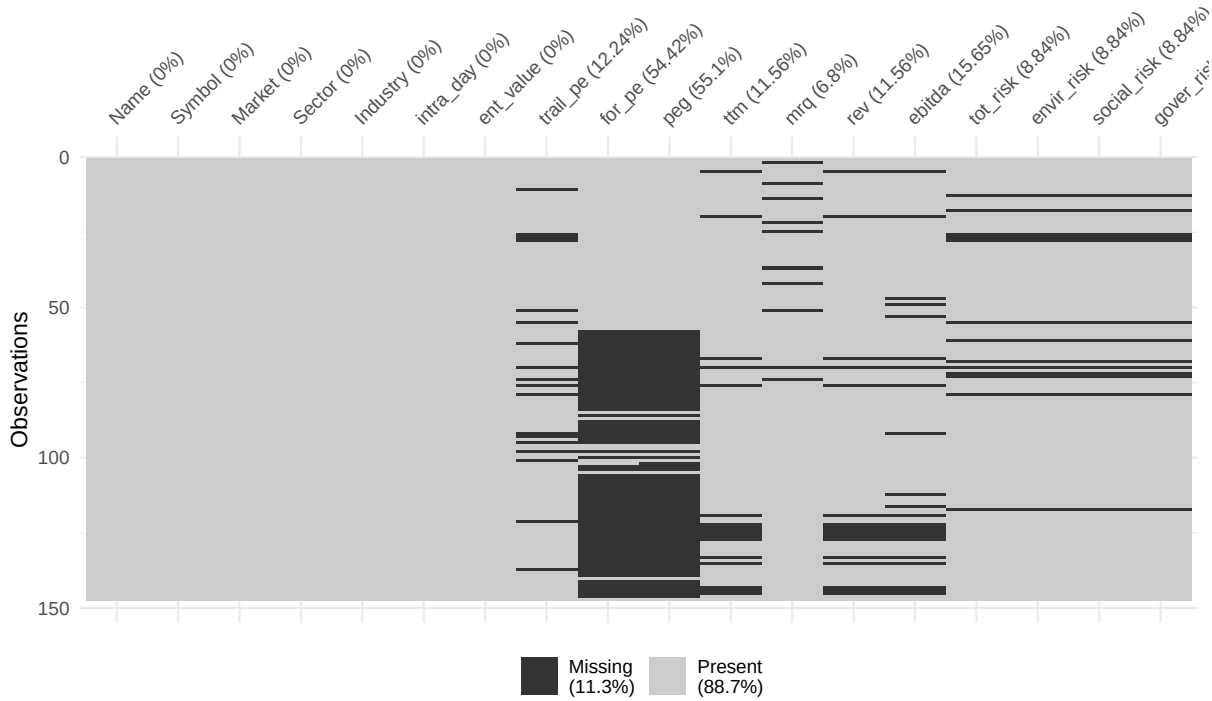


Figure 2: *The missing values of original data*

Table 3: *Summary table of original data*

Names	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

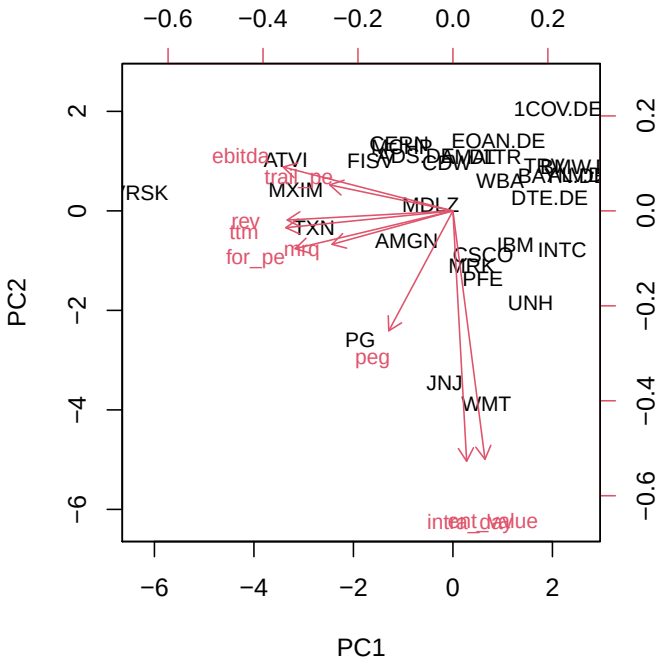


Figure 3: *(#fig:pca_cor)Correlation Biplot of Stock Value*

Referring to the correlation biplot Figure **??**fig:cor_va)we could notice that the the PC1 is positive correlated with the measurement of the company value indication which are **intra_day** and **ent_value** even if the correlation is pretty not strong. The PC2 is positive correlated with the stock earning ratio (ebitda and trail_pe) which means that the increasing in the measurement of the stock earning ratio will increase the PC2 slightly. The rest of the ratio are neither postive correlated with PC1 nor PC2, but we could notice that the other variables which are related to the price based evaluation of the

stock are pretty close to the PC2. The **peg** ratio could not be well explained by both PC1 and PV2. In the meanwhile, this plot also highlights that the two measurement of the company value have a really strong association with each other and do not have any association with other variables which related to the stock price and earning evaluation. Therefore, we could say that the market value of a company may not influence on their stock price and earning per share. However, the relationship between those stock price and earning measurement are quite strong.

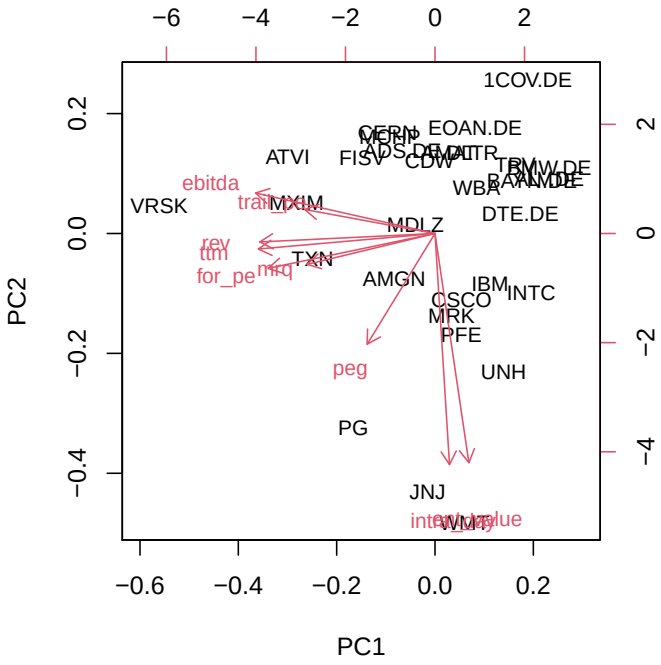


Figure 4: *(#fig:bi_dis)Distance Biplot of Stock Value*

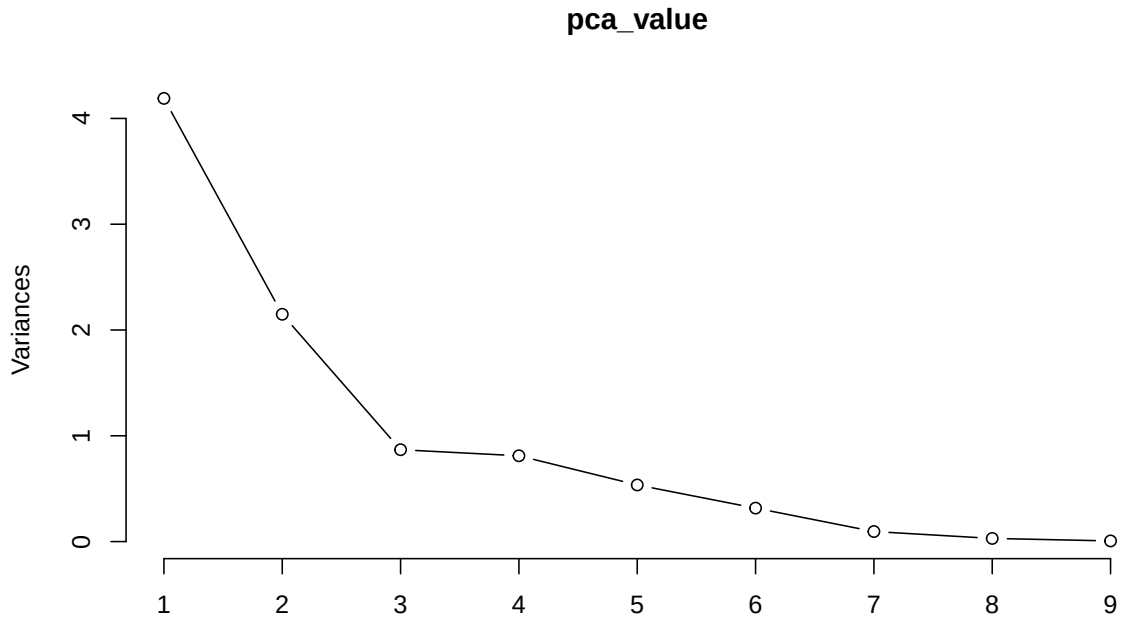
After we identify the relationship between those measurement of the variable, what we do next is to figure out the connection of each observation by using the distance biplot **??**fig:dis_va). What this biplot could tell us is that **Johnson & Johnson** and **Walmart** have a pretty high value of the company, and **Activision Blizzard**, **Texas Instruments Incorporated**, **Maxim Integrated Products** indicate the high earnings in the stock.

In the meanwhile, we also need to pay attention on the **Verisk analytics** the potential outlier for the PC1, and **Johnson & Johnson** and **Walmart** the potential outlier for PC2. Since we notice that the price and earning per share for **VRSK** are quite high the reason may due to that **VRSK** is mainly a data analytics and risk assessment firm. The mainly provide the consulting service instead of the selling goods. Therefore, compaing with the multinational retail company, the might not have a really

large firm size, as for a financial sector, their stock value is postively correlated with the service they provide.

JNJ and **WMT** perform just in the opposite way. As for the multination company, their profit mainly come from the selling products. Therefore, they need to continuously increase their market share to maintain their market profit.

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.0468 1.4658 0.93188 0.90054 0.73163 0.56248 0.30811
## Proportion of Variance 0.4655 0.2387 0.09649 0.09011 0.05948 0.03515 0.01055
## Cumulative Proportion  0.4655 0.7042 0.80072 0.89083 0.95030 0.98546 0.99600
##                            PC8     PC9
## Standard deviation     0.17270 0.07838
## Proportion of Variance 0.00331 0.00068
## Cumulative Proportion  0.99932 1.00000
```



Besides of the information we get from the biplot, there are also some of limitations in our value analysis. The first limitation is mainly due to our small sample space. Based on the 30 out of 147 variables, there are 70.42% of the overall variation could be explained by the first two principle.

Table 5: *Summary table of PCA*

	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

It indeed could explain the sufficient amount of the variables but if we take the whole variables into consideration, this might not be accurate enough and also not very representative. Therefore, alternative approach is required. Another limitation is that about the number of eigenvalue selection. Screeplot suggests that the better PC value we need to include is three and this is a little bit contradict to our biplot.

Overall we could say that PCA in the value analysis in our sample stocks could express some information, but it might not be really representative in the general stock market.

3.2.2 Risk Analysis

Besides evaluating the price and value of those stocks, the reports would also analysis the potential risk of each stock based on the risk score of stocks, as well as make the comparison. Firstly, for the purpose of improving the accuracy of PCA for analysing the risk of stocks, the report has compared the total ESG risk score with the sum of the rest three risk scores and filtered out the inconsistent observations, to ensure the consistency in order to avoid errors and improve the accuracy of analysis. Before executing PCA, which principal components need to be utilised should be considered. Table 5 shows the summary statistics of components. It is clear that PC1 and PC2 have explained almost 86% of the total variation of 4 variables. In addition, figure 5 is a scree plot which indicates the variance explanation of each principal component. According to the Kaiser's Rule, principal component of one and two would be selected because they all with a variance greater than 1. Therefore, PC1 and PC2 are the two principal components applied to the following analysis.

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

```
## [1] TRUE
```

Figure 6 is the distance biplot which shows the distance among each stock in the dataset, and implies the similarity between stocks. Based on the figure 6,
the stocks of **Verisk Analytics, Inc. (VRSK)** and **UnitedHealth Group Incorporated (UNH)** may be exactly same because they seem likely perfectly superimpose. Besides, **Allianz SE (ALV.DE)** and

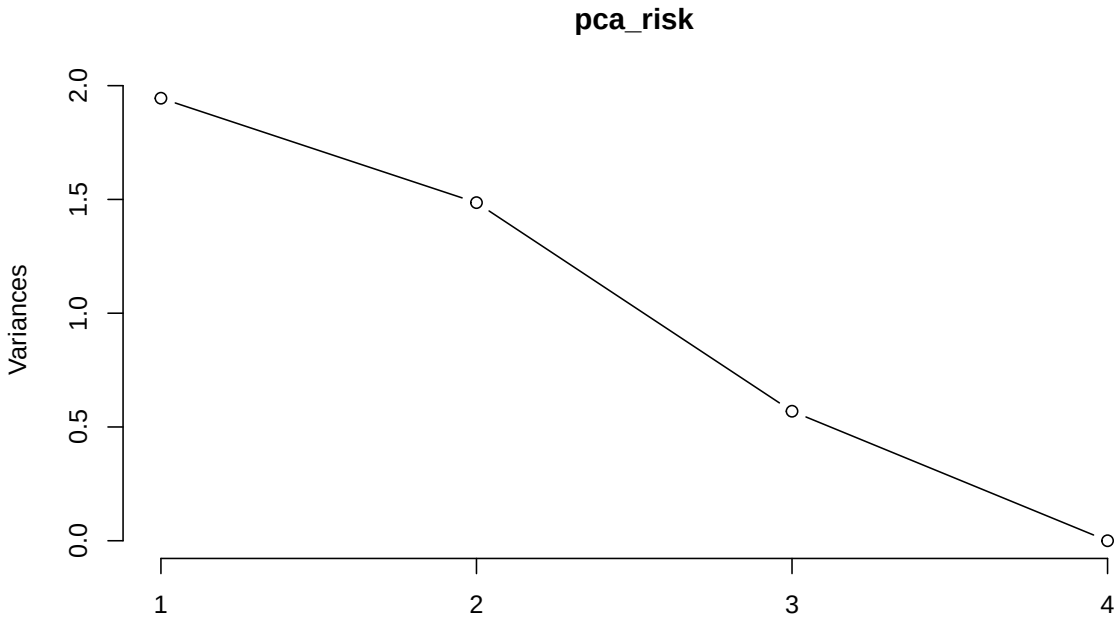


Figure 5: *Screeplot of PCs in PCA*

Dollar Tree, Inc. (DLTR), as well as **Cerner Corporation (CERN)** and **Fiserv, Inc. (FISV)** might be similar, because they are close to each other. While the stocks like **VRSK** and **Microchip Technology Incorporated (MCHP)**, or

CDW Corporation (CDW) and

Pfizer Inc. (PFE) might be different because they are far away from each other. In order to further analyse the correlation between each stock, a correlation biplot is required.

Figure 7 is the correlation biplot which explains the correlation between different risk scores, as well as the correlation between different stocks. Or even allow readers to compare stocks to different types of risk. According to figure 7, **VRSK**, **UNH**, **CERN**, and **FISV** are with the high values of social risk score, which could indicate that these four stock might have strong resilience against the social challenges and might perform better than the other stocks when facing the social problems. While, those stocks might not be good at facing the challenges from the internal or external environment because the angle between variable of social risk score and variable of environmental risk is close to 180 degree, which might imply the highly negative correlation approximately. On the contrary, **Covestro AG (1COV.DE)** and **MCHP** are the two stocks seem likely with the strongest abilities to face the environmental challenges. In addition, **MCHP** might also the stock with the highest score of governance risk, which could indicate good performance when meeting the governance challenges.

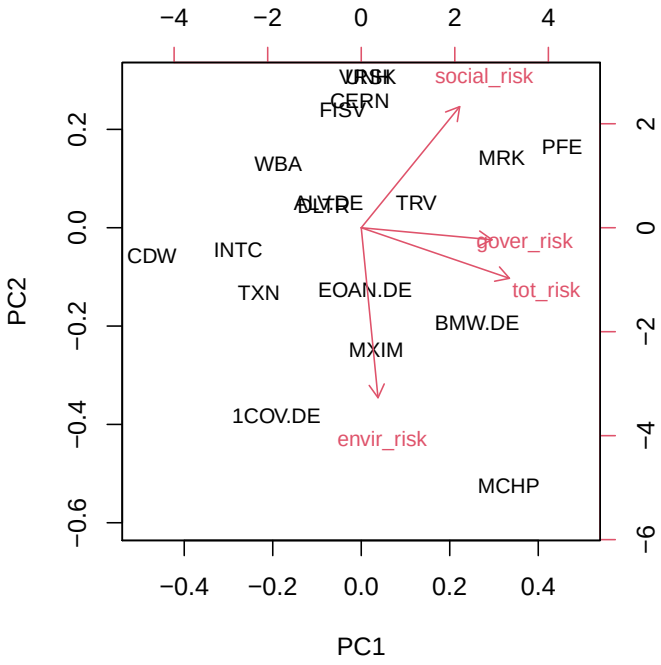


Figure 6: *Distance biplot of PCA of stocks' risk*

Meanwhile, **PFE**, **Bayerische Motoren Werke AG (BMW.DE)**, and **Merck & Co., Inc. (MRK)** are the another three stocks also perform well in governance challenges. While the performance of these three stocks might gradually decline. Because the projected positions of these three stocks along the axis of governance risk score reflect the gradually decreasing trend of approximate actual values. Furthermore, comparing stocks to total ESG risk score might give readers a comprehensive sense of the overall performance of each stock. In general, **MCHP** and **BMW.DE** are the stock with the best overall performance compared with other stocks, which indicate that they might be hard to be influences by internal and external challenges, and they have strong resilience when meeting that three risks. Therefore, there might not be significant fluctuations of these stocks when facing challenges, and could be seen as stable stocks. In contrast, the stock of **CDW** has a weak overall performance when facing challenges. Based on the approximation actual value which generates by projecting **CDW** to the axis of total risk score, it has a low value. It indicates that the risks might impact on **CDW** easily, and **CDW** might experience a significant fluctuation when facing risks.

3.3 Cluster Analysis

Prefer using Ward's method..

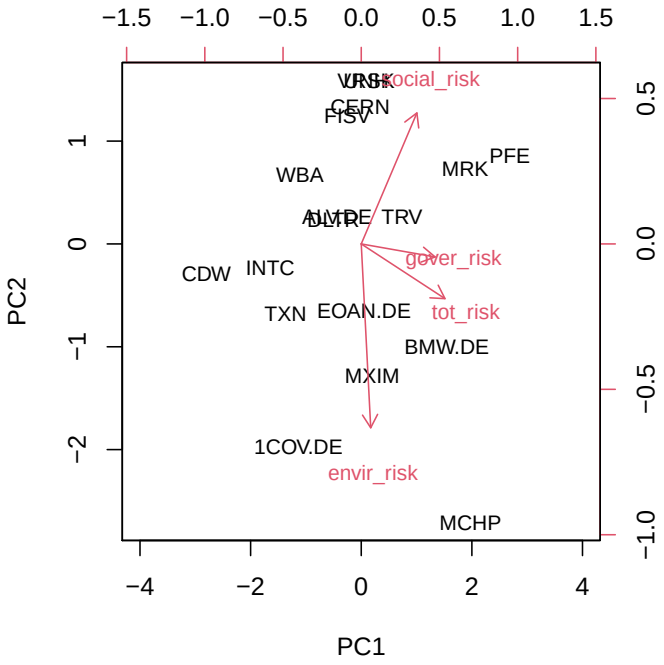
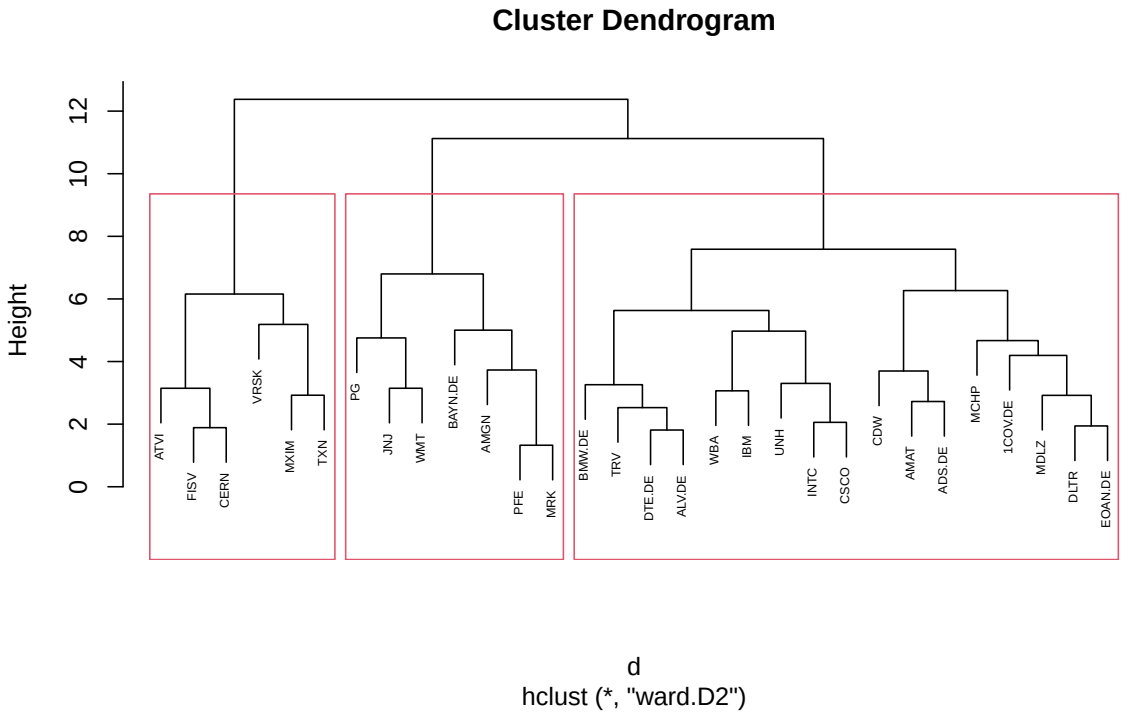
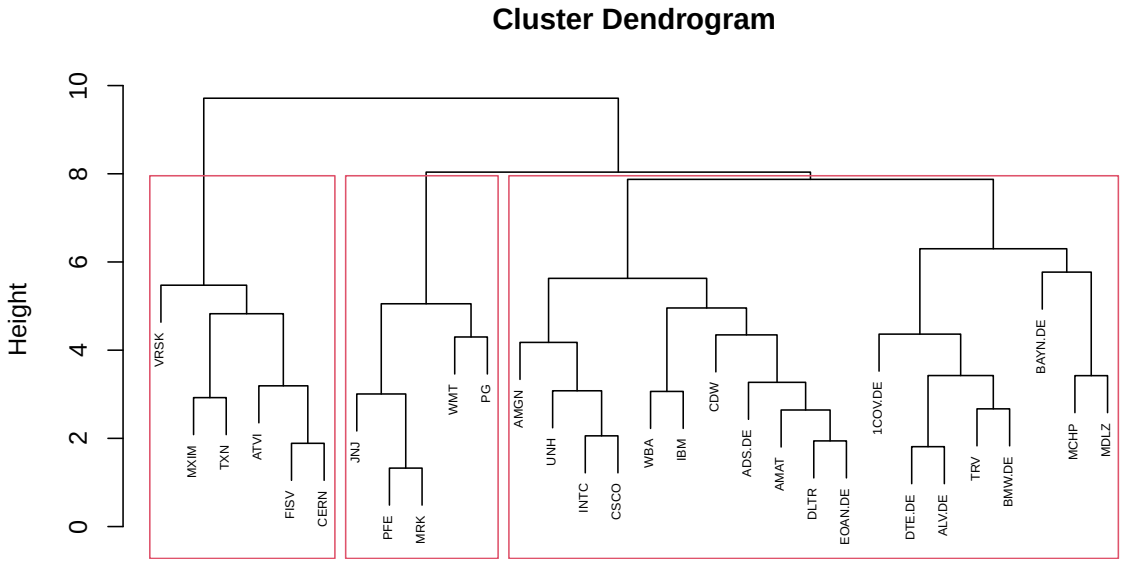
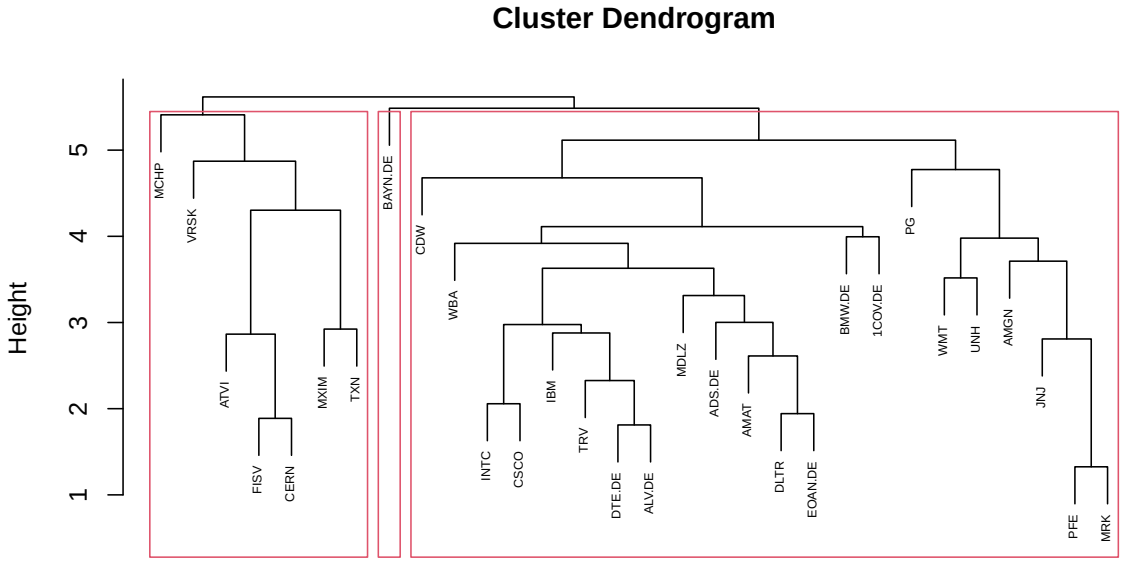
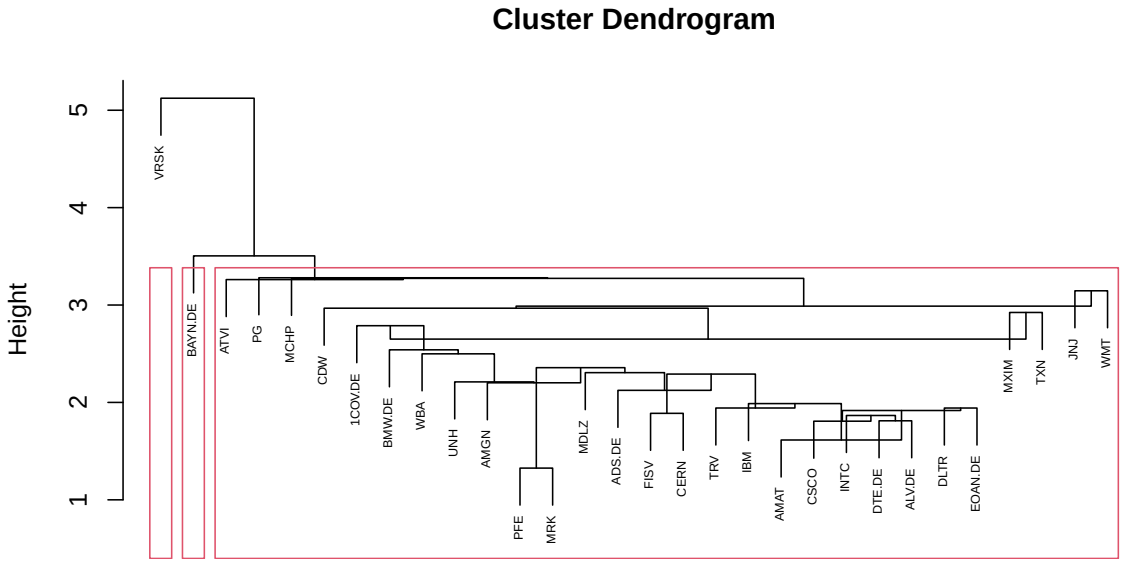


Figure 7: *Correlation biplot of PCA of stocks' risk*




Cluster Dendrogram
Height
0
2
4
6
8
10
VRSK
MXIM
TXN
ATVI
FISV
CERN
JNJ
PFE
MRK
WMT
PG
AMGN
UNH
INTC
CSCO
WBA
IBM
CDW
ADS.DE
AMAT
DLTR
EOAN.DE
1COV.DE
DTE.DE
ALV.DE
TRV
BMW.DE
BAYN.DE
MCHP
MDLZ
d
hclust (*, "complete")


Cluster Dendrogram
Height
1
2
3
4
5
MCHP
VRSK
ATVI
FISV
CERN
MXIM
TXN
BAYN.DE
CDW
WBA
INTC
CSCO
IBM
TRV
DTE.DE
ALV.DE
MDLZ
ADS.DE
AMAT
DLTR
EOAN.DE
BMW.DE
1COV.DE
PG
WMT
UNH
AMGN
JNJ
PFE
MRK
d
hclust (*, "average")




```
## [1] 0.7934295
```

```
## [1] 0.4481954
```

```
## [1] 0.09190793
```

4 Conclusions

5 Acknowledgement