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

3 October 2020

Department of Econometrics & Business Statistics

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

ABN: 12 377 614 012

Contents

1 Introduction

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 investors hard to make their decision. Therefore, this report would apply linear combination (LC) to combine all the variables into a index and utilise principal component analysis (PCA) to evaluate the performance of stocks.

In addition, this report will also consider the accuracy of PCA and will discuss about the potential limitation of PCA in the stock performance evaluation. Based on the result, the comparative analysis with Clustering 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 limitations in analysis.

The appendix will contain some notes which would be helpful in understanding our reports.

2 Data Description

2.1 Description

The data which used in this report was sourced from Yahoo Finance. Table 1 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:

2.2 Limitation

In this part we will briefly introduce a couple of limitations in our dataset and those limitation will also be discussed in the following section.

- 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. In addition, 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.

Table 1: *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 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

- There is some inconsistency between total ESG risk score and sum of individual ESG risk score. This inconsistency would directly increase the error in our final output.

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

Based on the original dataset, we will firstly tidy it 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.

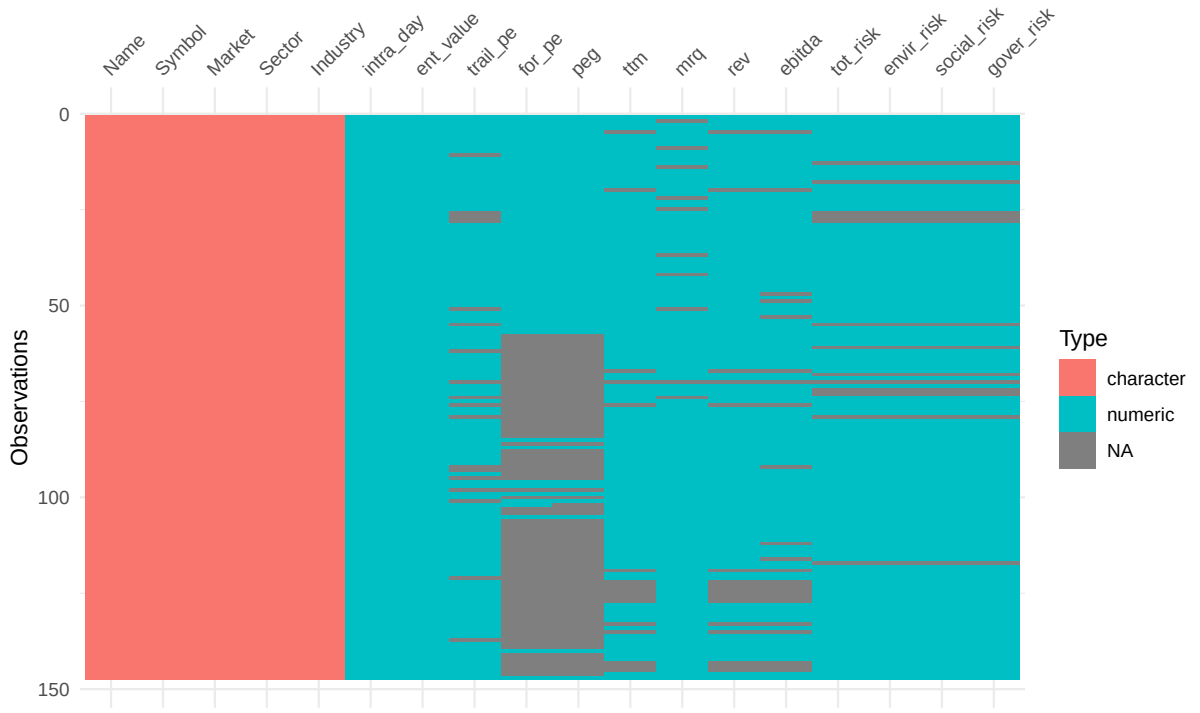

Figure 1: *The data structure of original data*

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

Most of the variables have the really small median and mean, but a extremely high maximum value. Those extreme value would definitely dominate our Principle Component Analysis and those outliers are shown below:

Table 2: *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

- 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

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

3.2 Principle Component Analysis

3.2.1 Value Analysis

Value analysis will be conducted by removing the outliers. It is necessary to filter out high influential outliers and we will standardised our data as well since it is based on different unit. Biplot and interpretation will also be provided.

Referring to the correlation biplot Figure 2 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**. 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 positive correlated with PC1 nor PC2, but we could notice that the other variables

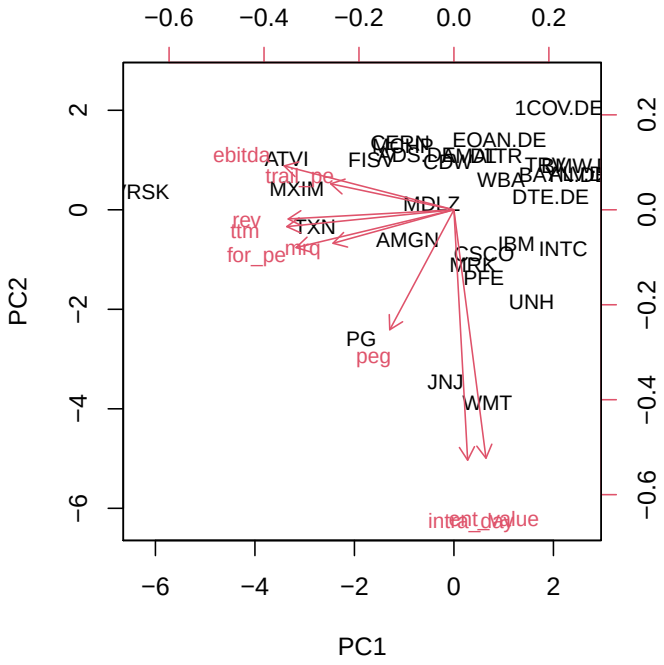

Figure 2: *Correlation Biplot of Stock Value*

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.

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.

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.

Meanwhile, we notice that the **Verisk analytics (VRSK)** the potential outlier for the PC1, and **JNJ** and **WMT** the potential outlier for PC2. We could explain the reason by identify the characteristic of these firm. **VRSK** is a data analytics and risk assessment firm. They mainly provide the consulting service instead of the selling goods. Therefore, being a financial service sector, they will not have a large firm size, but the stock value and EPS will be higher. **JNJ** and **WMT** perform in the opposite

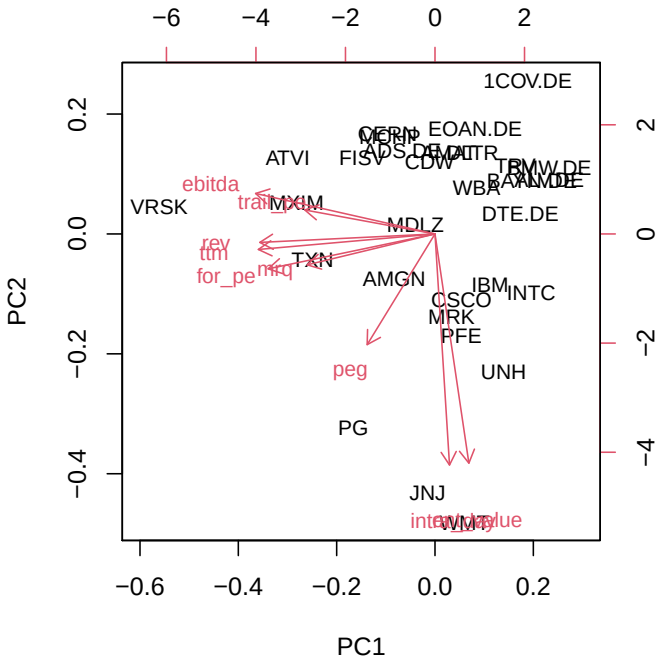

Figure 3: *Distance Biplot of Stock Value*

Table 3: *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

way because they mainly generate profit by selling goods. The continuously increasing market share will keep their market profit in a high level.

The limitation for the value analysis also exists. - 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 principle (Table 3). The really small space size is not representative and also would not accurate enough to explain the whole stock market condition. - There is some contradictory in PC selection in Screeplot (Figure 4) and biplot. Therefore, we need to consider alternative approach to make sure of the accuracy of our suggestion.

3.2.2 Risk Analysis

In this part, we will discuss 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 result.

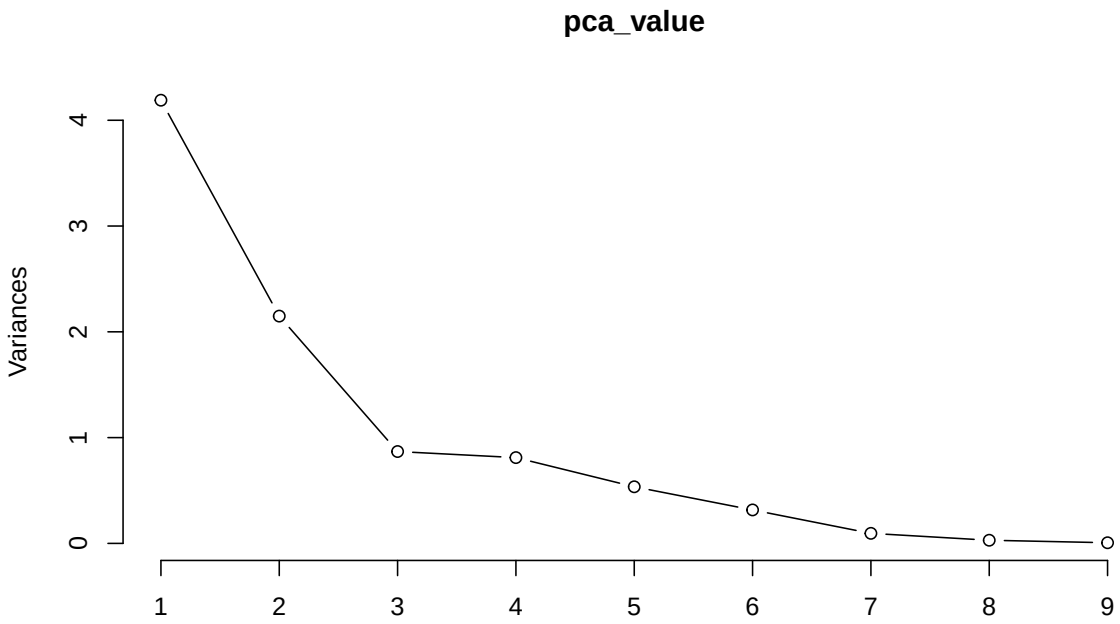

Figure 4: *Screeplot of PCs in PCA for value analysis of stocks*

Table 4: *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

Meanwhile, this report would also need to consider which PCs should be used. Table 4 shows the summary statistics of components. It is clear that PC1 and PC2 have explained almost 86% of the total variation of 4 variables. Besides, figure 5 also suggests that 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 exactly 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, 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 analysing the correlation between each stock, a correlation biplot is required.

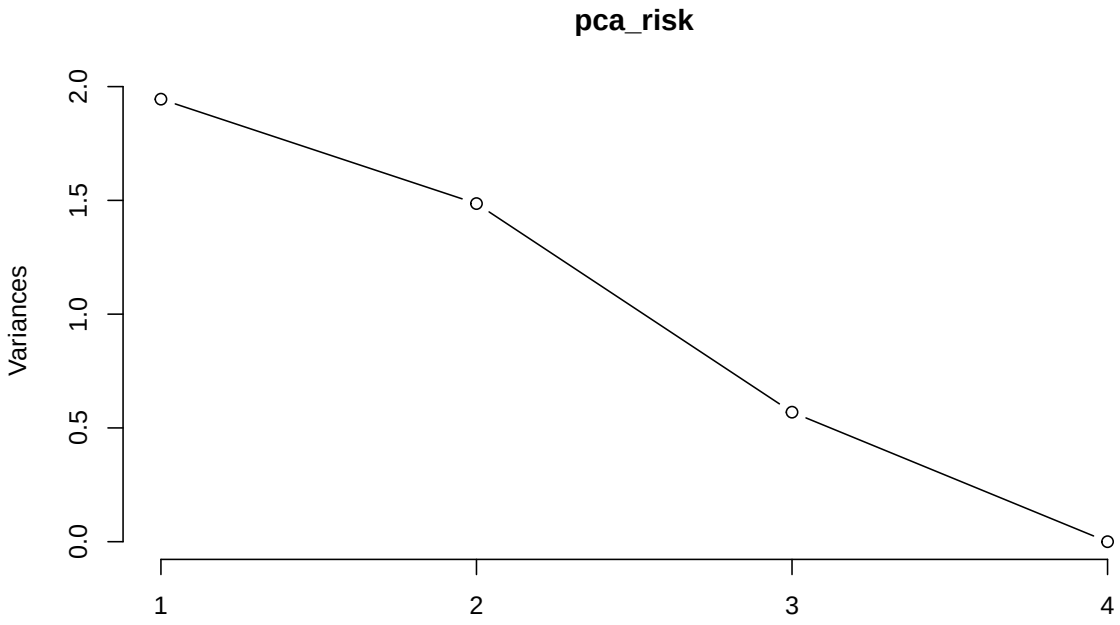

Figure 5: *Screeplot of PCs in PCA for risk analysis of stocks*

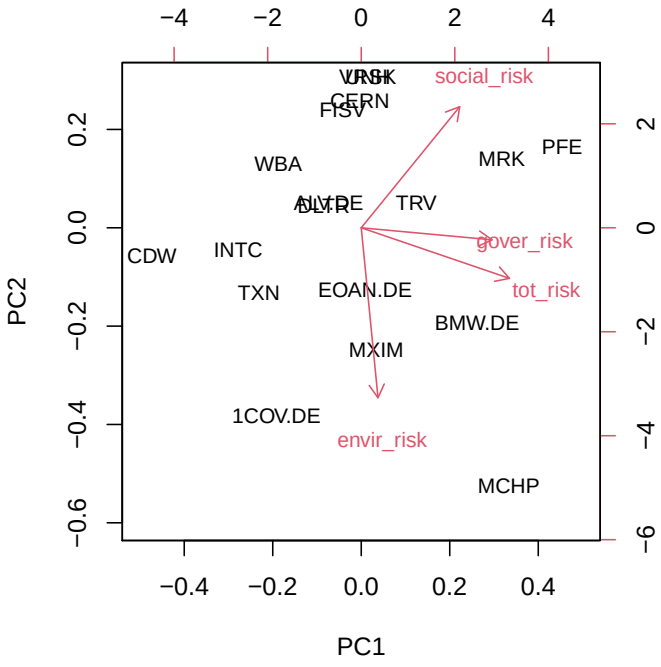

Figure 6: *Distance biplot of PCA of stocks' risk*

Figure 7 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. **VRSK**, **UNH**, **CERN**, and **FISV** with the high values of social risk score, which indicate that these four stocks might perform better than the others when facing the social challenges. While, those stocks might not be good at facing the challenges from environmental risks because the angle between social risk score and environmental risk socre is close to 180 degree, which imply a highly negative correlation approximately. In contrast, **Covestro AG (1COV.DE)** and **MCHP** have the strongest abilities to face the environmental challenges. Meanwhile, **MCHP** also has the highest score of governance risk, which indicates a good performance when meeting the governance challenges. Besides, **PFE**, **Bayerische Motoren Werke AG (BMW.DE)**, and **Merck & Co., Inc. (MRK)** also perform well in governance challenges. While because of the projected positions of these three stocks along the axis of governance risk score are gradually decreasing, the approximate actual values of stocks performance might gradually decline.

In general, based on the total ESG risk score, **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 challenges, and have strong resilience when meeting risks. Therefore, there might not be significant fluctuations of them when facing challenges, and could be stable. In contrast, the stock of **CDW** has a weak overall performance when facing challenges because the approximation actual value in the axis of total risk score is very low. It indicates that the risks might impact on **CDW** easily, and **CDW** might experience a significant fluctuation when facing risks.

3.3 Cluster Analysis

Using the hierarchical clustering analysis with 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, we use the Euclidian distance to find the distance between all pairs of observations. We employ the Ward's methodology to sort the clusters. And the resulting of clusters are shown in the dendogram, which is a tree-like diagram that displays the sequences of merges or splits. Based on the Figure 8, the two and four clusters solutions are not stable. Hence, the three cluster solution is stable which is shown in 9.

From the dendogram, there are three different clusters. Table 5 shows the first cluster of stocks, table 6 shows the second cluster of stocks, and table 7 shows the third cluster of stocks.

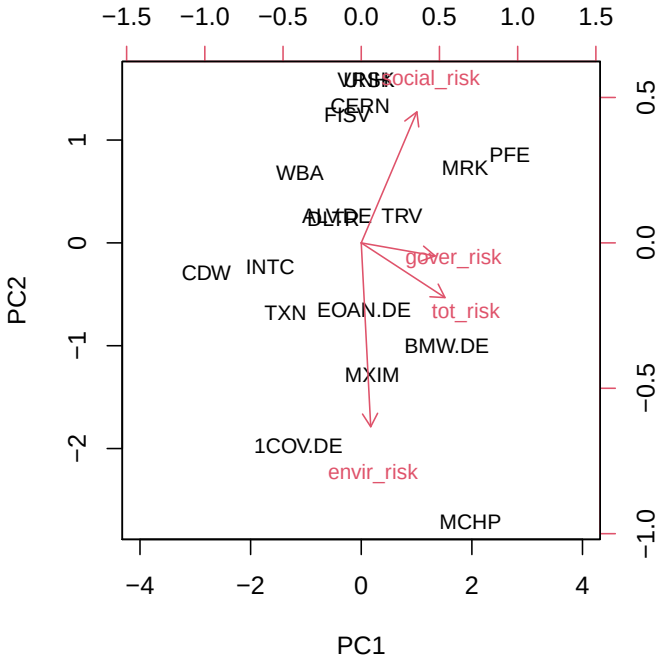

Figure 7: *Correlation biplot of PCA of stocks' risk*

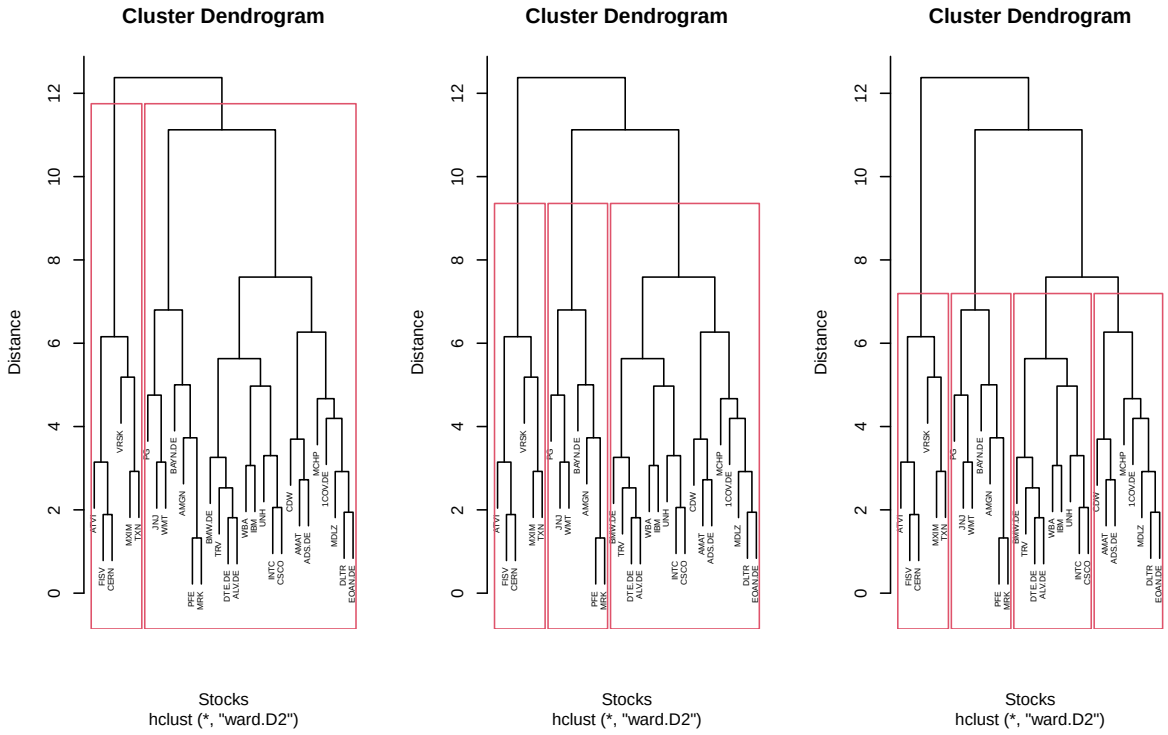

Figure 8: *Choosing clusters*

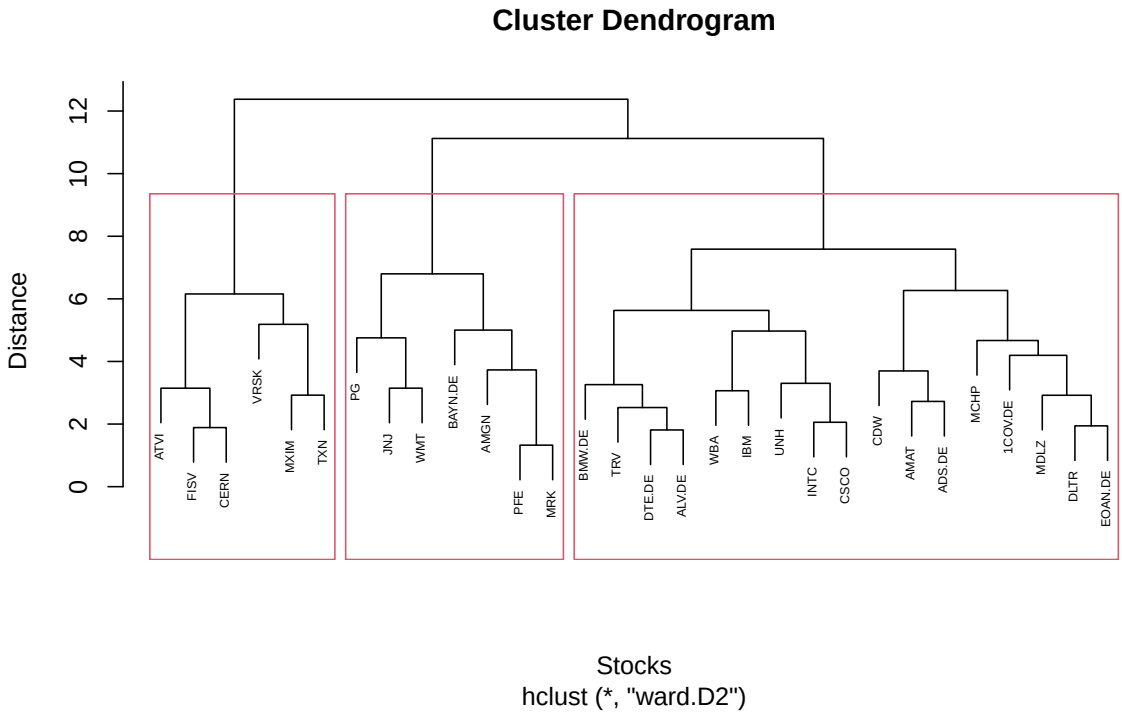

Figure 9: *Dendogram using Ward methodology and taking Euclidian distances*

Table 5: *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 6: *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 7: *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**, **MXIM** are better in price and earning. **VRSK**, **JNJ** and **WMT** could be considered as the special cases since they outperform in their own area. What ESG risk analysis provides us is that **MCHP** and **BMW.DE** have a great performance in overall anti-risk, some other companies are 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 in 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 companies.
- 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

The data could be downloaded from Yahoo Finance. Meanwhile, the report uses the template called **Monash Consulting Report** which could use by downloading the package called MonashEBSTemplates. In addition, the programming language used to analyse the stocks is R (4.0.2) (R Core Team, 2020).

Following packages has been included in our Rmd file:

- package dplyr (1.0.1) (Wickham et al., 2020),
- package ggplot2 (3.3.2) (Wickham, 2016),
- package tidyverse (1.3.0) (Wickham et al., 2019),
- package mclust (5.4.6) (Scrucca et al., 2016),
- package visdat (0.5.3) (Tierney, 2017),
- package gridExtra (2.3) (Auguie, 2017),
- package kableExtra (1.1.0) (Zhu, 2019),
- package tibble (3.0.3) (Müller & Wickham, 2020).

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.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 risks analysis. We still could not deny fact that in general the small sample size would lead to the bias in output. Therefore, our report use the cluster analysis as alternative approach The agglomerative method indicates that stable solution is three cluster. 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 8 indicates that the complete linkage method has a relatively high level of agreement with the Ward's method.

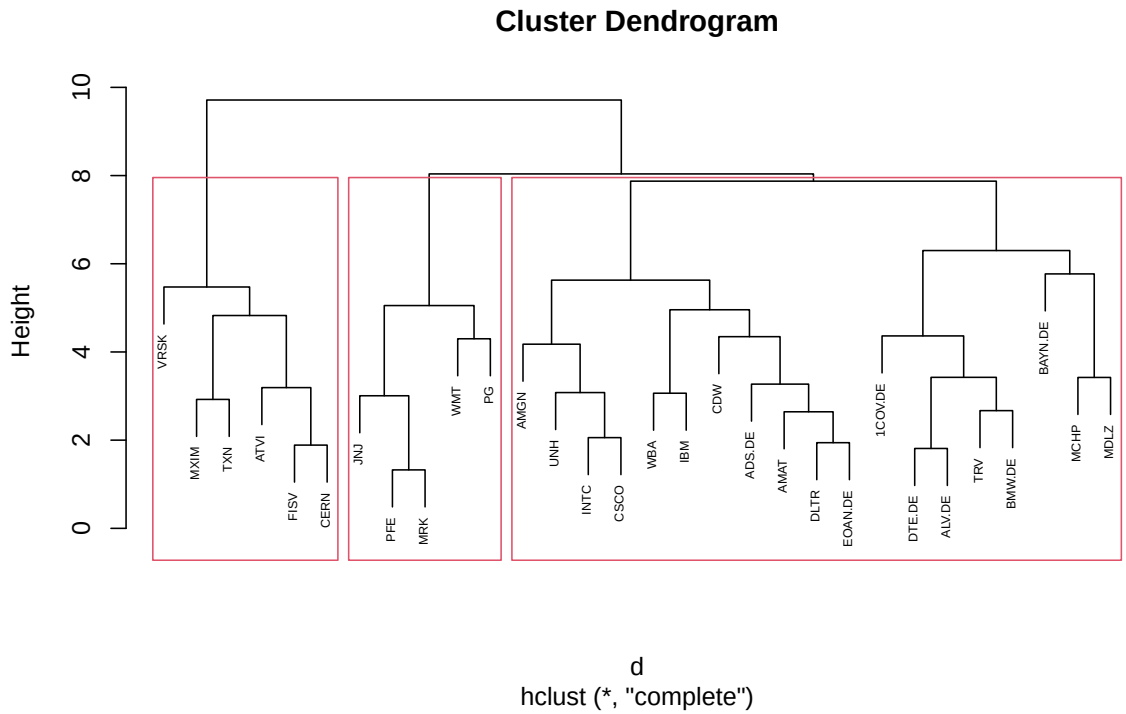

Figure 10: *Cluster dendrogram of complete linkage method*

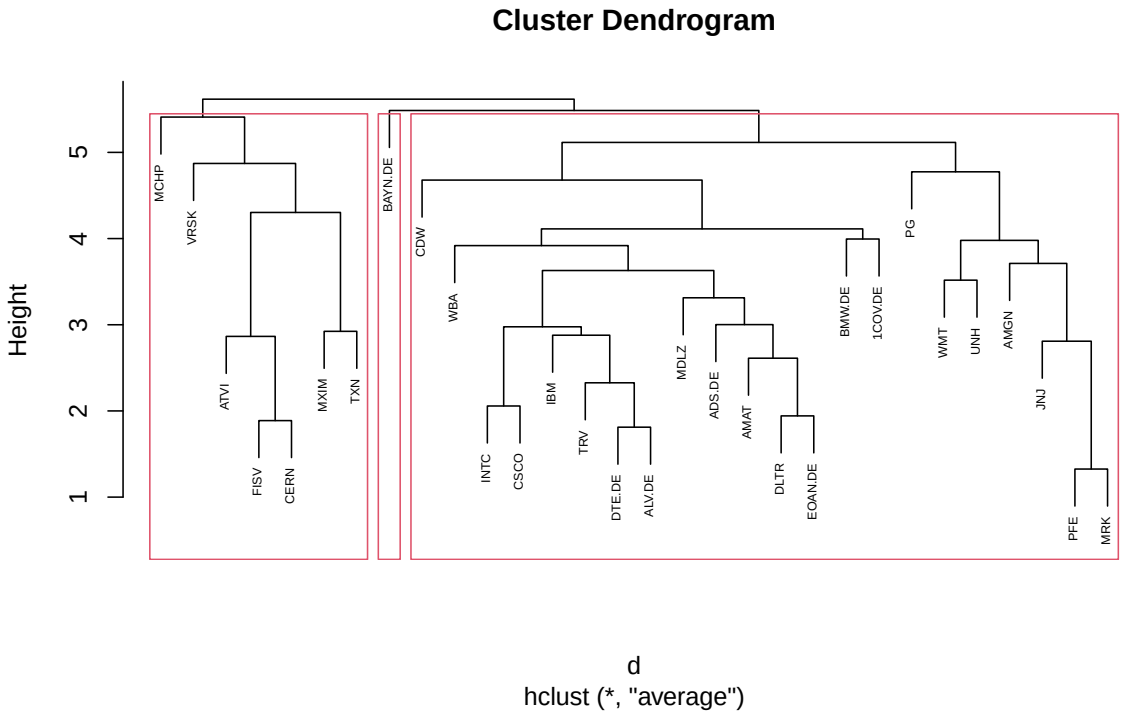

Figure 11: *Cluster dendrogram of average linkage method*

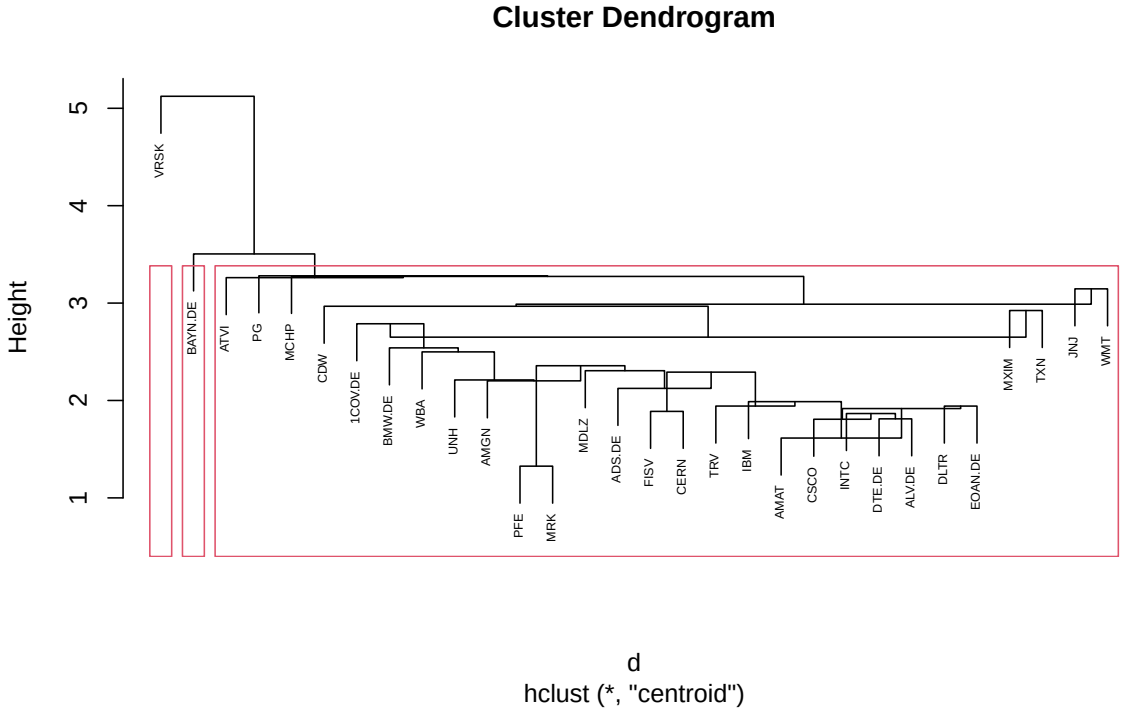

Figure 12: *Cluster dendrogram of centroid method*

Table 8: *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