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

2 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 readers hard to make the comparison. Therefore, this report would apply linear combination (LC) to combine all the variables into a index and utilise principal component analysis (PCA) 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 plays an important role in evaluating the potential values and risks of each stock, which utilises the small number of the PCs to explain the general variation in the original data.

In addition, this report will also consider the PCA method whether is a good measurement in the stock pricing and risk measuring. By investigating the measurement of stock value and stock risk to further find our the potential limitation that PCA might face.

Furthermore, the performance of PCA will also be compared with another method which is Clustering Analysis to check the accuracy of our result. 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

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. 1. This dataset contains a lot of missing value which would cause some bias in our final result. 2. This dataset does not contain enough observations. The insufficient sample space will make our final result become unreliable. In addition, if we further

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

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. 3. 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

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.

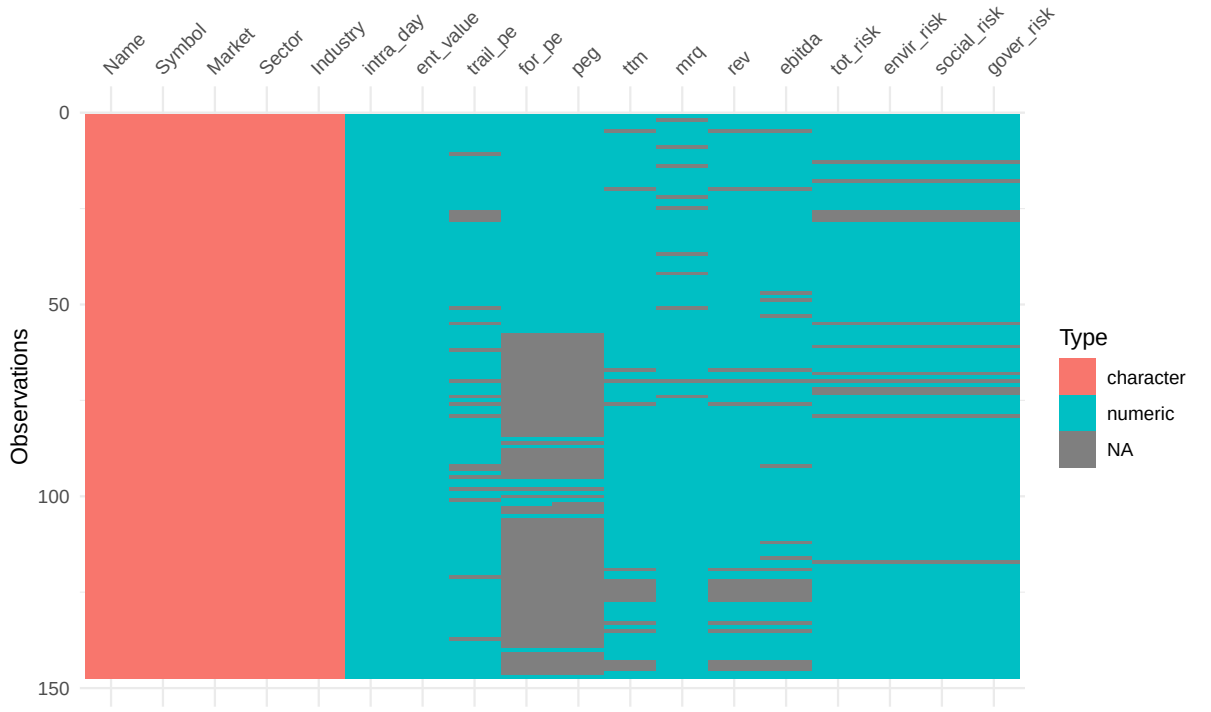

Figure 1: *The data structure of original data*

Meanwhile, by summarising original variables, table 2 indicates that the initial 147 observations have up to 102 missing value.

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

Besides, table 2 shows the information about the outliers. Most of the variables have the really small median and mean, but a extremely high maximum value as well. We could say that the extremely value would definitely dominate our Principle component analysis. The outliers are as following:

- 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 remove, 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.

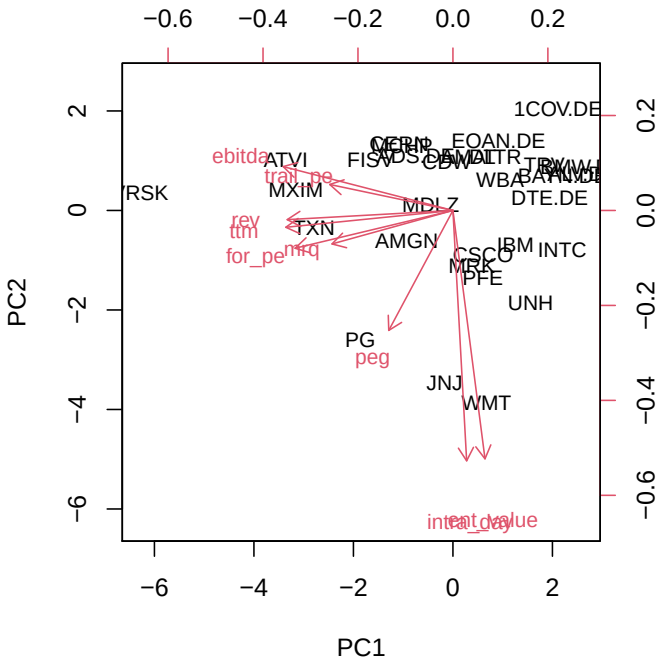

Figure 2: *Correlation Biplot of Stock Value*

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** 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 positive 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.

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 3. What this biplot could tell us is that **Johnson & Johnson (JNJ)** and **Walmart (WMT)** have a pretty high value of

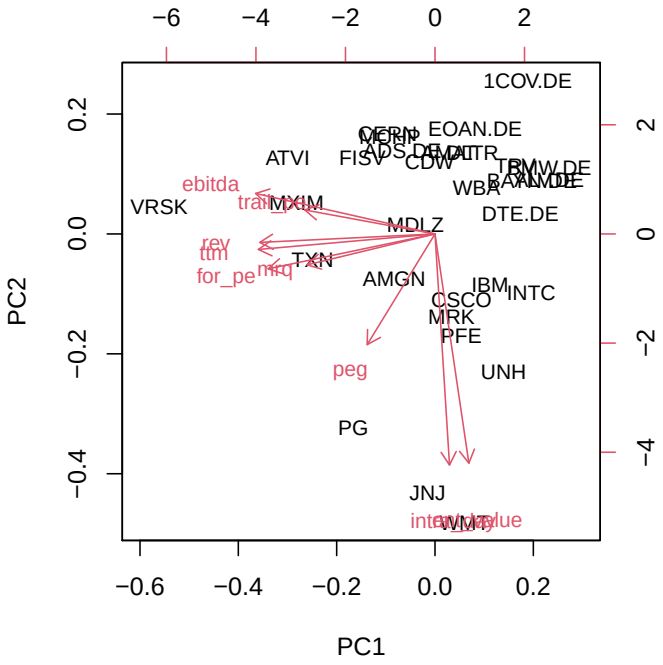

Figure 3: *Distance Biplot of Stock Value*

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

the company, and **Activision Blizzard (ATVI)**, **Texas Instruments Incorporated (TXN)**, **Maxim Integrated Products (MXIM)** indicate the high earnings in the stock.

In the meanwhile, we also need to pay attention on the **Verisk analytics (VRSK)** the potential outlier for the PC1, and **JNJ** and **WMT** 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, comparing with the multinational retail company, the might not have a really large firm size, as for a financial sector, their stock value is positively correlated with the service they provide.

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

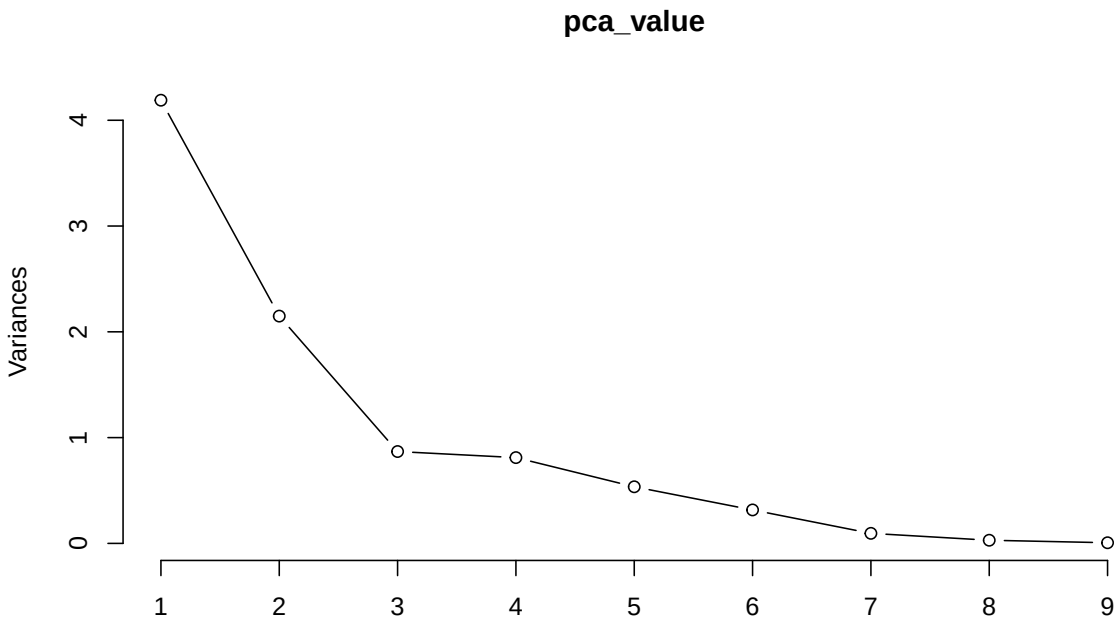

Figure 4: *Screeplot of PCs in PCA for value analysis of stocks*

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 3). 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 (Figure 4) 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. Before executing the risk analysis, this report has compared the total ESG risk score with the sum of the rest three risk scores and filtered out the inconsistent observations in order to improve the accuracy of PCA for risk analysis and avoid errors.

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. In addition, 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.

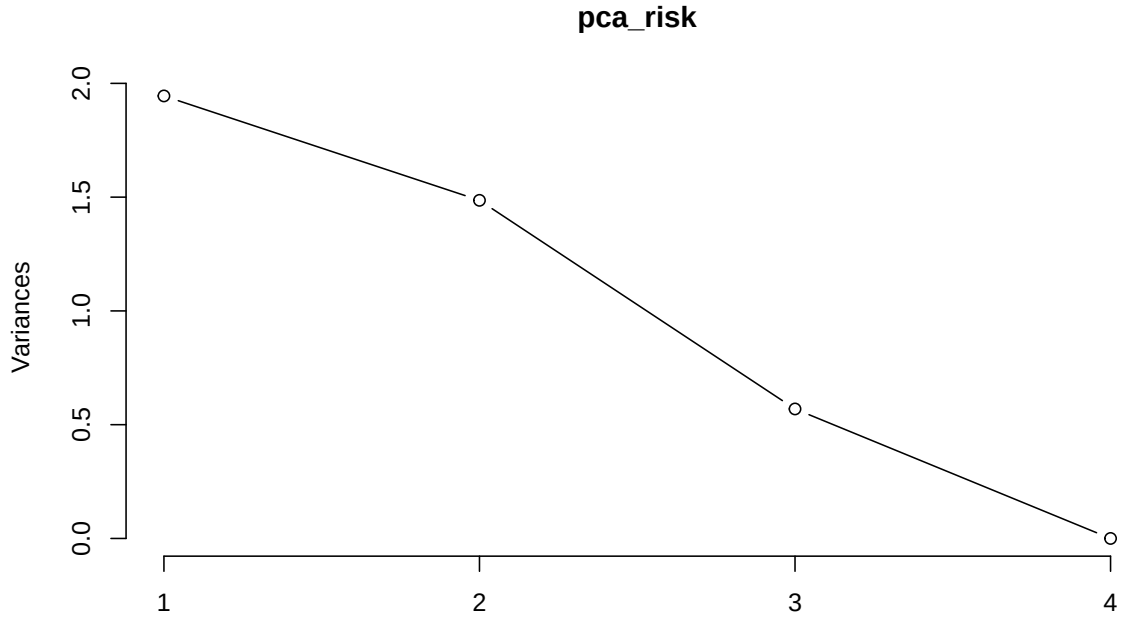

Figure 5: *Screeplot of PCs in PCA for risk analysis of stocks*

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 **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 analyse the correlation between each stock, a correlation biplot is required.

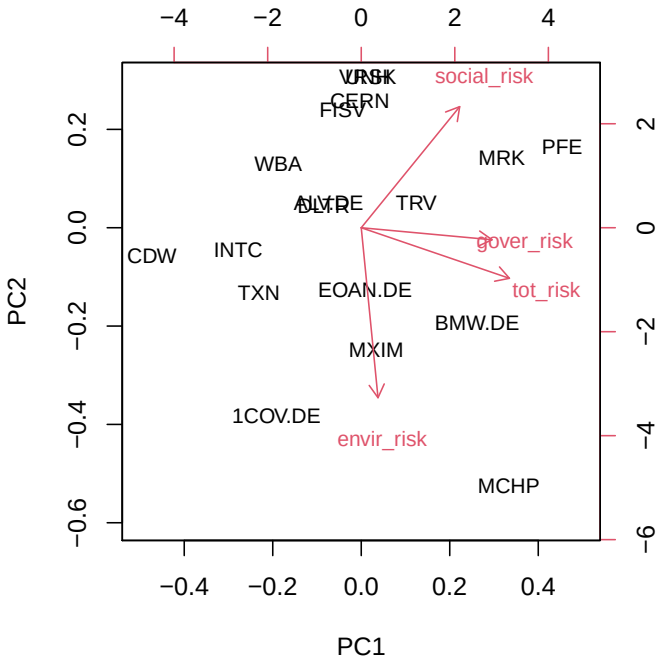

Figure 6: *Distance biplot of PCA of stocks' risk*

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. 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 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 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 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.

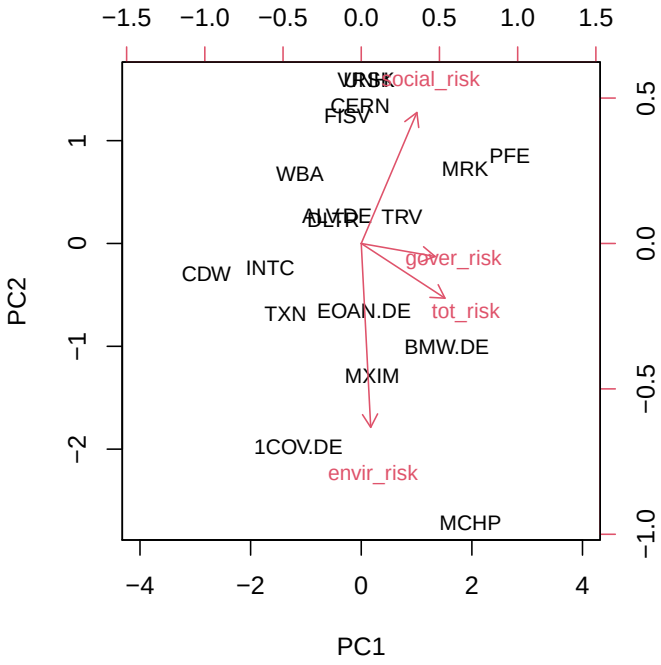

Figure 7: *Correlation biplot of PCA of stocks' risk*

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.

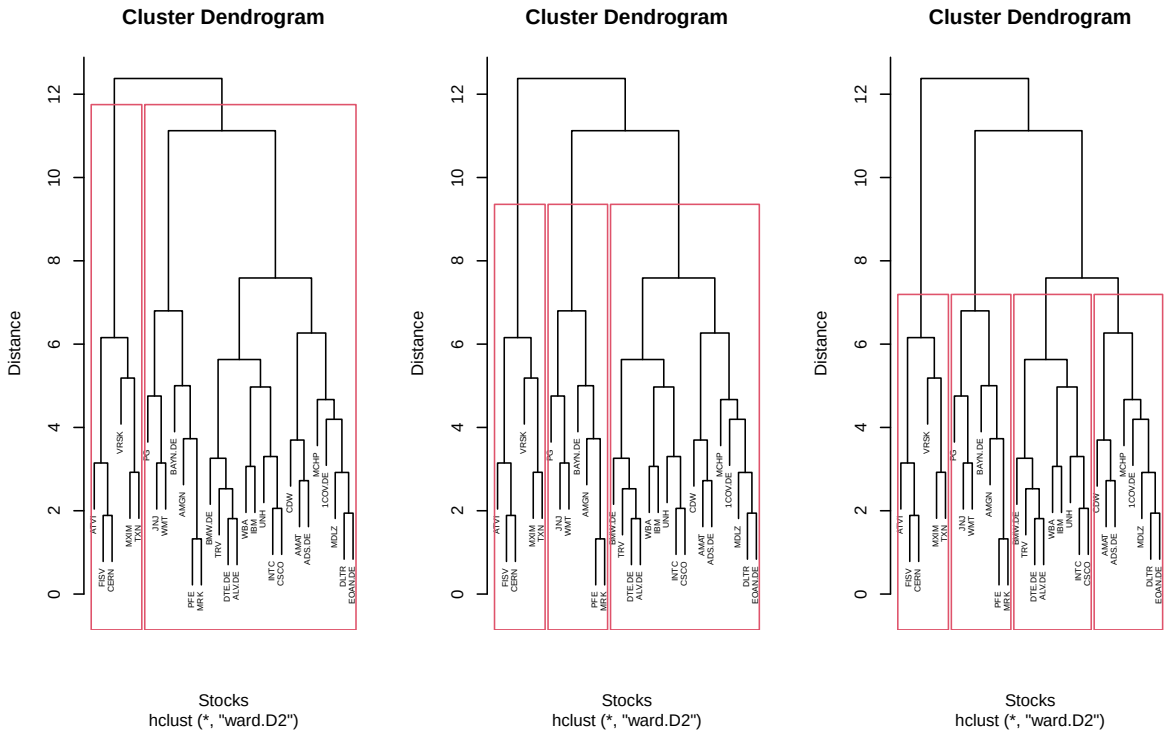

Figure 8: *Choosing clusters*

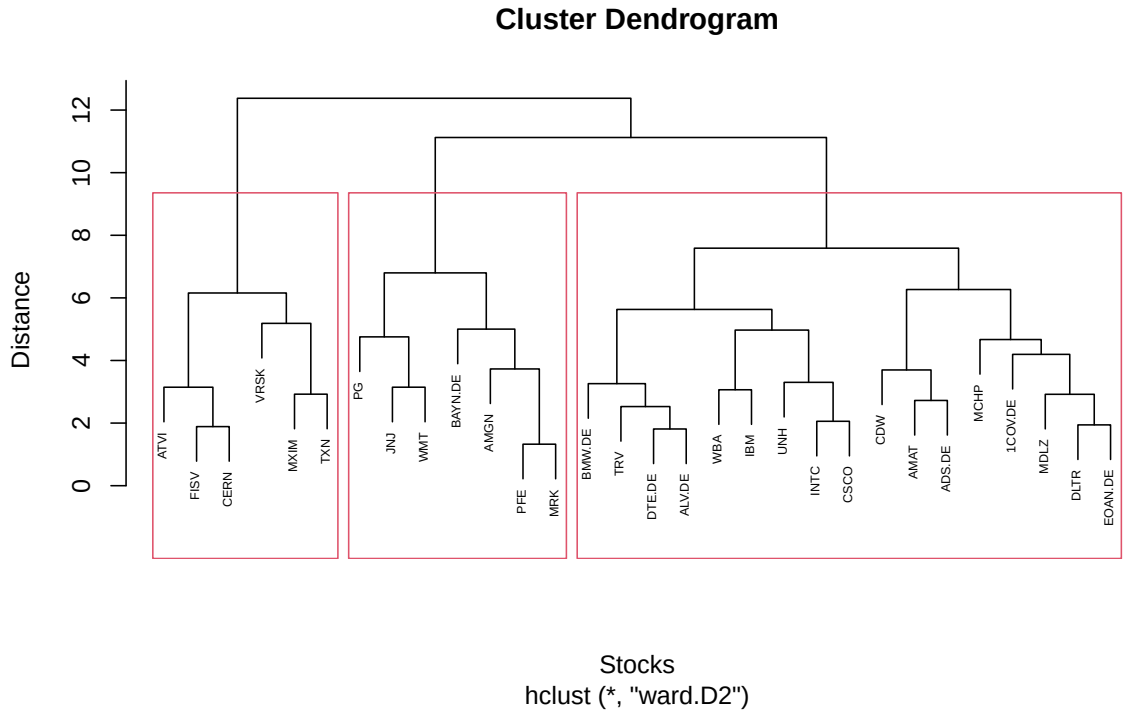

Figure 9: *Dendogram using Ward methodology and taking Euclidian distances*

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

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

We also using other methods to do the cluster analysis, such as 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.

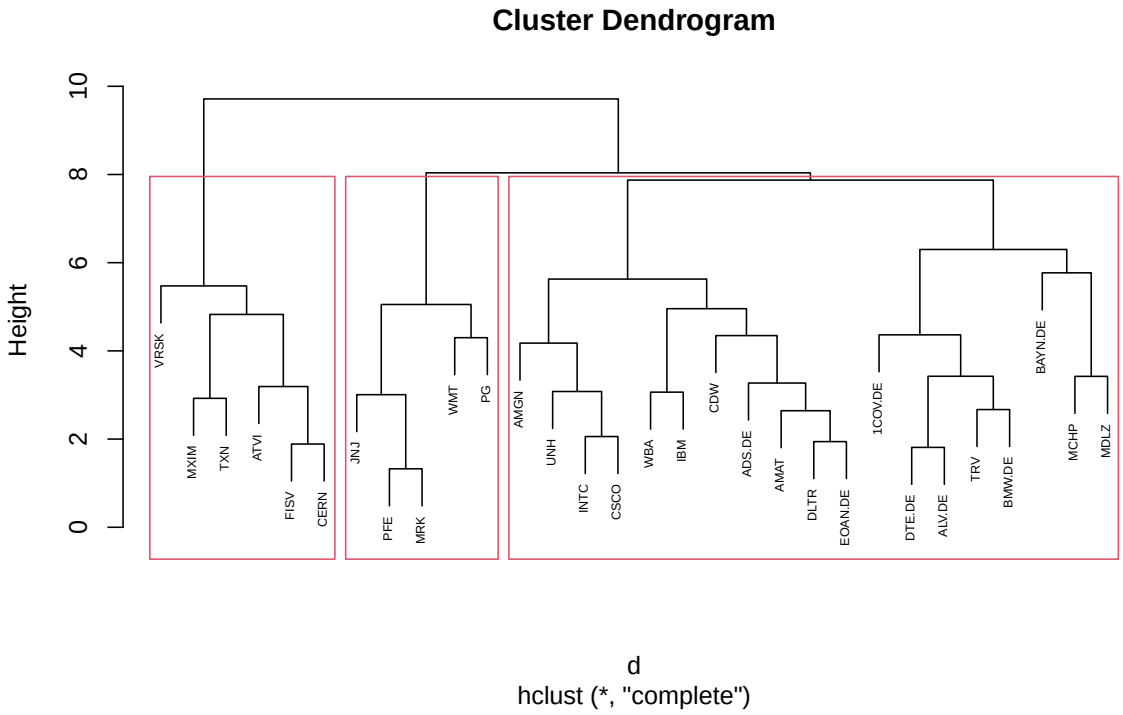

Figure 10: *Cluster dendrogram of complete linkage method*

4 Conclusions

In conclusion, **JNJ** and ** WMT** have high values of the company, while ** ATVI, TXN, MXIM** show the high earnings in the stock market. Besides, **VRSK** shows a high value in both price and earning per share, which might due to the high performance of services gives investors a positive market prospect. Meanwhile, **VRSK** also has high score of social risk which might also because the characteristics of service industry gives the company high performance to meet the social risk. Furthermore, because **JNJ** and **WMT** are the multinational companies and the main profit would

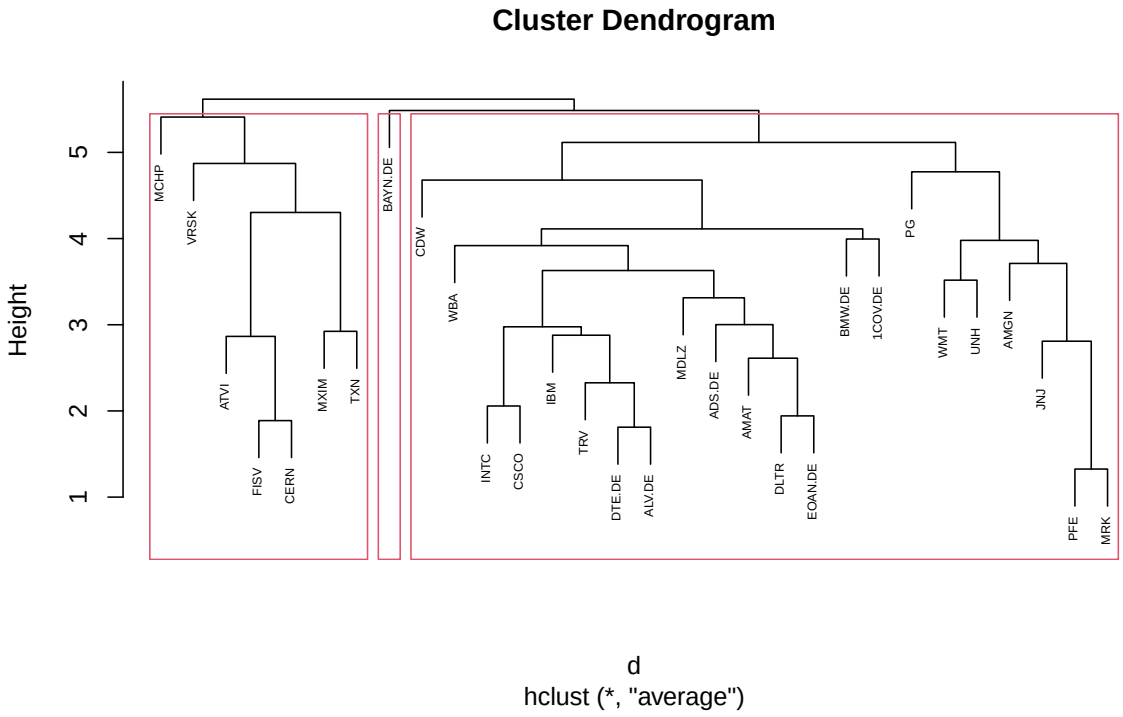

Figure 11: *Cluster dendrogram of average linkage method*

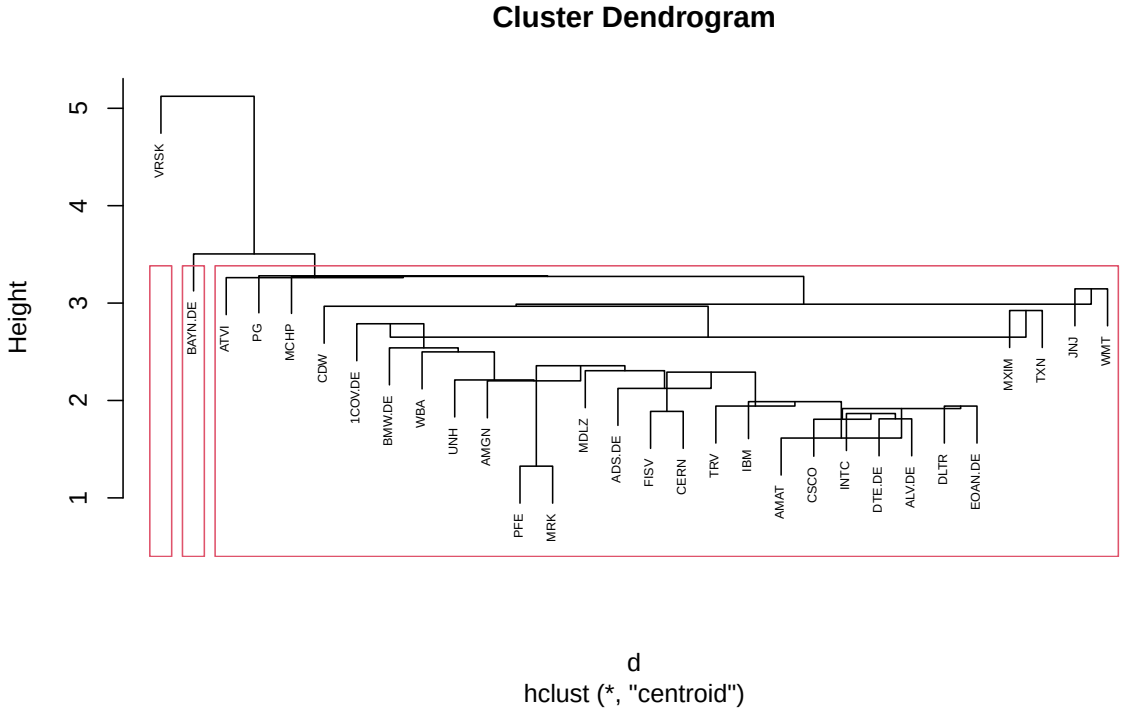

Figure 12: *Cluster dendrogram of centroid method*

come from products selling, they should continuously increase their market share to keep the market profit.

On the other hand, the risk score of each stock indicates the performance of meeting the challenges from risks. **MCHP** and **BMW.DE** show the best overall performance of anti-risks. **UNH**, **CERN**, and **FISV** are the another three companies have the high performance in social risk, while they are not resilient enough when meeting the challenges from environmental risk compared with **1COV.DE** and **MCHP**. Besides, **PFE**, **BMW.DE** and **MRK** show the high performance in meeting the governance challenges. Overall, the internal and external risks would all impact on the value of stocks which generate the fluctuation of prices in the stock market and might influence the confidence of investors. Companies need to improve the ability of self-resilience and anti-risks so than enhance the performance when facing different types of risks.

However, because of the small sample space and the incomplete eigenvalue selection, the sample might not be representative and the biplot might be incomplete based on the Screeplot in the values analysis of stocks. Even though the biplot is suitable because the Screeplot only indicates two useful PCs in risks analysis. The sample size, however, is still too small to provide sufficient and accurate results. Therefore, the report uses another method to analyse the stocks which is cluster analysis. According to agglomerative method, we found that the stable solution is three cluster. Besides, we also found that the complete linkage method is relatively highly agree with the Ward's method.

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 and Wickham 2020).

References

Auguie, B (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3. https://CRAN.R-project.org/package=gridExtra.

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

Wickham, H, R François, L Henry, and K Müller (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.