MONASH University

MONASH BUSINESS SCHOOL

ASSIGNMENT COVER SHEET

Students' name	(Name) **Aarathy Babu, Dewi Amaliah, Priya Dingorkar, Rahul Bharadwaj**		
ID number	**31230008, 31251587, 31292917, 31322239**	**Phone**	
Unit name	**High Dimensional Data Analysis**	**Unit code**	**ETF5500**

Title of assignment	**High Dimensional Data Analysis Assignment 2 – Group Assignment**		
Lecturer/tutor	**Ruben Loaiza-Maya**		
Is this an authorised group assignment? ✓ **Yes** ☐ **No** If this submission is a group assignment, each student must attach their own signed cover sheet to the assignment.			
Has any part of this assignment been previously submitted as part of another unit/course? ☐ **Yes** ✓ **No**			
Tutorial/laboratory day & time			
Due date **20/09/2021**		**Date submitted** **20/09/2021**	

All work must be submitted by the due date. If an extension of time to submit work is required, a Special Consideration Application (In-semester Assessment Task) must be submitted.

Has an extension been approved? **Yes** ☐ **No** ☐ **If yes, please give the new submission date ……./……./…….**

Please note that it is your responsibility to retain copies of your assessments.

Intentional plagiarism or collusion amounts to cheating under Part 7 of the Monash University (Council) Regulations.

Plagiarism: means taking and using another person's ideas or manner of expressing them and passing them off as one's own. For example, by failing to give appropriate acknowledgement. The material used can be from any source (staff, students or the internet, published or unpublished works).

Collusion: means unauthorised collaboration with another person on assessable written, oral or practical work and includes paying another person to complete all or part of the work.

Where there are reasonable grounds for believing that intentional plagiarism or collusion has occurred, this will be reported to the Chief Examiner, who may disallow the work concerned by prohibiting assessment or refer the matter to the Associate Dean Teaching and Learning.

Student Statement:

- I have read the University's Student Academic Integrity Policy and the University's Student Academic Integrity: Managing Plagiarism and Collusion Procedures.
- I understand the consequences of engaging in plagiarism and collusion as described in Part 7 of the Monash University (Council) Regulations.
- I have taken proper care of safeguarding this work and made all reasonable effort to ensure it could not be copied.
- I acknowledge that the assessor of this assignment may for the purposes of assessment, reproduce the assignment and:
  - i. provide to another member of faculty; and/or
  - ii. submit it to a plagiarism checking service; and/or
  - iii. submit it to a plagiarism checking service which may then retain a copy of the assignment on its database for the purpose of future plagiarism checking.
- I certify that I have not plagiarised the work of others or participated in unauthorised collaboration when preparing this assignment.

Signature **Aarathy Babu, Dewi Amaliah, Priya Dingorkar, Rahul Bharadwaj** *Date* **20/09/2021**

Privacy Statement

The information on this form is collected for the primary purpose of assessing your assignment. Other purposes of collection include recording your plagiarism and collusion declaration, attending to course and administrative matters and statistical analyses. If you choose not to complete all the questions on this form it may not be possible for Monash University to assess your assignment. You have a right to access personal information that Monash University holds about you, subject to any exceptions in relevant legislation. If you wish to seek access to your personal information or inquire about the handling of your personal information, please contact the University Privacy Officer: privacyofficer@adm.monash.edu.au

Assignment 2 Report

Aarathy Babu, Dewi Amaliah, Priya Dingorkar, Rahul Bharadwaj

1 Data Description

In this study, we will use a variety of high-dimensional analysis approaches like Multidimensional Scaling, Principle Component Analysis, and Clustering to extract useful insights from the data. This report is based on data from US businesses that declared bankruptcy between 1980 and 2000. The information is coming from the UCLA-LoPucki Brankruptcy Research Database. This datasets has 21 variables and 436 firms with their description explained below.

- Name: Name of the firm
- Assets: Total assets (in millions of dollars)
- CityFiled: City where filing took place
- CPI U.S CPI at the time of filing
- DaysIn: Length of bankruptcy process
- DENYOther: CityFiled, categorized as Wilmington (DE), New York (NY) or all other cities (OT)
- Ebit: Earnings (operating income) at time of filing (in millions of dollars)
- Employees: Number of employees before bankruptcy
- EmplUnion: Number of union employees before bankruptcy
- FilingRate: Total number of other bankrupcy filings in the year of this filing
- FirmEnd: Short description of the event that ended the firm's existence
- GDP: Gross Domestic Product for the Quarter in which the case was filed
- HeadCityPop: The population of the firms headquarters city
- HeadCourtCityToDE: The distance in miles from the firms headquarters city to the city in which the case was filed
- HeadStAtFiling: The state in which firms headquarters is located
- Liab: Total amount of money owed (in millions of dollars)
- MonthFiled: Categorical variable where numbers from 1 to 12 correspond to months from Jan to Dec
- PrimeFiling: Prime rate of interest on the bankruptcy filing date
- Sales: Sales before bankruptcy (in dollars)
- SICMajGroup: Standard industrial clasification code
- YearFiled: Year bankruptcy was filed

2 Preliminary Data Analysis

Before carrying out further analysis of the data, let us conduct some preliminary data analysis. Throughout our strategy, we have tried to retain the data as much as possible while maintaining high data quality and credibility.

From the summary shown below, we have 6 character variables and 15 numeric variables.

```
## Rows: 436
## Columns: 21
## $ Name              <chr> "Combustion Equipment Associates, Inc.", "Penn-Dixie~
## $ Assets            <int> 531, 552, 1897, 821, 4097, 1200, 1141, 2628, 1456, 1~
## $ CityFiled         <chr> "New York", "New York", "Cleveland", "New York", "Sa~
## $ CPI               <dbl> 84.8, 81.0, 84.0, 93.2, 87.0, 94.0, 93.7, 87.9, 94.9~
```

```
## $ DaysIn            <int> 1157, 696, 1170, 1545, 792, 1099, 1343, 2238, 881, 1~
## $ DENYOther         <chr> "NY", "NY", "OT", "NY", "OT", "OT", "OT", "NY", "OT"~
## $ Ebit              <dbl> 13.831140, -13.521542, 102.647226, 71.496993, 176.43~
## $ Employees         <int> 2400, 4191, 9685, 1116, 1400, 5225, 32000, 1900, 172~
## $ EmplUnion         <int> NA, 2975, 5800, NA, NA, NA, NA, NA, NA, NA, 8531, NA~
## $ FilingRate        <int> 3, 3, 3, 5, 5, 5, 5, 5, 13, 13, 13, 13, 13, 13, 13, ~
## $ FirmEnd           <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", ~
## $ GDP               <dbl> 42.067, 41.346, 41.296, 43.083, 42.891, 42.613, 42.6~
## $ HeadCityPop       <dbl> 7071639, 7071639, 58056, 418532, 683472, 1185802, 11~
## $ HeadCourtCityToDE <int> 106, 106, 435, 241, 2514, 435, 2383, 106, 653, 1238,~
## $ HeadStAtFiling    <chr> "NY", "NY", "MI", "PA", "CA", "MI", "CA", "NY", "IL"~
## $ Liab              <dbl> 309.6648, 377.9007, 1201.9985, 751.4130, 4872.2510, ~
## $ MonthFiled        <int> 10, 4, 9, 9, 1, 12, 11, 2, 4, 12, 5, 6, 11, 8, 2, 8,~
## $ PrimeFiling       <dbl> 14.00, 20.00, 11.50, 19.50, 20.00, 15.75, 16.00, 19.~
## $ Sales             <dbl> 357537044, 900139474, 3662349812, 423926972, 6016918~
## $ SICMajGroup       <chr> "38 Measuring, Analyzing and Controlling Instruments~
## $ YearFiled         <int> 1980, 1980, 1980, 1981, 1981, 1981, 1981, 1981, 1982~
```

It can be observed that there are quite a number of empty values present in `FirmEnd` which are essentially `NULL` values. Therefore, we have converted these into`NA` values.

The data credibility issues are checked by confirming if the `DaysIn`, `EmplUnion`, `Employees`, `HeadCourtCityToDE`,`MonthFiled`, `YearFiled` and `HeadCityPop` are non-negative values. There are observations where the `EmplUnion` values are more than `Employees` which was removed from the data and that certain companies have 1 `Employees` and 1 `EmplUnion` values as shown below, which is suspicious but since there is not any concrete evidence that these observations pose data credibility issues, these observations were not excluded for the analysis.

Table 1: Credibility issues in Employees and EmplUnion

Name	Employees	EmplUnion
Residential Resources Mortgage Investments Corp.	1	1
Mortgage & Realty Trust (1990)	1	1
Promus Companies Inc. (Harrahs Jazz Co. only)	1	3000

We have separated `SICMajGroup` into a new factor variable `SIC` and its meaning in the `SICMajGroup` so as to make it more identifiable without the lengthy name.

The missing values in the data has been visualized as shown in figure 1.

It can be observed that `FirmEnd` has the highest number of missing values, followed by `EmplUnion`. The strategy employed is to remove the variables `FirmEnd` and `EmplUnion`. As the variable `FirmEnd` depicts the description of the end of Firm's existence. It doesn't provide significant value to the analysis and it can be excluded. Similarly `EmplUnion` is removed due to the fact that `Employees` and `EmplUnion` are closely related and `EmplUnion` is be a subset of `Employees`. Therefore, removing `EmplUnion` which has too many missing values would not affect our analysis significantly.

The missing values of `DaysIn` in 4 companies were encoded based on the publicly available data and imputation. Values were encoded for **AP Industries** and **Daisy Systems Corp.** (See Appendix for more information). However, the data for **Hunt International Resources Corp.** and **McCrory Corp.** was not available, therefore we have imputed the variable, based on median of the `DaysIn` in the industry classification they belong to.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    31.0   248.0   509.0   635.5   867.5  3730.0
```

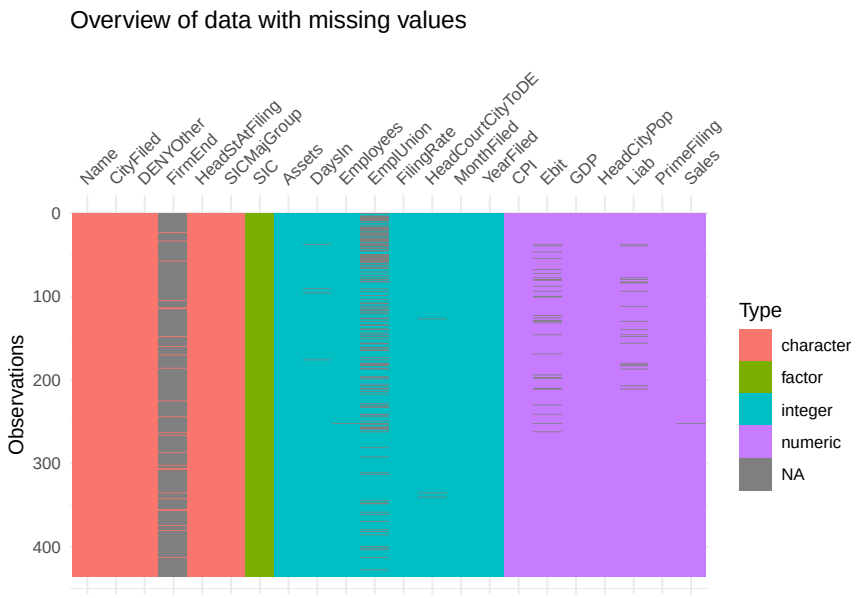

Figure 1: Overview of missing values in the data

Table 2: Firms with missing values in DaysIn

Name	DaysIn
Hunt International Resources Corp.	NA
AP Industries, Inc.	NA
Daisy Systems Corp.	NA
McCrory Corp.	NA

The summary statistics of the variable after imputation, suggests no suspicious outliers or anomalies as the bankruptcy can be a lengthy ordeal.

The missing values in `HeadCourtCityToDE` shown in the table below are imputed using the values in `CityFiled`, `DENYOther`, and `HeadStAtFiling`. Considering the publicly available data on headquarter address and the `CityFiled`, the distances between these cities were found and imputed into the data accordingly (see Appendix).

Table 3: Missing values in 'HeadCourtCityToDE'

Name	HeadCourtCityToDE	CityFiled	DENYOther	HeadStAtFiling
Divi Hotels, N.V.	NA	Miami	OT	Aruba
Loewen Group, Inc.	NA	Wilmington	DE	Canada
Philip Services Corp. (1999)	NA	Wilmington	DE	Canada

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   248.0   707.0   925.5  1318.0  2942.0
```

Exploring the summary statistics, it was observed that the minimum distance is 1, when inspected the state of headquarters and the city filed is in the same state. Therefore it does not pose a data credibility issue.

With regards to the `Employees` variable, it was observed that there is a single observation that is missing data on its employees. Under closer examination of the `Sales` Variable, we observed that it was the same firm that had missing data on `Sales` as well. On closer inspection of this firm, the presence of missing values

Table 4: Firms with headquarters and the city filed in the same state

Name	HeadCourtCityToDE	CityFiled	DENYOther	HeadStAtFiling
Phoenix Steel Corp.	1	Wilmington	DE	DE
Columbia Gas System Inc.	1	Wilmington	DE	DE

on the variable `Ebit` was also found, therefore we remove this observation considering the fact that this single observation has missing values of these three variables.

Table 5: Firms with missing values in Sales, Employees, and Ebit

Name	Sales	Employees	Ebit
County Seat, Inc.	NA	NA	NA

The missing values in `Liab` and `Ebit` was treated by dropping the missing observations, as the missing values in each of the variables were below 10% and out of the 39 rows where either one of the two variables were missing, 8 of the observations have missing values on both `Liab` and `Ebit`. We believe it is more reasonable to drop the missing values than impute them as imputation could mislead the analysis.

`DENYOther`, `MonthFiled` and `YearFiled` ought to be factor as mentioned in the data description therefore are converted to factor from numeric variables as shown in figure 2

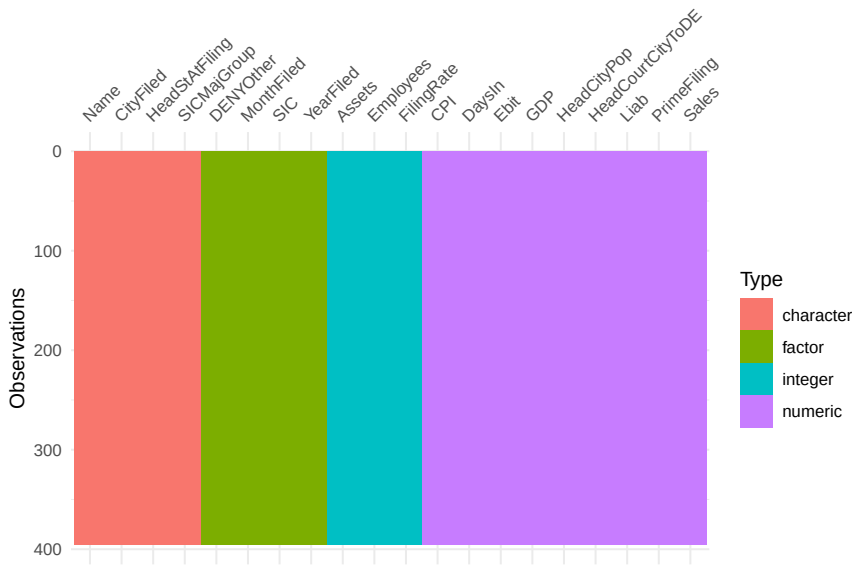

Figure 2: Overview of Cleaned Data

The data was then checked for outliers, even though we haven't found suspicious outliers in majority of the variables (see Appendix), outliers were found in `Ebit`, `Liab` , `Assets` and `Sales` as shown below in figure 3 and 4. Interestingly, these values belong to a single firm called **Texaco Inc.**. This will be discussed further in the sections below.

In order to gain insights from the data, we have further explored it. Below shown is a correlation plot. It is clear from the plot that `HeadCityPop`and `HeadCourtCityToDE` have no correlation with any of the other variables. Therefore, we omit these two variables from further analysis.

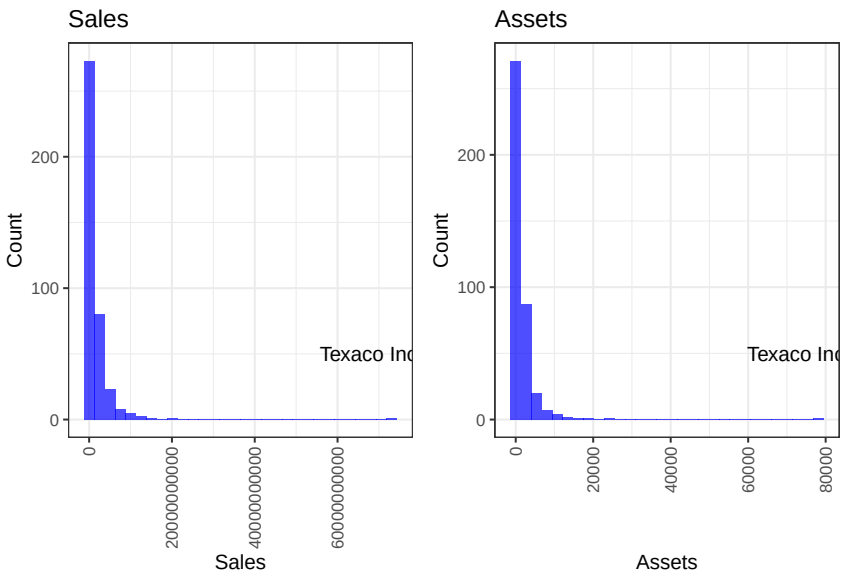

Figure 3: Presence of Outliers in Sales and Assests

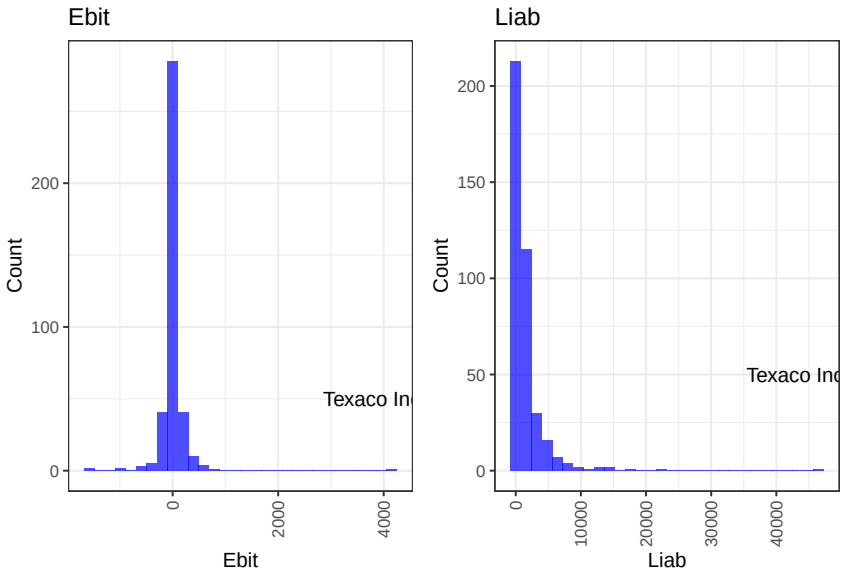

Figure 4: Presence of Outliers in Ebit and Liab

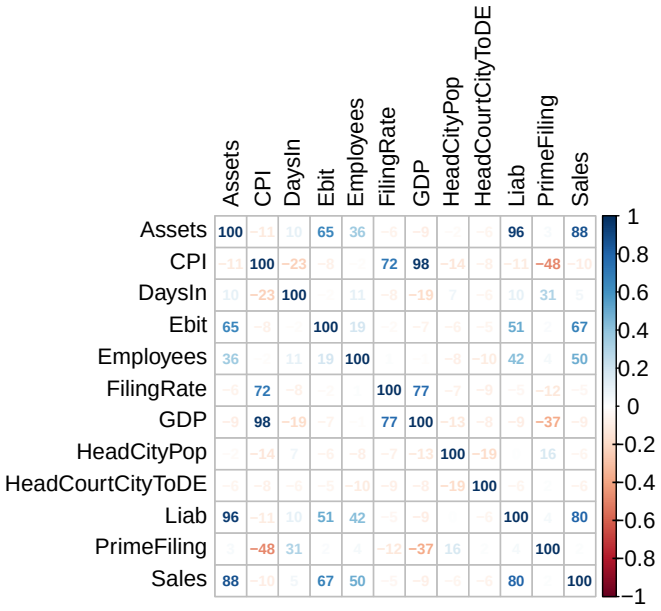

Figure 5: Correlation plot of numeric variables in the cleaned data

3 Multidimensional Scaling (MDS)

MDS is a statistical method to represent multidimensional data into lower-dimensional (2D) data. Thus, MDS is relevant to represent bankruptcy data in two-dimensional visualisation. This method uses distance to do the job. Hence, we limit the MDS only to incorporate numerical variables so that we can use Euclidean distance or is known as classical MDS. We will also only incorporate numerical variables directly related to bankruptcy. Those variables are: `Assest`, `DaysIn`, `Employees`, `CPI`, `Ebit`, `Liab`, `FillingRate`, `GDP`, `PrimeFilling`, and `Sales`. These variables has different unit of measurements, hence we standardise it.

3.1 Classical MDS

Figure 6 conveys that Texaco Inc (Tin.), Baldwin-United Corporation (B-UC), Federated Department Stores, Inc. (FDSI), LTV Corp. (1986) (LTCV.(1) are potential outliers. On closer inspection of the data, we find that these firms have the largest assets. Moreover, Texaco Inc. also has high operating income, sales, and liability.

As mentioned previously, the aim of MDS is to visualise the firms in 2D scatter plot. However, this objective will be less clearly achieved in Figure 6 since too many observations overlapped each other. Hence, we decide to exclude Texaco Inc. and re-conduct classical MDS. This gives us a clearer visualisation as follows:

Figure 7 suggests that the visual representation of the rest firms other than Texaco, Inc. remains the same. B-UC, LTCV.(1, and FDSI are still far apart from other firms. It implies that our MDS is pretty robust. However, since it gives a clearer visualisation, we will use the data without Texaco, Inc. in the rest of MDS analysis. It also implies that most firms that filed for bankruptcy have similar characteristics since they tend to be plotted near or even overlapped with each other. We can also see that some firms are spread out. It means that these firms have different profile.

3.2 Goodness of Fit

In this part, we inspect the MDS's Goodness of Fit. If two GoF values are equal, which is the ideal condition if we use Eucledean distance, then we can conclude that the strain is minimised and the solution is optimal. Here is the GoFs of the MDS:

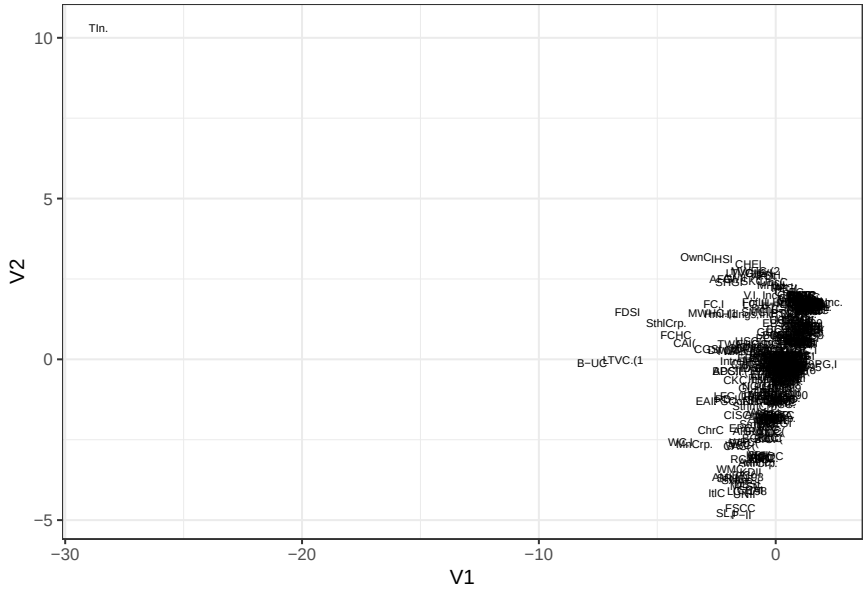

Figure 6: Classical MDS solution for bankruptcy data. The x and y-axis represent the new variables as the result of MDS. Some outliers observed in thedata

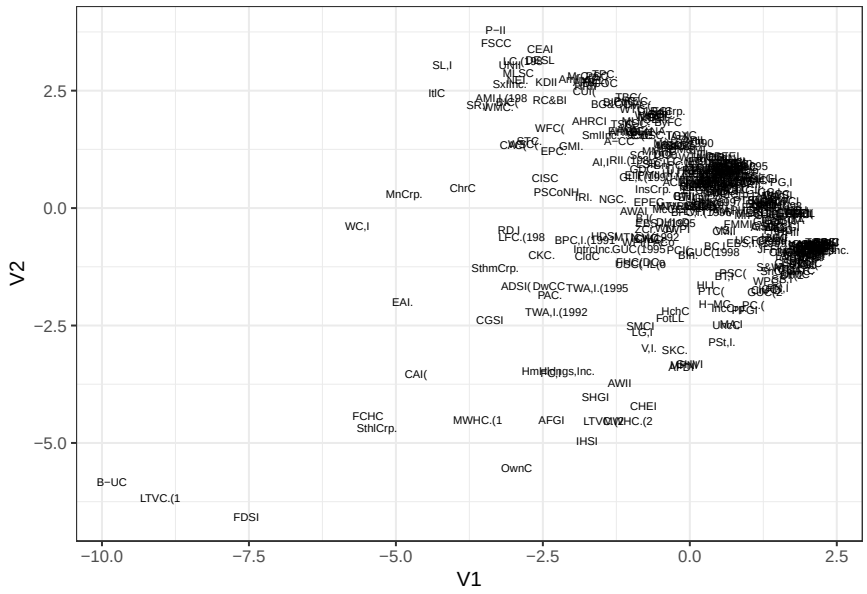

Figure 7: Classical MDS solution for bankruptcy data after excluding Texaco Inc. The x and y-axis represent the new variables as the result of MDS. We get a clearer visualisation compared to the previous MDS result

```
##  num [1:2] 0.579 0.579
```

We find that the GoF_1 and GoF_2 are equal. Hence, our MDS is optimal. We also find that all the eigenvalues are positive (see Appendix).

3.3 Comparison with non-Classical MDS

Next, we compare the classical MDS with non-classical MDS (Sammon mapping). The stress function could be used to indicate the accuracy of representation. The lower, the better the accuracy.

```
## Initial stress        : 0.12510
## stress after   2 iters: 0.12082
```

```
## [1] 0.1208242
```

We find that the stress is relatively low (0.121), thus non-classical MDS also produce fairly accurate representation of the bankruptcy data. Moreover, the plot (see Appendix) also produce relatively similar result when compared with the classical MDS. Hence, we can conclude that the result is fairly robust with the change of methodology.

3.4 Visualisation with Categorical Variable

This section will show the MDS solution by also take the categorical variables into account. Too keep the report concise, we displaye some categorical features in the Appendix and only display interesting finding in this subsection.

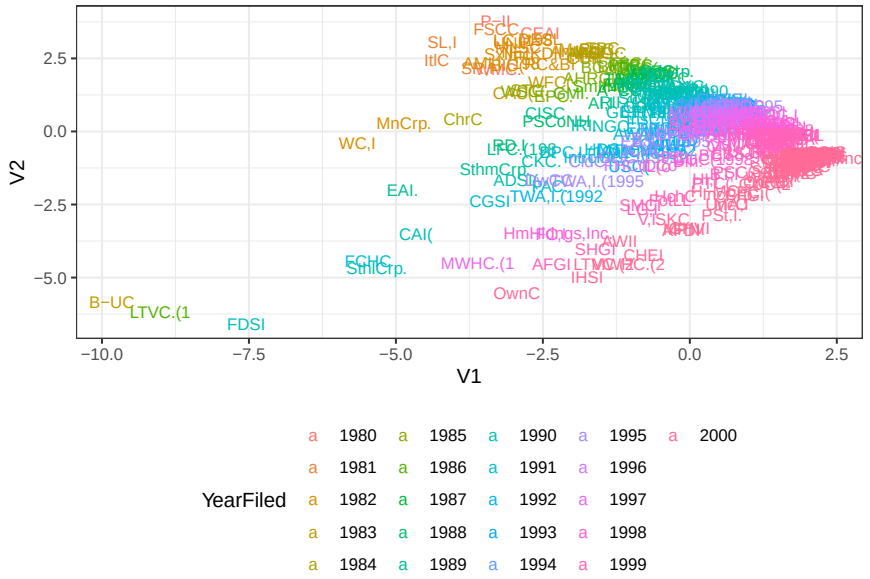

Figure 8: Classical MDS solution plotted by year when the bankruptcy filed.

The classical MDS solution plotted by year as shown in Figure 8 shows that there is pattern regarding the year. Firms who filed for bankruptcy in the same year tend to be similar each other. This could be because in the same year, CPI, filing rate, and prime interest are pretty similar. This is an interesting finding since we could infer that macroeconomic ,i.e, market condition could profile firms who filed for bankruptcy.

4 Principal Component Analysis (PCA)

In this section, we perform PCA, a dimension reduction method by transforming a large set of variables into a smaller one while retaining the variance of the data. We display the PCA without the Texaco Inc. as done in MDS (see Appendix for PCA including the outlier observation). We standardized our variables to ensure the results are not sensitive to the units of measurement. Thus, giving us more accurate analysis.

4.1 The PCA

```
## Importance of components:
##                             PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.7767 1.6219 1.0429 1.0133 0.90027 0.82613 0.56680
## Proportion of Variance 0.3157 0.2631 0.1088 0.1027 0.08105 0.06825 0.03213
## Cumulative Proportion  0.3157 0.5787 0.6875 0.7902 0.87124 0.93948 0.97161
##                            PC8     PC9    PC10
## Standard deviation     0.48852 0.17761 0.11705
## Proportion of Variance 0.02387 0.00315 0.00137
## Cumulative Proportion  0.99548 0.99863 1.00000
```

From the output above we learn:

- Proportion of variance explained by the first four PCs together is now 67.99%.
- Proportion of variance explained by the first and second PC alone is 26.41% and 22.24%, respectively.
- Kaiser's rule, which choose the PC's whose variance and standard deviation greater than 1, suggests to choose 4 PCs. However, we will inspect it further using the scree plot.

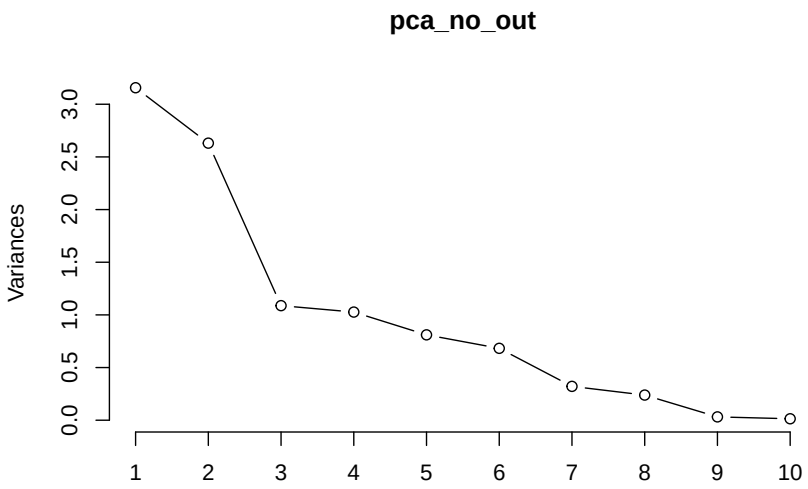

Figure 9: Screeplot of PCA on bankruptcy data

- Interestingly the scree plot and the Kaiser's rule do not agree with each other. However, we refer to scree plot since it is more distinguishable to show the elbow structure.
- The scree plot suggests to include 3 PCs which explain 68.75% variations of the data. However, we cannot visualise it in biplot, so we plot the first two PC's as they explain 57.87% of variation in our data.

From figure 10, we can infer the following:

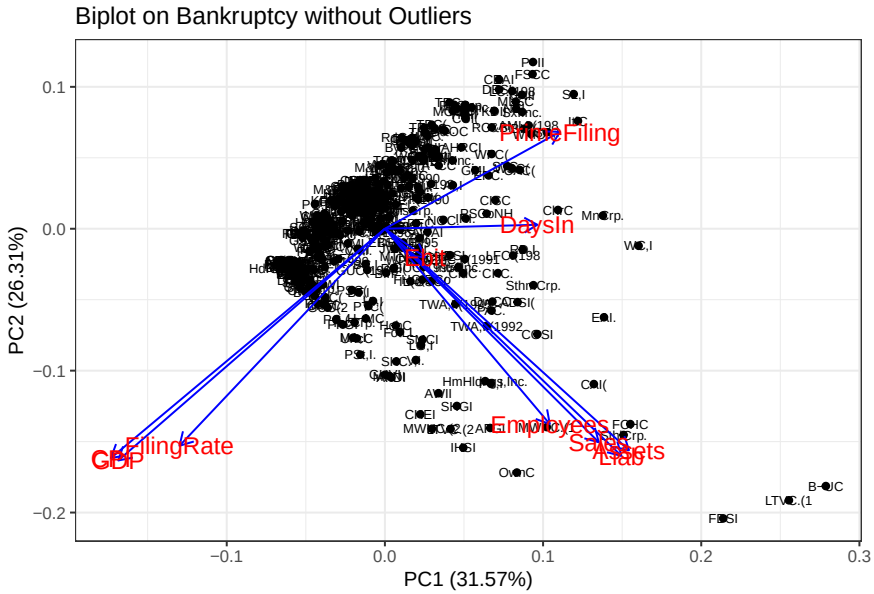

Figure 10: PCA Biplot without Outliers

- The firms are spread out in different direction showing different properties with the surrounding variables. For instance, two companies that show similar characteristics are FDSI and LDVC and the ones that show very dissimilar characteristics to each other are FDSI and DESL.
- The further away these vectors or variables are from a PC origin, the more influence they have on that PC. For instance taking a closer a look at the left quadrant, we infer that CPI has more influence, followed by `GDP` and `FilingRate` whereas `Ebit` has the least influence among all the variables.
- `Employees` and `GDP/CPI` shows an angle of almost 90°. Thus, there is no correlation between these variables.
- `PrimeFiling` is negatively correlated with `GDP`, `CPI`, and `FilingRate` since the angle is almost 180°.
- `CPI` and `GDP`are highly positively correlated. `Assets`, `Liab`, `Ebit`, and `Sales` are all also positively correlated as the angle between them is nearly zero.

Figure 11 shows a clearer visuals of the angles between variables so that we can distinguish between the variables that are positively, negatively, or not correlated at all.

5 Cluster Analysis

Clustering is the process of grouping the observations into categories or groups of observations based on their similarities. It helps us identify the commonalities and similar characteristics within the data which help us answer business questions.

In this case, we can validate if there are any similarities among the companies that are going bankrupt. This helps us understand the most possible reason for organisations going bankrupt and possibly help prevent the same in the future.

There are two main types of clustering namely, hierarchical and non-hierarchical clustering. Non-hierarchical clustering with the number of clusters known ex ante are useful if the number of clusters can be determined beforehand. We will mainly focus on Hierarchical clustering method.

5.0.1 Hierarchical Clustering:

Ward D2 method gives the best clustering outcome with reasonable sized clusters so we compare all other clusters with this method (See Appendix for other methods).

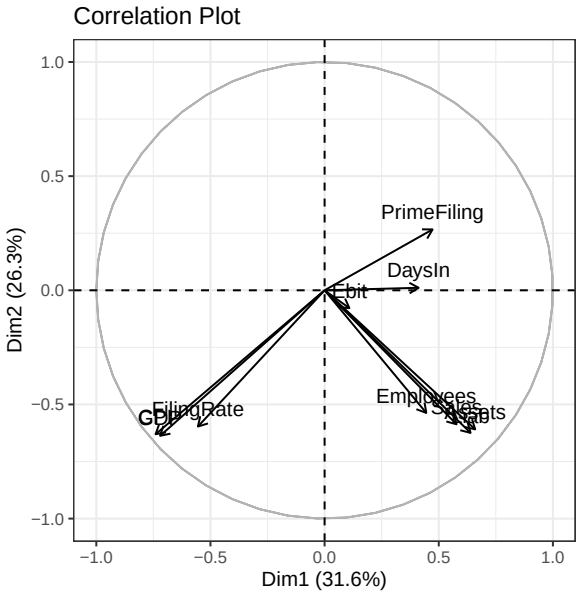

Figure 11: PCA Correlation Plot

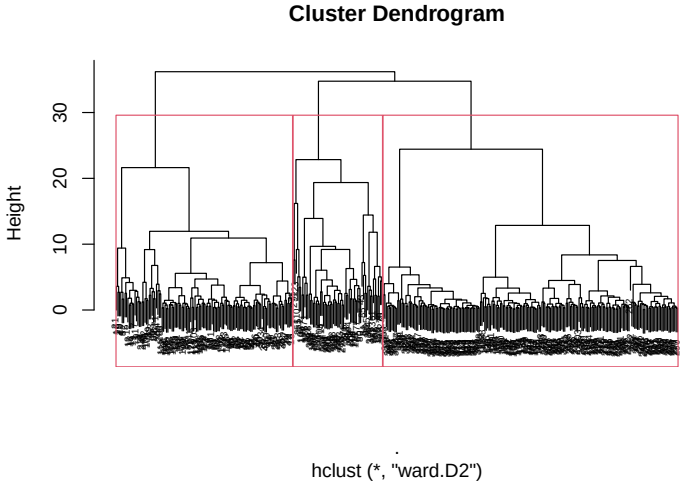

Figure 12: Dendrogram with Ward's menthod and 3 clusters.

Figure 12 shows that there is a big range of Height tolerance on y-axis for k = 3 with the next values at 34 and and 23 for k = 2 and k = 4, respectively (see Appendix for height with different k). Thus, we cluster the data into three clusters as it has the highest range which keeps our clustering stable.

5.0.2 Choosing the best method:

Average Method:

```
## [1] 0.05736127
```

Centroid Method:

```
## [1] 0.04238993
```

Complete Method:

```
## [1] 0.1293648
```

The complete linkage method has the highest agreement with Ward D2 between the different types of clustering. If all the clustering methods give almost similar results, then we can safely say that the data has some evident observable patterns. If not, the data does not have any solid patterns to cluster it based on similarities. In our case, there is no solid evidence of clustering results from the different methods. Each method gives an almost new result. Thus we can say that from the given data it is hard to find patterns that are making the organizations go bankrupt. The closest method to Ward D2 is Complete Linkage method which is 44% similar to the Ward D2 method.

6 Limitations

- We only use numerical data in the MDS analysis due to the complexity of incorporating non-numeric data. However, we tried to also display that categorical variable when visualising the MDS result.
- The scree plot infer that our data was mostly explained by the the first three PCs. However, due to the limitations of biplot we have visualized our data using the first and the second PCs only. This mean they might be certain patterns in the data and to visualize third PCs we will further need to work on the structure of the data.
- Since our focus is on hierarchical clustering, it does not consider all possible clusters like in k-means clustering.

7 Appendix

7.1 Data Cleaning

Imputation of variable `HeadCourtCityToDE`

As per our research online, we came to the conclusion that the `HeadCourtCityToDE` for Divi Hotels, N.V. is 1126 miles where as for Loewen Group, Inc (British Columbia to Wilmington) and Philip Services Corp. (Ontario to Wilmington) is 2942 and 1234 miles respectively.

Imputation of variable `DaysIn`

- `DaysIn` can be encoded equivalent to 121 days for AP Industries, Inc.
- `DaysIn` can be encoded equivalent to 1944 days for Daisy Systems Corp.

Dealing Missing Values in `Sales` and `Employees`

Table 6: Missing values in Sales

Name	Sales
County Seat, Inc.	NA

Table 7: Missing values in Employees

Name	Employees
County Seat, Inc.	NA

Checking Outliers

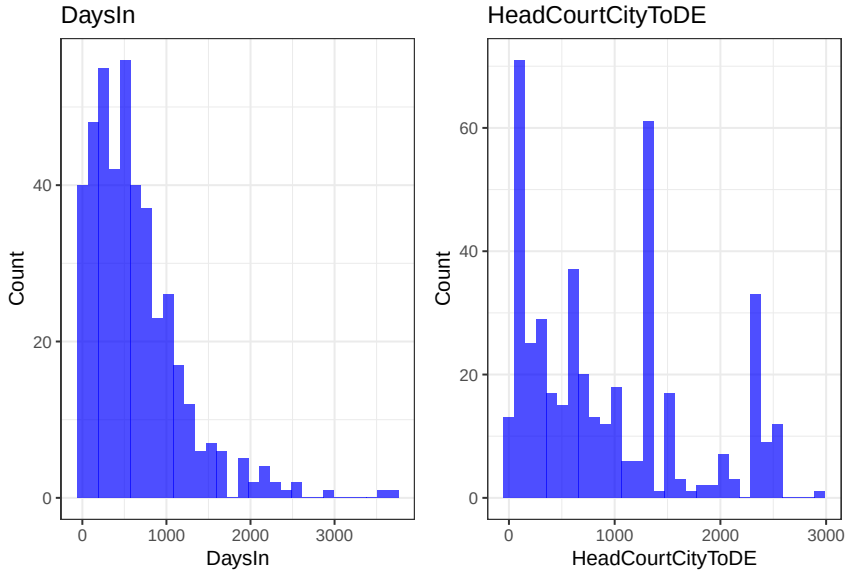

Figure 13: The figure indicates that there isnt any outliers in the variables DaysIn and HeadCourtCitytoDE

7.2 MDS

Eigenvalues of classical MDS

```
## [1] -0.000000000002189025
```

Since the values has e-12, it is reciprocal to 2 with 12 trailing zeros. Hence, even though it looks negative, it is very close, even indistinguishable from zero. That is why the value of GoF_1 and GoF_2 are equal.

MDS plot using Sammon mapping

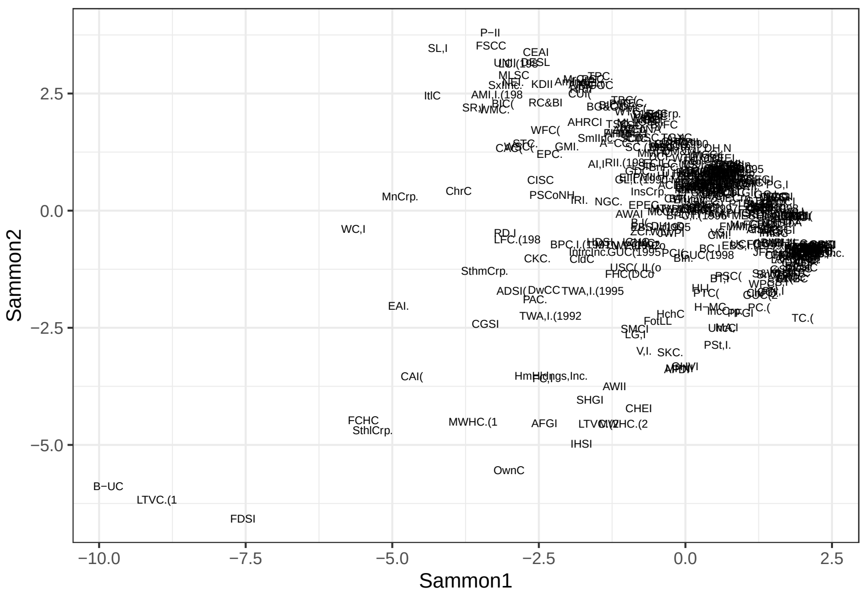

Figure 14: MDS solution using Sammon mapping

Additional plots of MDS based on the city where the bankruptcy filed

Figure 15 shows no specific pattern of bankruptcy regarding the city where it is filed. The firms who similar to each other (as seen in the overlapped text) could filed for bankruptcy in different city. Besides, firms who are potentially outliers (B-UC, LTCV.(1, and FDSI) are not filed their bankruptcy in Delaware.

Additional plots of MDS based on industry

Figure 16 shows the classical MDS solution by industry classification. Note that in the original data, there are 55 industry. This number is too big to be plotted, hence we collapse some industry which has the similar sector, for example manufacture, mining, construction, and finance.

Figure 16 suggests that there is no clear specific pattern of the firm bankruptcy regarding the industry. Wholesale and retail firms is bit more spread out. Manufacture industry is also observed to be spread out everywhere and could be because this industry has many observations. Further, B-UC and SthmCrp. are observed to be relatively further apart from the other real estate firms since they have bigger assets.

7.3 PCA

PCA performed on the complete data including outliers

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.8997 1.6691 1.0772 0.90927 0.83477 0.69985 0.49186
## Proportion of Variance 0.3609 0.2786 0.1160 0.08268 0.06968 0.04898 0.02419
## Cumulative Proportion  0.3609 0.6395 0.7555 0.83821 0.90790 0.95687 0.98107
##                            PC8     PC9    PC10
## Standard deviation     0.39471 0.14267 0.11477
## Proportion of Variance 0.01558 0.00204 0.00132
```

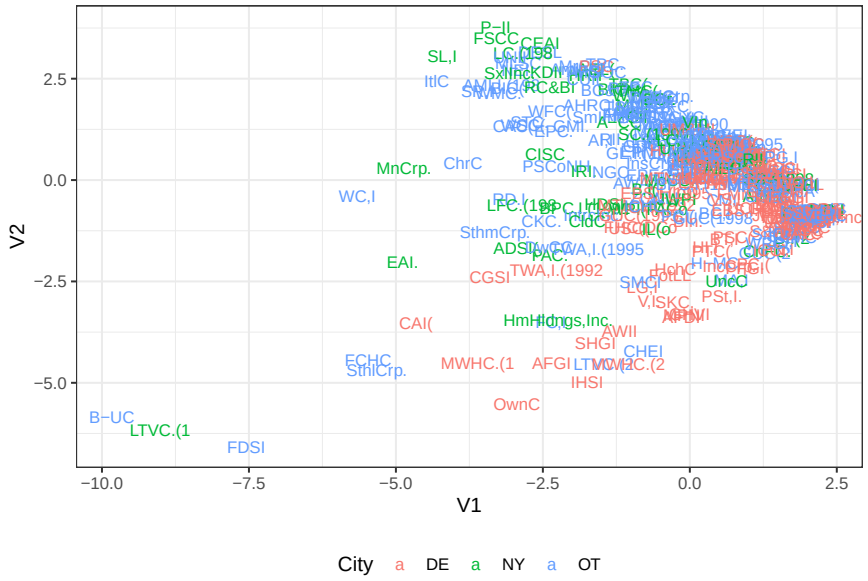

Figure 15: Classical MDS solution plotted by city where the bankruptcy filed.

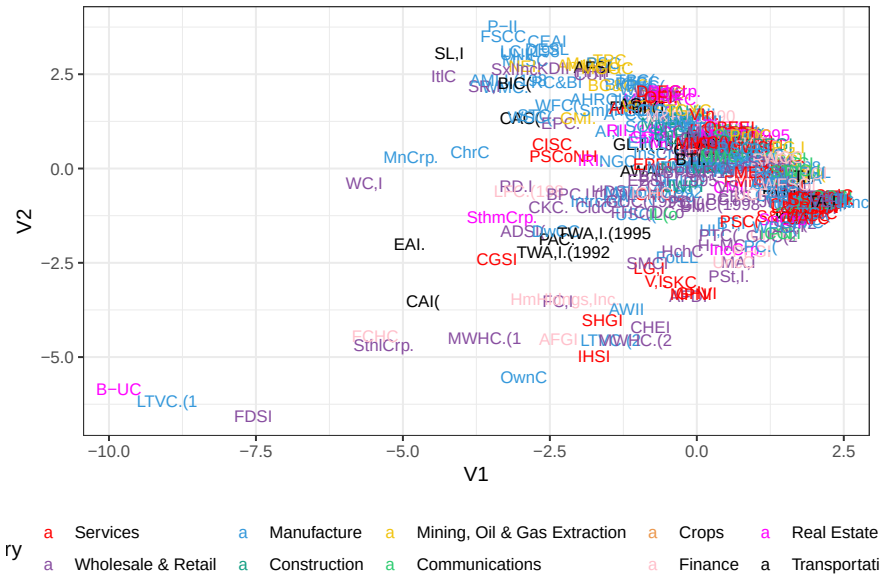

Figure 16: Classical MDS solution plotted by industry.

```
## Cumulative Proportion  0.99665 0.99868 1.00000
```

- Using the `summary` function we can infer the following:
  - Proportion of variance explained by the first four PCs together is 73.28%
  - Proportion of variance explained by the first and second PC alone is 30.10% and 23.64% respectively
  - Using kaisers rule, we choose those PC's whose variance and standard deviation is greater than 1, in our bankruptcy data we will choose 4 PC's.
- Let us now plot use the scree plot to find the total number of PC's that best explain our data.

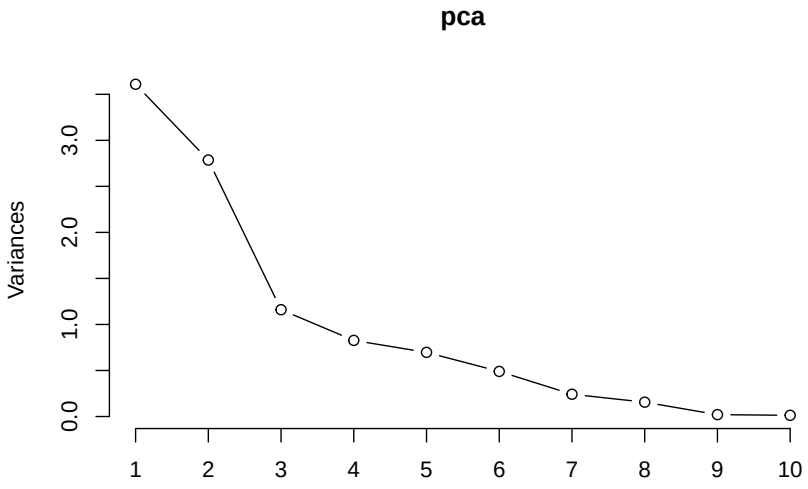

Figure 17: Scree Plot

- Using the scree plot we infer that our bankruptcy data is explained by the first three PC's. Also note this is different to kaisers rule.
- Let us now plot a biplot, that will help us infer intersting features about our data. A PCA biplot shows both PC scores of samples (dots) and loadings of variables(vectors).

7.4 Cluster Analysis

The tolerance values for different number of clusters + when k = 2 height is around 34 + when k = 3 height is around 29 + when k = 4 height is around 23 + when k = 6 height is around 21 + when k = 7 height is around 18

Plots for Average, Complete, and Centroid

8 Acknowledgment

Our sincere gratitude goes out to Ruben Loaiza-Maya and the our tutor Ari Handayani for their guidance and support. Their culminating efforts have placed us in a situation where we can produce this report collectively while showcasing our competence to employ diverse solutions that can be used with high dimensional data.

9 References

Loaiza-Maya, Ruben. (2021). Cluster Analysis. High Dimensional Data Analysis Lecture Notes.

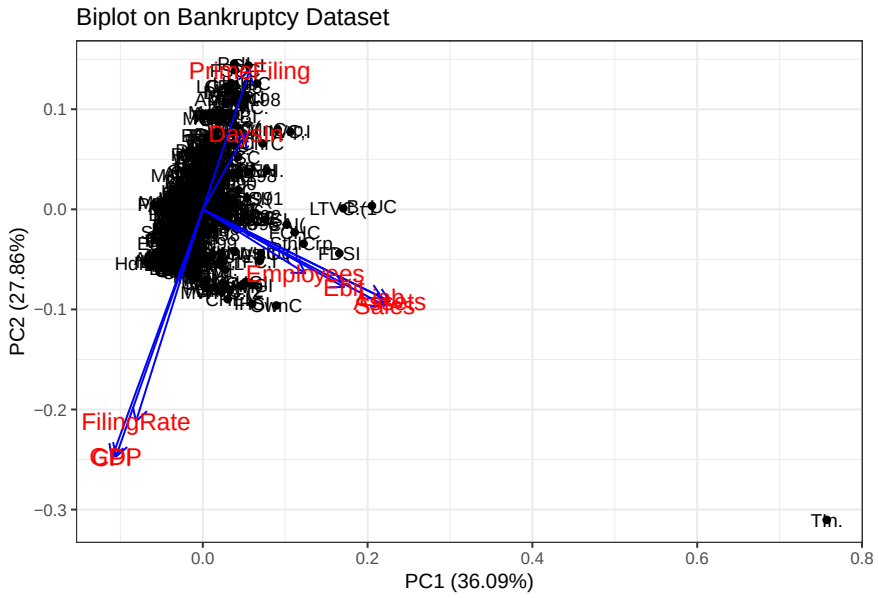

Figure 18: PCA Biplot

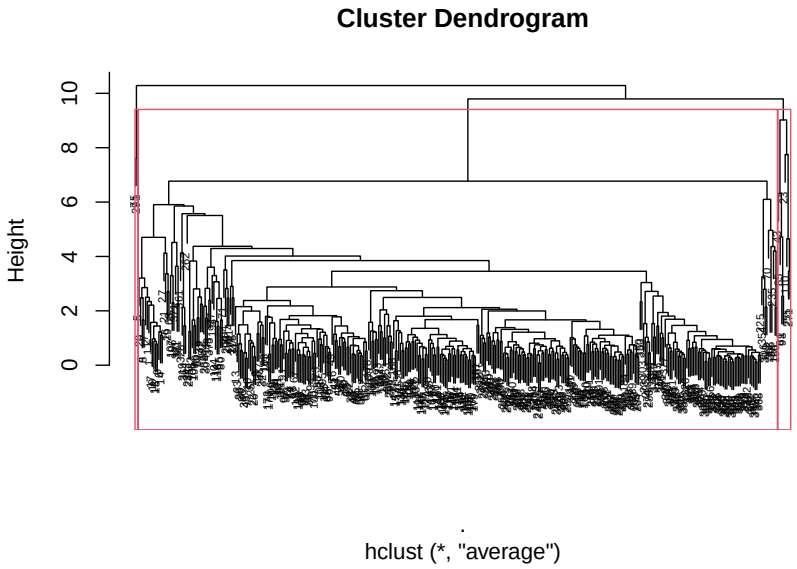

Figure 19: Dendrogram with average method

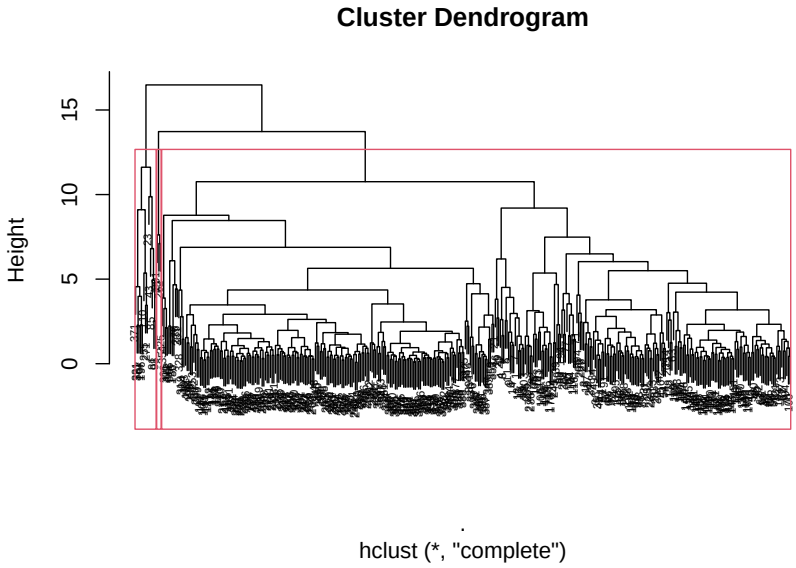

Figure 20: Dendrogram with complete method

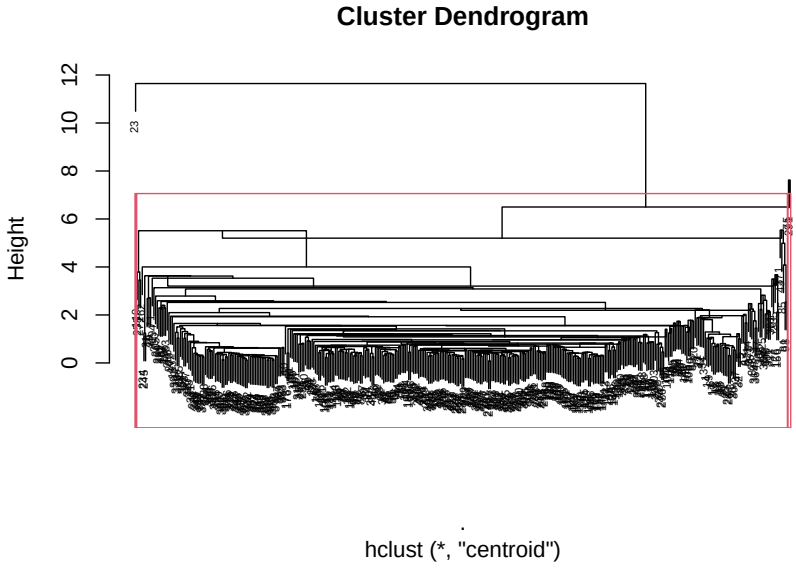

Figure 21: Dendrogram with centroid method

Loaiza-Maya, Ruben. (2021). Multidimensional Scaling. High Dimensional Data Analysis Lecture Notes.

Loaiza-Maya, Ruben. (2021). Principal Component Analysis. High Dimensional Data Analysis Lecture Notes.

10 R Packages Used

Allaire JJ and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2020). rmarkdown: Dynamic Documents for R. R package version 2.5. URL https://rmarkdown.rstudio.com.

Grolemund G., Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL https://www.jstatsoft.org/v40/i03/.

Horikoshi M., Yuan Tang (2016). ggfortify: Data Visualization Tools for Statistical Analysis Results. https://CRAN.R-project.org/package=ggfortify

Kassambara A., Fabian Mundt (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra

Le S., Julie Josse, Francois Husson (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, 25(1), 1-18. 10.18637/jss.v025.i01

Mahto A (2019). splitstackshape: Stack and Reshape Datasets After Splitting Concatenated Values. R package version 1.4.8. https://CRAN.R-project.org/package=splitstackshape

Pedersen T.L. (2020). patchwork: The Composer of Plots. R package version 1.1.1. https://CRAN.R-project.org/package=patchwork

Scrucca L., Fop M., Murphy T. B. and Raftery A. E. (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models The R Journal 8/1, pp. 289-317

Slowikowski K (2021). ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. R package version 0.9.1. https://CRAN.R-project.org/package=ggrepel

Tierney N (2017). "visdat: Visualising Whole Data Frames." *JOSS*, *2*(16), 355. doi: 10.21105/joss.00355 (URL: https://doi.org/10.21105/joss.00355), <URL: http://dx.doi.org/10.21105/joss.00355>.

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Wickham, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7. https://CRAN.R-project.org/package=dplyr

Xie, Y. (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.33.

Xie, Y. (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.33.

Zhu, Hao (2021). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra