# High Dimensional Data

9/18/2021

## Contents

# 1 Introduction

The purpose of this assignment is to conduct a preliminary analysis on a set of US bankruptcy data in order to better understand and potentially categorise different types of bankruptcy. Before this analysis takes place, a brief data description is provided, followed by some data wrangling and exploration to first clean the data, and then visualise some of the patterns and characteristics present in it.

The main body of the analysis has been broken down into two parts. The first will conduct multidimensional scaling (MDS), a technique that aims to produce a low dimensional representation of a higher dimensional space by attempting to produce pairwise Euclidean distances that are as similar as possible to the original distances. This technique will help visualise which observations are similar from the data and whether any different groups can be observed.

The second stage will involve undertaking a principal components analysis (PCA) which reduces the number of variables by producing a linear combination of the original variables that explains the largest amount of variance. This will provide a good comparison to the results of the MDS while also providing greater insight, as the loadings assigned to each principal component can be examined.

Following this analysis, the results from each will be discussed and compared in the conclusion.

## 1.1 Data Description

The subject data of this analysis is based on a data set containing information at the time of filing on 436 organisations from the United States that went bankrupt between 1980 and 2000. The original source for this data is the UCLA-LoPucki Bankruptcy Research Database.

The data set provided contained the following 22 variables (**Variable Name** (Variable Type): Variable Description) :

- **Name** (character): Name of the firm
- **Assets** (integer): Total assets (in millions of dollars)
- **CityFiled** (character): City where filing took place
- **CPI** (numeric): U.S CPI at the time of filing
- **DaysIn** (integer): Length of bankruptcy process
- **DENYOther** (character): CityFiled, categorized as Wilmington (DE), New York (NY) or all other cities (OT)
- **Ebit** (numeric): Earnings (operating income) at time of filing (in millions of dollars)
- **Employees** (integer): Number of employees before bankruptcy
- **EmplUnion** (integer): Number of union employees before bankruptcy
- **FilingRate** (integer): Total number of other bankrupcy filings in the year of this filing
- **FirmEnd** (character): Short description of the event that ended the firm's existence
- **GDP** (numeric): Gross Domestic Product for the Quarter in which the case was filed
- **HeadCityPop** (numeric): The population of the firms headquarters city
- **HeadCourtCityToDE** (integer): The distance in miles from the firms headquarters city to the city in whic the case was filed
- **HeadStAtFiling** (character): The state in which firms headquarters is located
- **Liab** (numeric): Total amount of money owed (in millions of dollars)

- **MonthFiled** (integer): Categorical variable where numbers from 1 to 12 correspond to months from Jan to Dec
- **PrimeFiling** (numeric): Prime rate of interest on the bankruptcy filing date
- **Sales** (numeric): Sales before bankruptcy (in dollars)
- **SICMajGroup** (character): Standard industrial clasification code
- **YearFiled** (integer): Year bankruptcy was filed

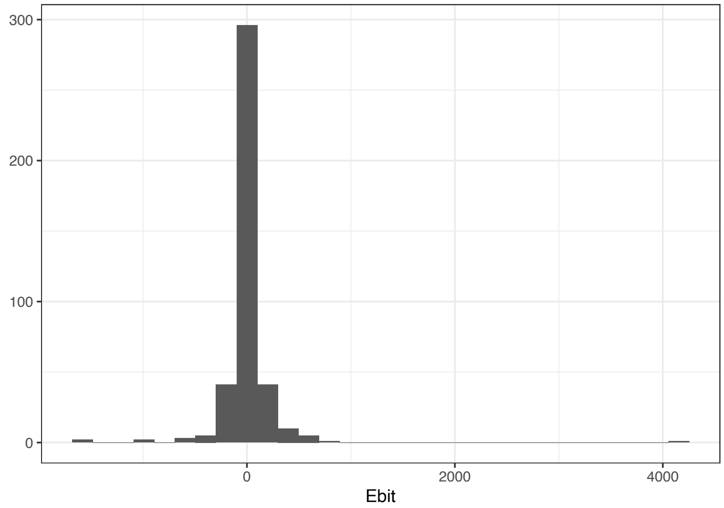
# 2 Preliminary Data Analysis

## 2.1 Data Wrangling

Before conducting the analysis, it is first important to check for any missing values, outliers, or other credibility issues as this may impact the findings later on.

Beginning with the outliers first, histograms of each of the variables were produced. Of these, three displayed seemingly significant outliers; Ebit, Sales, and Liab. The histograms for both Ebit and Sales are shown below.

Figure 1: Histogram for Ebit



As can be seem from the plots above, there was a single firm with both sales and earnings far higher than any other in the data set that nonetheless went bankrupt. This is unusual as it is unlikely that a firm that is selling a lot of products and generating a lot of income would go bankrupt. More surprisingly, when the data was examined, both these potential outliers were from the same firm, Texaco Inc. (shown below).

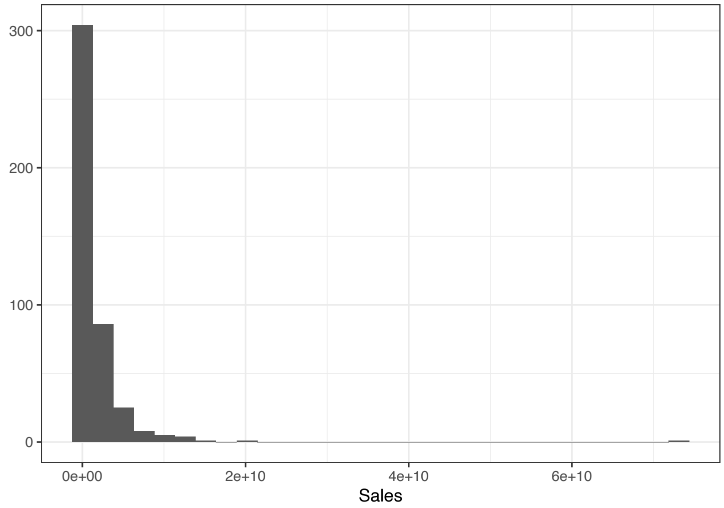
Further research regarding this particular bankruptcy revealed it was for legal rather than financial purposes. Thus, it was considered to be removed from the data. However, our group decide to keep it since we will use logarithmic transformation[1] on our dataset to handle such issue, including outliers.

Table 1: Outlier for Ebit and Sales

| Name | Ebit | Sales |
|---|---:|---:|
| Texaco Inc. | 4222.978 | 73130213265 |

[1] Our discussion about logarithmic transformation is included in the Appendix

Figure 2: Histogram for Sales



The third variable to reveal a potential outlier was liability. A histogram of this variable and the associated firm are shown in the output below.

Figure 3: Histogram for Liability

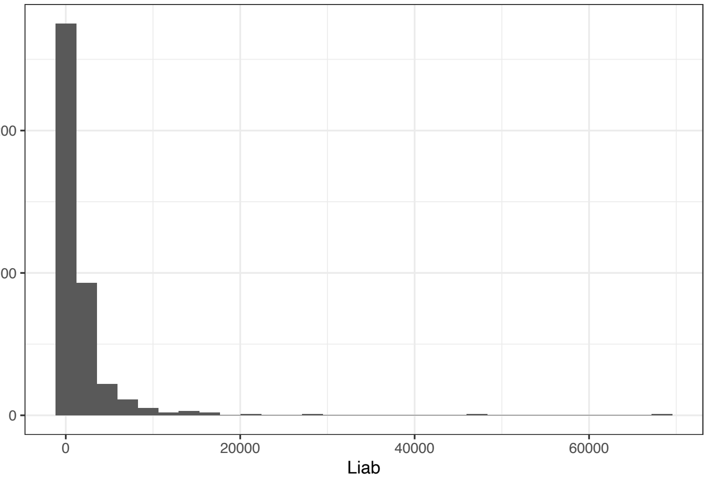


Table 2: Outlier for Liability

| Name | Liab |
|---|---:|
| First RepublicBank Corp | 68403.04 |

The highly positive value is less unusual than that for sales or earnings as it is possible, and very likely, that a firm with a large amount of liability would go bankrupt. Additionally, no further research revealed any alternative reasons as was the case with Texaco. Thus, this observation was not removed.

The next stage involved preforming data validation, ensuring the credibility of what was recorded. This included checking for implausible values. Exploration unveiled that for the Promus Companies Inc., there were actually more employees recorded as belonging to a union than there were employees total at the firm. As such, this data cannot be trusted and was also removed from the analysis.

Table 3: Implausible Employee values

| Name | Employees | EmplUnion |
|---|---:|---:|
| Promus Companies Inc. (Harrahs Jazz Co. only) | 1 | 3000 |

The Employees variable was also checked and it was found that there were a few firms with only a single employees, even though they still had assets and sales in the hundreds of millions of dollars. As this was deemed implausible these 6 companies, shown in the table below, were also removed prior to the analysis as they represented a small sample of the total data set. Furthermore, the next smallest firm had 28 employees, and the third smallest had 68. This significant drop-off provided further evidence for these observations removal.

Table 4: Counts for Employee variable

| Employees | Number of Observations |
|---:|---:|
| 1 | 6 |
| 28 | 1 |
| 68 | 1 |
| 74 | 1 |
| 80 | 1 |

Table 5: Observations with Only 1 Employee

| Name | Employees | Assets | Sales |
|---|---:|---:|---:|
| Residential Resources Mortgage Investments Corp. | 1 | 513 | 33211255.5 |
| Mortgage & Realty Trust (1990) | 1 | 1022 | 39131452.4 |
| EUA Power Corp. | 1 | 686 | 19696139.6 |
| NACO Finance Corp. | 1 | 328 | 41451879.8 |
| Commonwealth Equity Trust | 1 | 489 | 55755204.6 |
| Promus Companies Inc. (Harrahs Jazz Co. only) | 1 | 1095 | 479098.5 |

A similar examination was performed on the financial side, comparing a companies earnings against its sales prior to bankruptcy. This revealed two firms who generated an income several magnitudes higher than there recorded sales. This also is not plausible as earnings is a function of sales. Thus, these two firms were also removed.

Table 6: Implausible Ebit values

| Name | Ebit | Sales_Million |
|---|---:|---:|
| McLean Industries Inc. | 95.31978 | 29.05200 |
| Mortgage & Realty Trust (1990) | 99.67475 | 39.13145 |

The final wrangling stage involved checking for missing values, as any observations may not be included in the subsequent analysis. The table below shows the number of missing values for each variable in the data set. As can be seen, most are quite small compared to the total number of observations (436) and thus won't contribute to much data loss. However, the EmplUnion value has 133 missing records, which is over 30% of the total observations.

Table 8: Counts of EmplUnion variable

| Variable | Median |
|---|---:|
| EmplUnion | 1 |

Table 7: Missing observations per variable

| Variable | Missing Observations |
|---|---:|
| EmplUnion | 131 |
| Ebit | 26 |
| Liab | 16 |
| DaysIn | 4 |
| HeadCourtCityToDE | 3 |
| Employees | 1 |
| Sales | 1 |
| Name | 0 |
| Assets | 0 |
| CityFiled | 0 |
| CPI | 0 |
| DENYOther | 0 |
| FilingRate | 0 |
| FirmEnd | 0 |
| GDP | 0 |
| HeadCityPop | 0 |
| HeadStAtFiling | 0 |
| MonthFiled | 0 |
| PrimeFiling | 0 |
| SICMajGroup | 0 |
| YearFiled | 0 |
| Period | 0 |

To deal with this issue, one possible solution is to exclude the EmplUnion variable from the data set and not include it in the analysis.

An alternative solution would to be impute the median value and see if the results of the analysis are similar. Examining the table below reveals that the median value for this variable is 1. After having group discussion on this issue, our group decide to do median imputation since this approach is easy to implement and we can still maintain to get a complete dataset. In addition, because some variable with missing values are skewed, so median is a better representation of majority of the observation in variables[2] (Amballa 2020).

It is first important to examine the original high dimensional space before conducting any of the reduction techniques. The plot below shows the original numeric variables. Several key insights can be gained from examining this graphic.

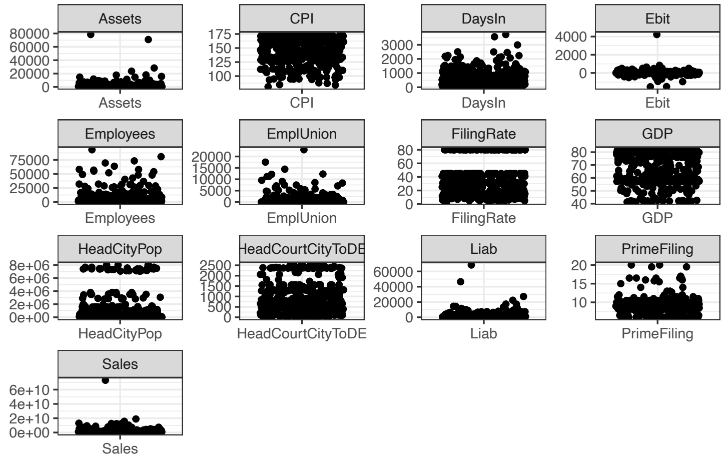
Firstly, most the firms that went bankrupt had low values for Assets, Employees, EmplUnion, Liab, and Sales. This makes sense as all of these indicate that a corporation is struggling financially, thus making bankruptcy necessary. However, there are some firms went bankrupt that were performing quite well at the time they collapsed. Thus, there bankruptcy may be due to another unexplained factor. Similarly, the Ebit is mostly centered around zero which indicates that most of the recorded bankruptcies occured when the company was making no profit or even losing money.

Secondly, there also seems to be a geographic factor at play with a large number of bankruptcies occurring in cities with a low population whearas there are fewer occurences in larger cities.

[2] This issue is also discussed in the Appendix

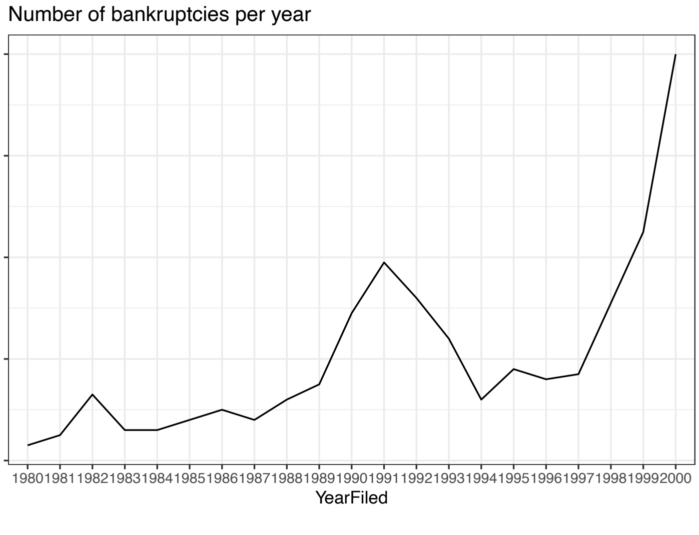
A third key takeaway is that the broader economy appears to play an important role in the number of bankruptcies. Bankruptcies occur most commonly when the CPI is high and the PrimeFiling rate is low, while GDP is more dispersed. This indicates that periods of inflation where money is lent with little interest contribute to bankruptcies.

Figure 4: Scatterplot Matrix for Original Numeric Variables



The same can be done for the temporal variables, month and year. Beginning with the year variable it appears that there are various peaks in bankruptcies in 1982, 1991 and then a sharp upwards trend into the year 2000. These period align up with a recession in the early 1980s, one in the early 1990s and the Dot-com bubble of the late 1990's and early 2000's. This supports the previous visualization which showed economic activity as a potential driving force behind bankruptcies. Meanwhile for the month variable, it appears that most bankruptcies occurred in the second half of the year before dropping off sharply in the new year. This may coincide with the end of financial year at the end of June as companies make the decision to declare bankruptcy after a poor reporting season.

Figure 5: Number of Bankruptcies per Year



The data can also be viewed geographically, as has been done below, which reveals there are quite a large number of bankruptcies in Texas, California and the north-east around New York State.

Figure 6: Number of bankruptcies per month

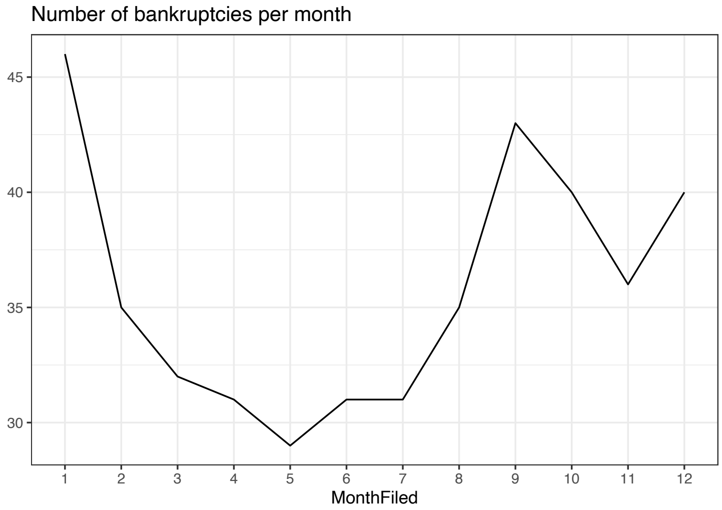
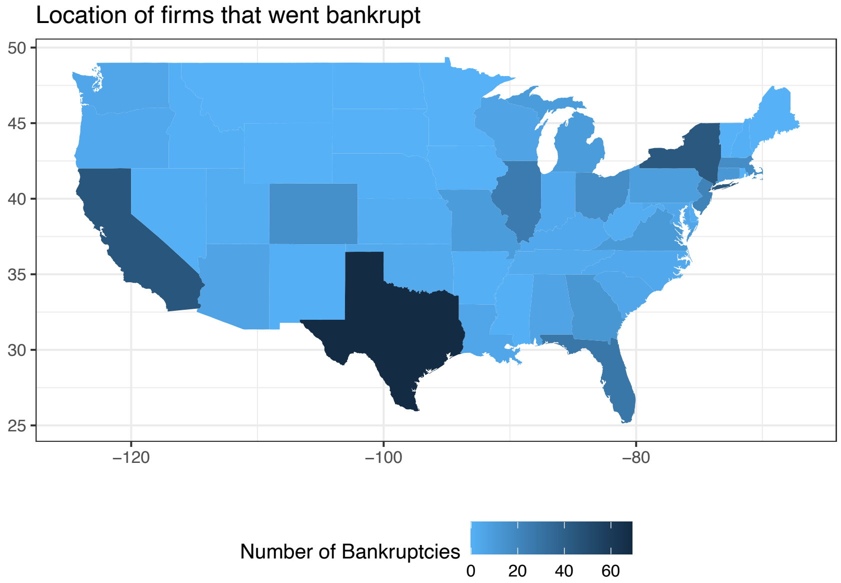


Figure 7: Distribution of bankruptcies by state



# 3 Analysis and Findings

## 3.1 Multi Dimensional Scalling (MDS) Analysis

This section will be discussing our analysis using Multi Dimensional Scaling (MDS) to project our dataset that contains many variables/dimensions to a low dimensional representation. For this purpose, we will be using pairwise 2D Euclidean distances in measuring metric variables in the data. The projection will guide us on finding similarity/dissimilarity between observations regarding to their distances.

### 3.1.1 Methodology

Using our tidy dataset as prepared in the Data Wrangling section, we only select metric variables and compute their Euclidean distances.

The next step is standardizing the metric variables because they are measured in different units, for example `Assets` is stated in Millions Dollars, `Sales` in Dollars, while `HeadCourtCityToDE` in miles. Standardizing variables is conducted by subtracting their values with the mean of each variable and then dividing the result by the variable's standard deviation (UCLA 2019). This can be done easily using `scale` function in R.

After all variables are standardized, we compute the distance using `dist` function to extract the Euclidean distances between every possible pair of the US Firm. Then, the data is ready to compute with the Classical MDS using `cmdscale` function. The results of the MDS computation are two new variables representing the high dimensional distances that come from the true data ($d_(ij)$ & low dimensional distances that come from the solution $\delta_(ij))$. These two distances are used to minimise Strain.

$$Strain = \Sigma_{i=1}^{n-1}\Sigma_{j>1} = (\delta_(ij) - d_(ij))^2$$

The next process will be plotting the distances into a scatter plot to identifying the similarity/dissimilarity between the observations. Lastly, to evaluate the result we will calculate the Goodness of Fit (GOF) of the solution. the GOF measurement are based on eigenvalues which should always positive for the Euclidean distance.

Considering nature of the data (i.e. some missing values and outliers), we have several different approaches for processing the data in MDS.[3] Nevertheless, we come up with one main approach that produces the highest GOF value of 0.4894409 for both GOF1 and GOF2 (both are equal) as well as a minimum eigenvalue that really close ro zero ($-2.3123692 \times 10^{-12}$). Furthermore, the chosen approach can be explained as follows:

a. Missing values are managed by imputing median values of each variable.

b. Some variables have really skewed distribution thus transformed by using logarithmic transformation. The log transformed variables can be seen at Figure 8

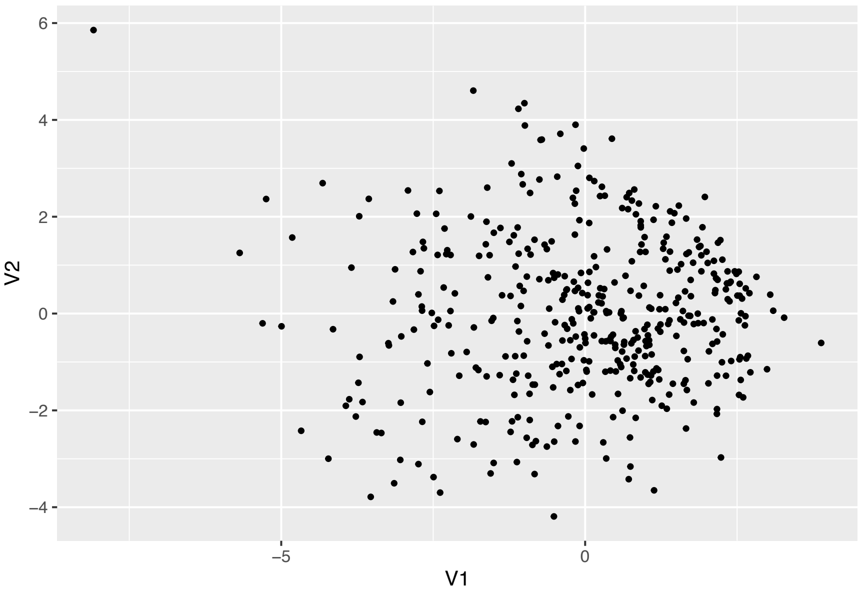
### 3.1.2 MDS Analysis by Category

In this section we will map the MDS distance into scatter plot and colored them by some variable that can clearly group observations based on their similarity/dissimilarity of the features.

After observing the MDS plot with the combination of several variables in the dataset, we decide to select the most interesting variables; **YearFiled** and **Sales**, on which we will focus and breakdown our analysis further.

[3] Full explanation of 5 possible approaches can be seen in the Appendix

Figure 8: MDS plot using Log Transformation



**3.1.2.1 Year of Case Filling** To observe the relationship between the MDS distances with variable YearFiled, we divide companies into 4 different time period groups depending on their year of bankruptcy filing:

a. Companies with filing year less than or equal to 1985 are grouped in "1980-1985,"

b. Companies with filing year more than or equal to 1986 but less than or equal to 1990 are grouped in "1986-1990,"

c. Companies with filing year more than or equal to 1991 but less than or equal to 1995 are grouped in "1991-1995,"

d. Companies with filing year more than or equal to 1996 but less than or equal to 2000 are grouped in "1996-2000."

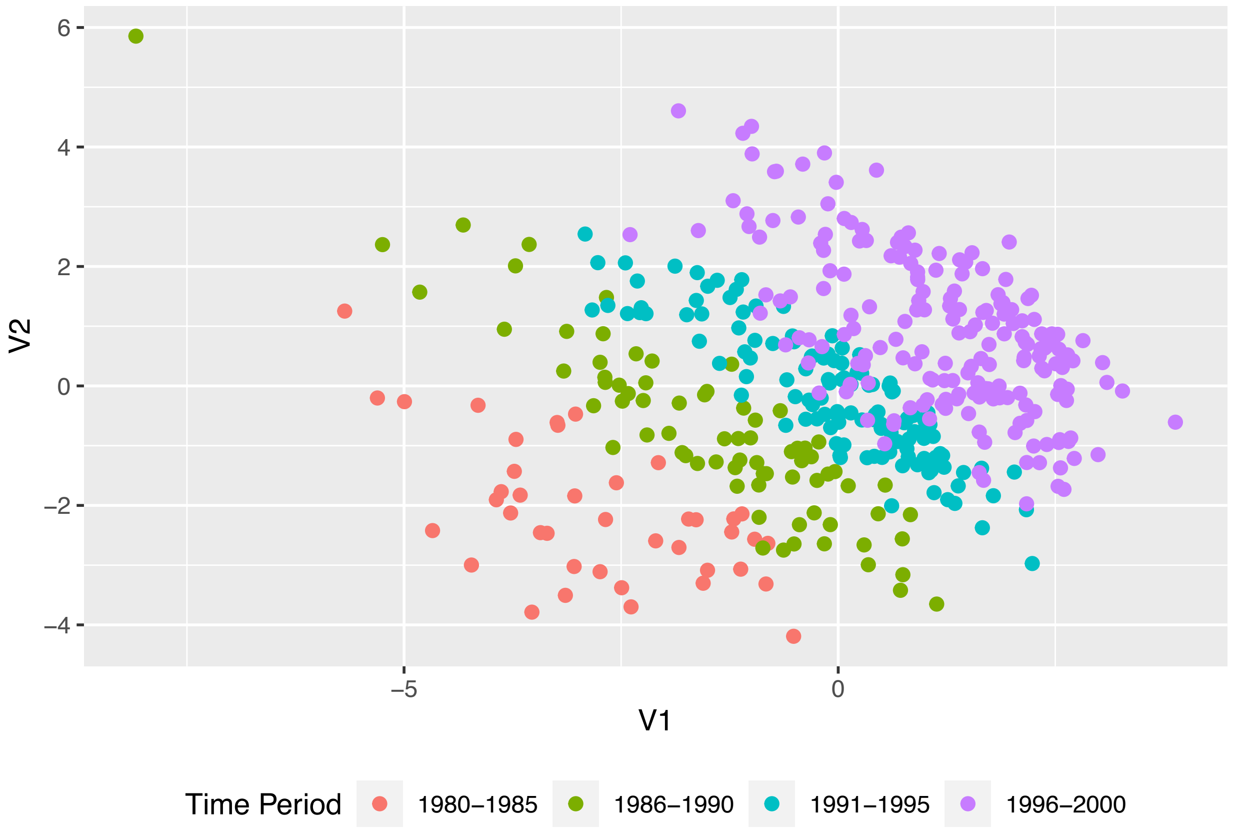
As we can see from figure 9, more recent time period seems to have larger number of bankruptcy cases. Moreover, some portions of companies that filing for bankruptcy in period 1996-2000 have similarity with several companies that filing for bankruptcy during 1991-1995 (indicated by overlapping purple and blue dots). On the contrary, companies that filing for bankruptcy in the period of 1980-1985 (shown by red color) tend to have more disperse features (lower similarity) than not only the other groups but also within the group itself. Among these four groups, the 1991-1995 seems to have smallest within-group distance.

From Fig. 9, we can also see that there is 1 company that is really separated away from the larger groups. This far distance can be articulated as having very high dissimilarity. The company is Texaco Inc.. After checking the dataset, its original attributes, i.e. assets, sales and liabilities, is gigantic to the most of companies.

For the largest time group (1996-2000), we can notice here that there are 2 observations that have far V1 distances from the group's center. These two companies are Towner Petroleum Co. and Texaco Inc.. Furthermore, as the period of 1996-2000 consist of largest cases, we will analyze it further in a particular section of Principal Component Analysis (PCA) to aid in interpretation of dimensions.

**3.1.2.2 Sales** In this part, we group companies based on their sales volume (in log) category. We create 4 category which can be explained as below:

Figure 9: MDS Analysis Coloured by Time Period



a. Companies with log sales volume below or equal to 25% quantile are cateorized as "low sales,"

b. Companies with log sales volume larger than 25% quantile but lower than or equal to 50% quantile are categorized as "medium sales,"

c. Companies with log sales volume larger than 50% quantile but lower than or equal to 75% quantile are categorized as "high sales,"

d. Companies with log sales volume greater than 75% quantile are categorized as "very high sales."

Table 9: Table of Sales Quantiles

| | Sales |
|---|---|
| 0% | 12.54784 |
| 25% | 19.75100 |
| 50% | 20.47723 |
| 75% | 21.15054 |
| 100% | 25.01551 |

The summary of the grouping can be seen in Table 10. We have quite balanced observations in the grouping.

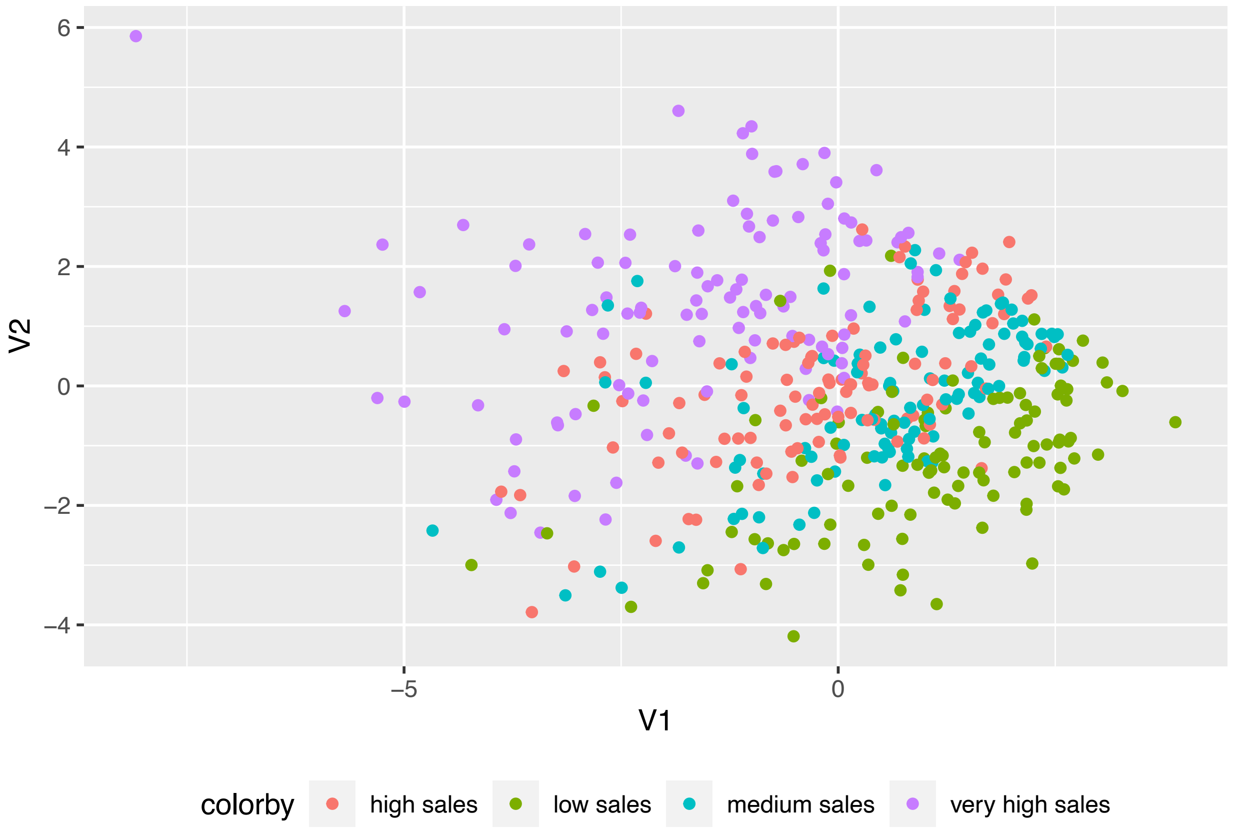
Table 10: Group of Companies by Sales

| | Companies |
|---|---|
| high sales | 107 |
| low sales | 108 |
| medium sales | 107 |
| very high sales | 107 |

The result of MDS analysis coloured by log sales volume is shown in figure 10. In the figure, we can see that companies with "very high" log sales volume looks more disperse than companies with "low," "high" and "medium" log sales volume. In addition, the group also quite stand out from the others (only few portion of purple dots overlapping with dots of other colors). This indicates that the "very high" group have larger variance of features/attributes that make within and between groups distances are large.

Moreover, we can also see the same furthest observation from the previous MDS analysis ( that representing the "very high" sales group.

Figure 10: MDS Analysis Coloured by Sales

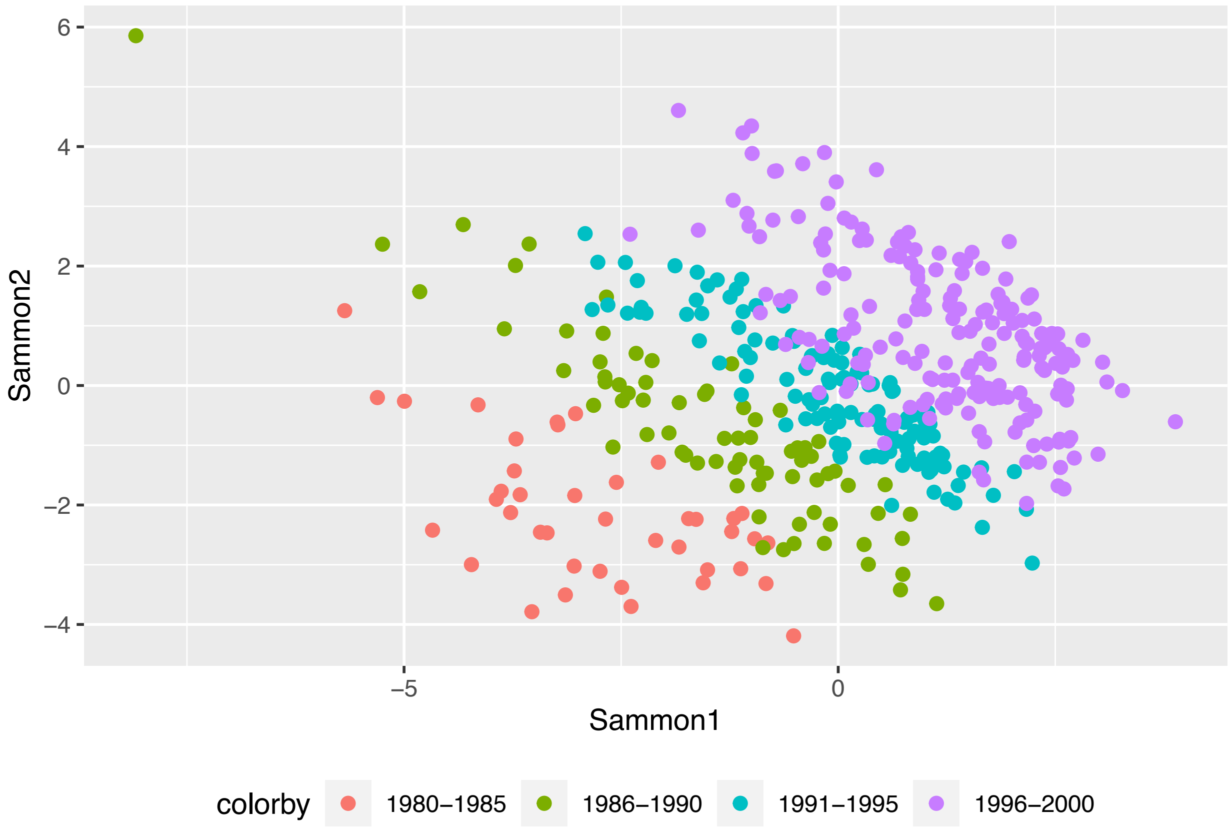


### 3.1.3 Sammon Mapping

In this section, we complement and compare our previous result from the classical MDS with Sammon mapping that considered as more robust solution. Sammon Mapping is type of MDS which preserve the local structure, not based on eigenvalues decomposition and a non linear mapping. Here we want to check the robustness of our clasical MDS. For this purpose, we are plotting the scatter plot of the fitted distances versus the observed distances.

```
## Initial stress        : 0.16273
## stress after   0 iters: 0.16273
```

Figure 11: Sammon Mapping Plot on the Log Transformed Data Colored by Time Period



Based of Figure 11 the plot has fairly similar pattern with the classical MDS, however we notice some observations in the 1996-2000 group separated in Sammon1. The conclusion that 1996-2000 groups are more diverse is perhaps a bit clearer when the Sammon mapping is used.

## 3.2 Principal Component Analysis (PCA)

In addition to the MDS, another method in dimension reductions is principal component analysis (PCA). In this section, the dimensionality of the bankruptcy dataset will be reduced by PCA. Following our analysis in the previous section, PCA will be conducted to the log transformed dataset for period 1996 to 2000. This period is selected as it has the highest numbers of bankruptcy filings compared to other time period.

The function `prcomp` from the `stats` package is used to compute the analysis of the PCA. In this instance, the variables are standardized as the variables used in the analysis are measured in different units. It is important to standardize the data before implementing PCA as the weights of the principal components are affected by different units of measurement (Jolliffe and Cadima 2016).

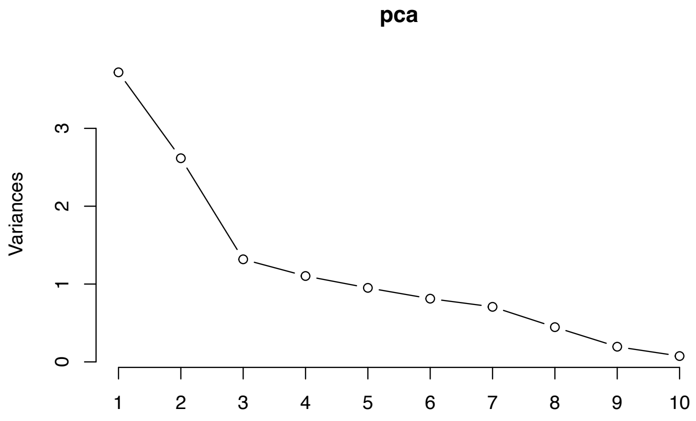
The result of the analysis are as follow:

```
## Importance of components:
##                          PC1   PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.929 1.617 1.1480 1.05042 0.97521 0.90125 0.84110
## Proportion of Variance 0.310 0.218 0.1098 0.09195 0.07925 0.06769 0.05895
## Cumulative Proportion  0.310 0.528 0.6379 0.72980 0.80905 0.87674 0.93569
##                            PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.66807 0.44180 0.27383 0.21455 0.09580
## Proportion of Variance 0.03719 0.01627 0.00625 0.00384 0.00076
## Cumulative Proportion  0.97289 0.98915 0.99540 0.99924 1.00000
```

### 3.2.1 Selecting the Number of Principal Components

Kaiser's Rule states that all principal components with a variance greater than 1 need to be selected. Following this rule, the first four principal components will be used for the analysis. The proportion of variance that is explained by the first four principal components is 72.98%.

Figure 12: Scree Plot of the Pricipal Component Analysis of the dataset



Another method that can be used to select the number of principal components is by looking at Scree plot. The number of principal components can be chosen by looking at where the plot flattens out. Figure 12 shows the Scree plot of the PCA of the bankruptcy dataset. In contrary to the Kaiser's rule, the Scree plot indicates that the numbers of principal components to be used for the analysis is three.

As the variance of PC4 is very close to 1, this report is going to rely on Scree plot in choosing the number of principal components. In addition, it is very clear depicted in the Scree plot that the curve flattens out after the third principal components.

### 3.2.2 PCA Rotations

The weights of each principal components are as follows:

Table 11: Principal Component Rotations of the 1996-2000 Bankruptcy Dataset

| Variables | PC1 | PC2 | PC3 |
|---|---:|---:|---:|
| Assets | 0.1939682 | 0.4689103 | -0.3483574 |
| CPI | 0.4721643 | -0.2216968 | -0.0385664 |
| DaysIn | 0.1342445 | -0.0283180 | -0.1161347 |
| Ebit | 0.1341896 | 0.1346624 | 0.5110734 |
| Employees | 0.2205726 | 0.4115236 | 0.2440794 |
| FilingRate | 0.4652208 | -0.2213541 | -0.0110513 |
| GDP | 0.4508524 | -0.2331650 | -0.0689473 |
| HeadCityPop | 0.0508928 | -0.0309242 | -0.5226993 |
| HeadCourtCityToDE | -0.0748551 | -0.0406411 | -0.2259733 |
| Liab | 0.1774761 | 0.4740109 | -0.3602922 |
| PrimeFiling | 0.3943043 | -0.1426590 | 0.0849223 |
| Sales | 0.1997279 | 0.4378089 | 0.2769929 |

From the result above, we can see that PC1 is more correlated (positvely) with GDP and CPI variables compared to the other variables. This indicates that the first principal component measures the economic condition in the United States at the time of filings. The larger the value of CPI and GDP, the smaller the value of the first principal component. In other words, PC1 value is higher when the economic conditions of the US (as indicated by GDP and CPI) is better.

On the other hand, PC2 are more associated with Employees, Sales, and Liab. This suggests that the second principal component is related to the state of the company before bankruptcy. The value of PC2 is higher when the value of Employees, Sales, and Liab are higher This means that PC2 is higher for larger companies.

In comparison to other variables, the third principal component is more associated with Ebit and HeadCityPop variables. Unfortunately, it is hard to interpret this principal component as there is nothing that is earnings at time of filing and the population of the firm's headquarter city. The value of PC3 is bigger when HeadCityPop is lower, and Ebit is higher.

### 3.2.3 Biplots

Figure 13 shows the distance biplot of the first and second principal components. Variables Liab, Assets, Sales, Employees have strong positive association with each other. In addition to those, variables PrimeFiling, GDP, CPI, and FilingRate are also positively correlated with each other. On the other hand, HeadCourtCitytoDE and Ebit have a strong negative correlation with each other.

There are many companies that are similar to each other, for example, Montgomery Ward Holding Corp and Flagstar Companies, Inc. In contrast, there are also companies that are very different from one another, for example National Energy Group and Owens Corning.

One company that is more isolated from the rest is Orbcomm Global, LP. This company is characterized by an association with low values of assets, sales, employees, and liability meaning it is a fairly small company (at least compared to others in the dataset). In addition, it is also characterized by high values of prime PrimeFiling, CPI, and FilingRate. High values of CPI, prime interest rate, and number of other bankruptcy filings mean that the economic conditions at the time of bankruptcy was filed was bad.

The biplot of PC1 and PC3 is shown in Figure 14. It can be seen from the graph that variables PrimeFilling, FilingRate, CPI, and GDP have a strong association with one another. On the other hand, Ebit and HeadCourtCitytoDE have a strong negative correlation with one another.

As shown, in the graph, Harvard Industries and Sun Television and Appliances are similar to one another. On the other hand, Ithaca Industries and Montgomery Ward Holding Corp are very different to one another.

Figure 13: Biplot of PC1 and PC2

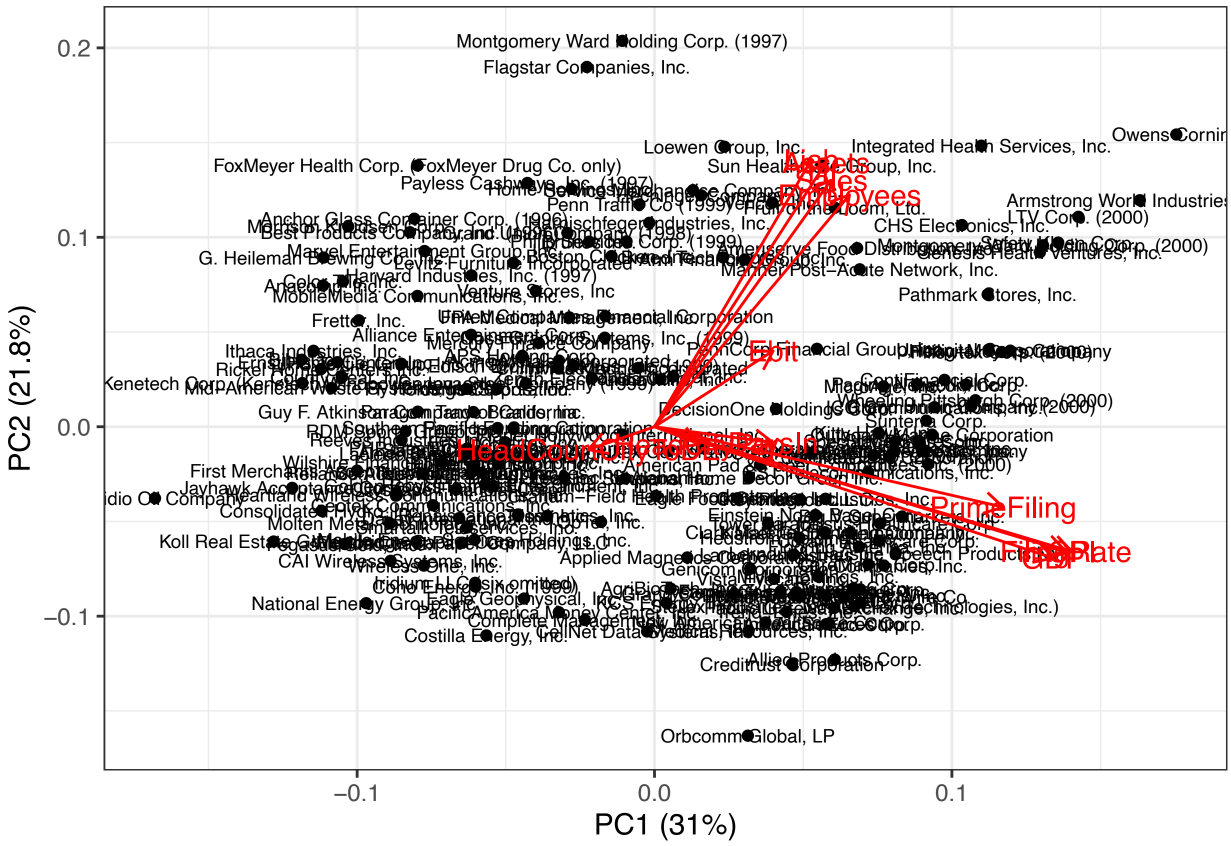
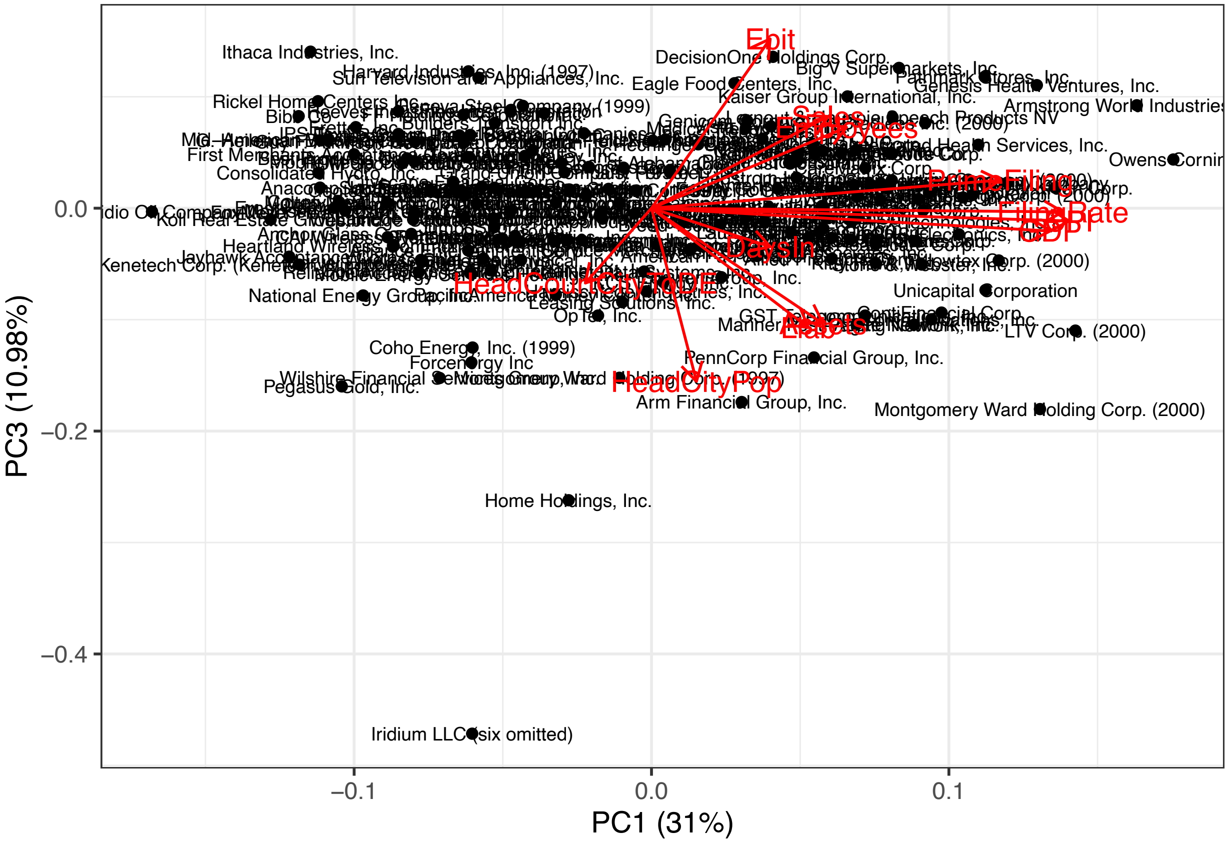
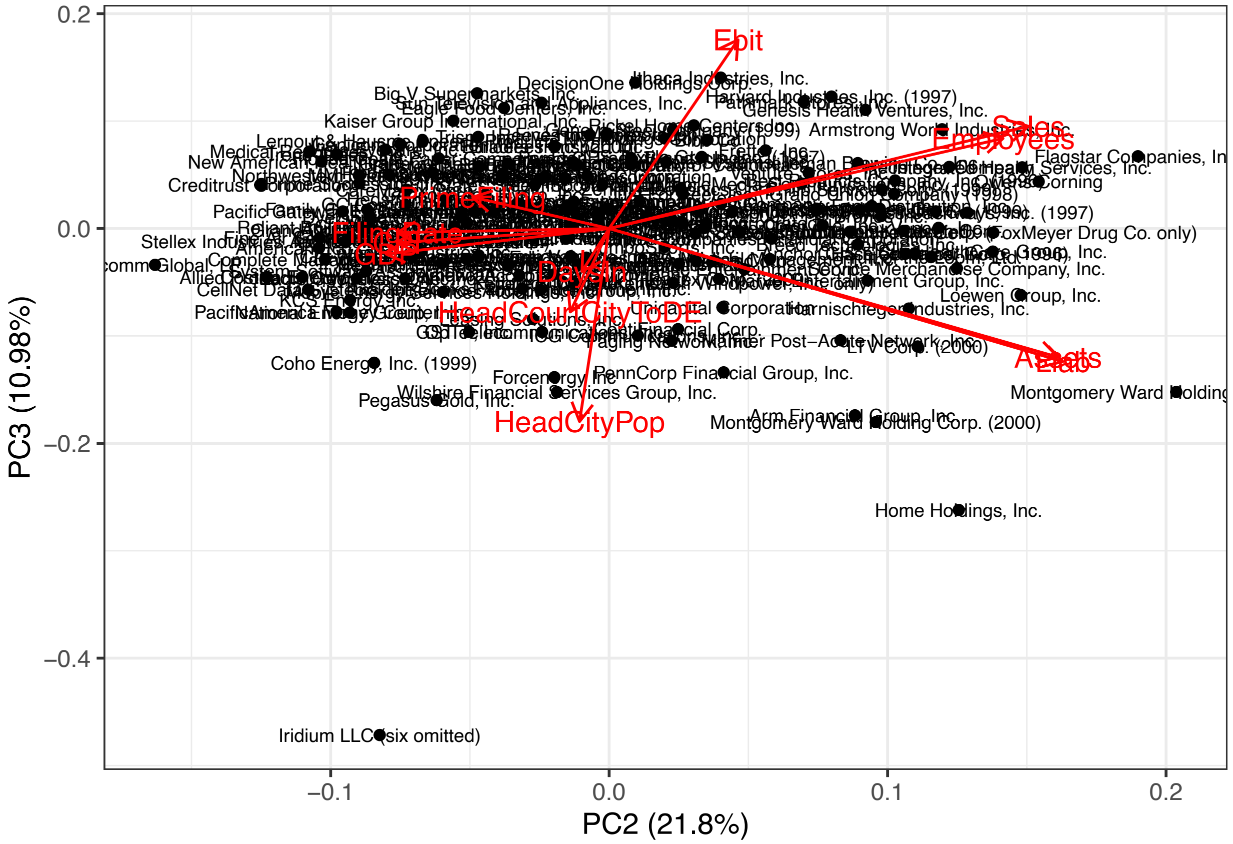


Figure 14: Biplot of PC1 and PC3



One outlier based on PC1 and PC3 is Iridium LLC. The biplot shows that this company has extremely high value of HeadCourtCityToDE meaning that t he company's headquarter city is located far away from the city in which the case was filed. On the other hand, it is characterised by an association with low values of Ebit.

Figure 15: Biplot of PC2 and PC3



Lastly, the biplot for PC2 and PC3 is shown by figure 15. As shown in the graph, Sales and Employees have a strong positive association with each other. Ebit and DaysIn, however, have a strong negative association with one another.

The graph shows that Arm Financial Group and Montgomery Ward Holding Group are similar to one another. On the contrary, Home Holdings and Big V Supermarkets are different to one another.

Similar to the biplot from PC1 and PC3, this graph also shows that Iridium LLC is very different to the rest of the observations. In this instance, the company is characterized by an association with high value of HeadCityPop meaning that the population of the firm's headquarter city is high. In addition, it can also be characterised by an association with low Ebit. This means that amount of earning before income tax at the time the case was filed is low.

## 3.3 Conclusion & Limitation

### 3.3.1 Conclusion

The preliminary data analysis and data wrangling are really important steps before proceeding to data analysis, especially if the data is not in the ideal condition as shown by our raw bankruptcy dataset that contains many missing values and skewed distributions.

Regarding the MDS analysis, the solution of classical MDS seems to have a robust result since its plot produces similar pattern with the Sammon Mapping that considered as more robust solution. Furthermore, by categorizing with several variables i.e. “YearFiled” and “Sales,” we can see clearer similarity/dissimilarity of the observations. Our finding is that more recent time period seems to have larger number of bankruptcy cases, and some portions of companies that file for bankruptcy in the period 1996-2000 tend to have more similarities with companies that filing for bankruptcy during 1991-1995. Grouping by Sales variable, the “very high” sales group seems to have a larger variance of features/attributes that make within and between groups distances are large.

In addition to MDS, another method we implement for this report for dimensionality reduction is PCA. Based on our analysis on the period 1996-2000 subset, three principal components are required to sufficiently explain the dataset. Based on the PC1 and PC2 biplot, we can see that variables `Liab`, `Assets`, `Sales`, `Employees` have strong positive association with each other. On the other hand, variables `Ebit` and `HeadCourtCitytoDE` have strong negative association with each other.

### 3.3.2 Limitations

- Our MDS solution seems to have moderate value of GOF hence probably not really the best solution to minimise Strain.
- Our approach of imputing the median is solely based on the assumption that the missing observations most likely look like the majority of the observations in the variable. Thus, if the assumption is incorrect our imputed data may falsely representing the original/real distribution.
- In this report we are not conducting any particular treatment for any correlated variables when doing the PCA analysis. Thus, our result of PCA analysis may be not optimally summarizing high dimensionality of the dataset in case there are linearly correlated variables. As explained by Dong and McAvoy (1996), PCA method is sometimes inadequate when there is non-linearity detected in the dataset.

# 4 Apendix

There are 5 different approaches that we consider for the MDS analysis:

1. Using original dataset without any treatment and construct the MDS. The plot result is in Figure 16.
2. In terms of missing values, we consider to impute the missing values with the median for each variable that has missing value, then compute the MDS. The reason of doing median imputation because it is easy to implement to get a complete dataset. Also because of some variable with missing values are skewed, so median is a better representation of majority of the observation in variables (Amballa 2020). The MDS plot of this approach is shown in Figure 17.
3. Since `EmplUnion` has a lot of missing values, we consider to remove it from the original dataset, keeping others with NA's then construct the MDS as shown in Figure 18.
4. Two observations (The Texaco.Inc and First Republic Bank Corp) that have extreme attributes[4] are excluded from the dataset. Removing extreme values can help us in observing the visualization of the solution. Figure of the log transformed variable can be seen at Figure 8.
5. Lastly, We also consider to remove outliers (The Texaco.Inc and First Republic Bank Corp) based on the imputed data. The reason to remove them is to focus on middle and smaller company closely, since it hard to interpret them if the outliers still exist. The result is in Figure 19.

After doing above treatments, we take the GOF and minimum eigenvalues of each treatment and compare their values. Here are the result that we obtain.

We can see based on Table 12, the original dataset and by removing `EmplUnion` (both of treatment still has NA's) has the minimum eigenvalues with big negative value ( -396.0989 and -342.8894).

Table 12: Summary of GOF and Eigen

| treatment | GOF1 | GOF2 | MinEig |
|---|---|---|---|
| ori | 0.3742273 | 0.4394227 | -396.0989 |
| impute | 0.4651144 | 0.4651144 | 0.0000 |
| no_union | 0.4242896 | 0.4766834 | -342.8894 |
| log | 0.4894409 | 0.4894409 | 0.0000 |
| no_out | 0.4675692 | 0.4675692 | 0.0000 |

We also noticed that both the GOF1 is not equal to GOF2. With the goodness of fit is different and the eigenvalues is negative, means that the classical MDS may not minimise Strain. It minimises a slightly different function of the distances. We also can conclude that the distance input is not Euclidian.

Another thing happened for impute, log, and no outlier treatment. Each of their GOF1 are similar with the second. Impute and no outlier are quite close. They have extremely small eigenvalues which can be assumed indistinguishable from zero. Based on the result we can say that the Classical MDS minimised the Strain. This is to be expected since the input distance is Euclidean.

After we measure the GOF, we noticed that the Log approach has the biggest GOF than others. Hence, we choose this scenario as our main approach.

[4] The Texaco.Inc has the biggest asset in millions of dollars recorded, while First Republic Bank Corp has the biggest liability which is total amount of money owed in million dollars

Figure 16: Plot of original dataset with Outliers of Texaco Inc. and First Republic Bank Corp

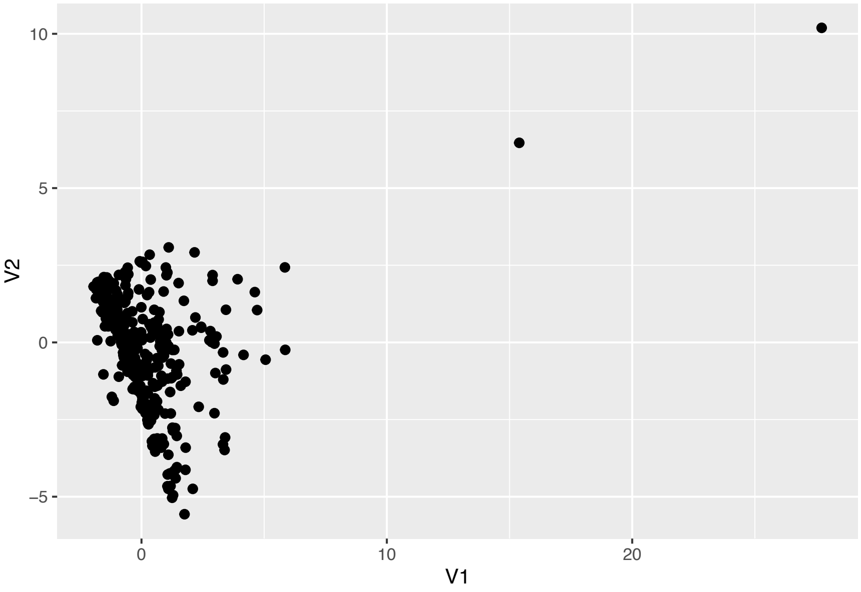


Figure 17: MDS Plot with Imputing Median

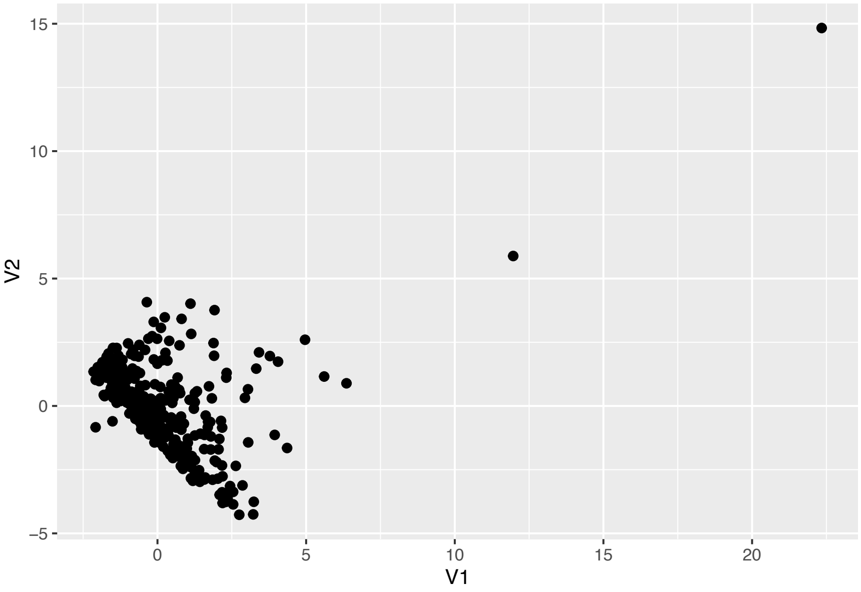


Figure 18: MDS Plot with Removed Employed Union

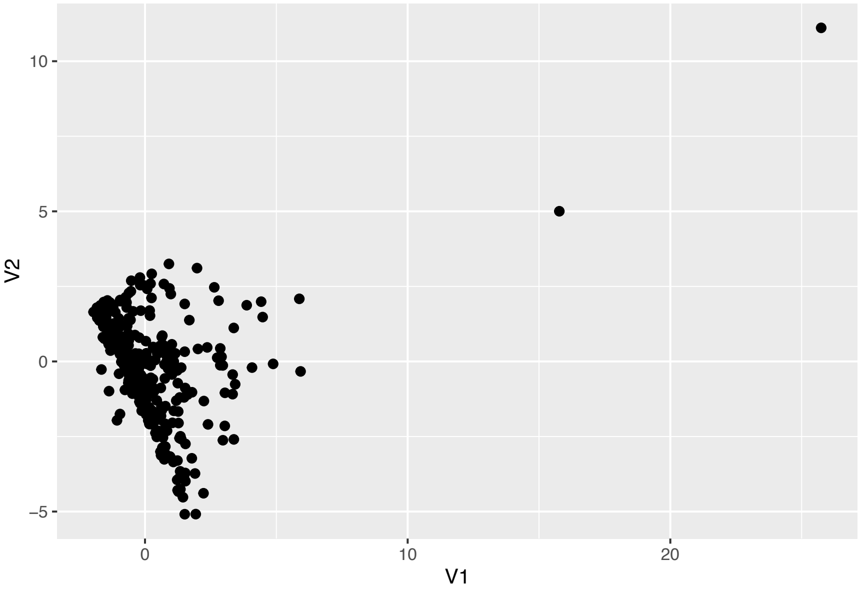
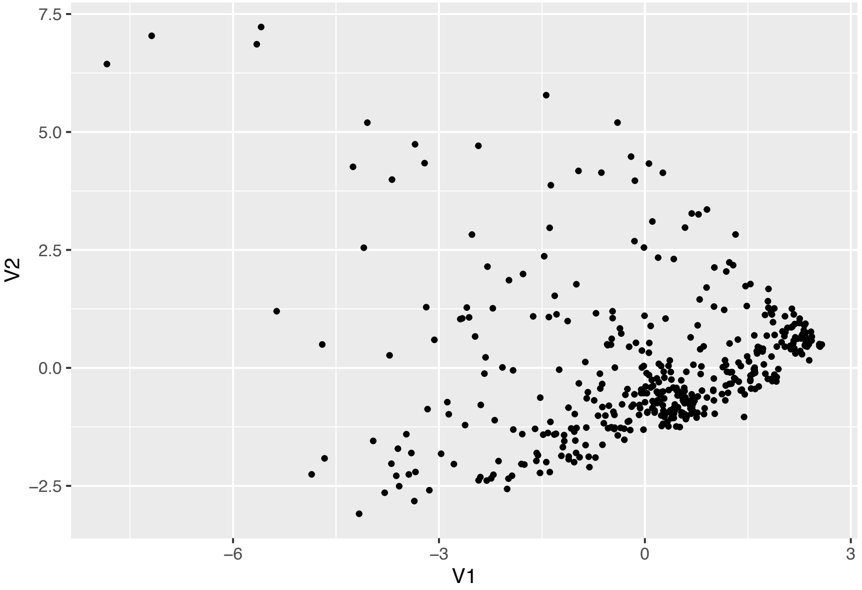


Figure 19: MDS Plot with Removed Outliers



# References

Amballa, Arun. 2020. "Feature Engineering Part-1 Mean/ Median Imputation. | by Arun Amballa | Analytics Vidhya | Medium." https://medium.com/analytics-vidhya/feature-engineering-part-1-mean-median-imputation-761043b95379.

Dong, Dong, and Thomas J McAvoy. 1996. "Nonlinear Principal Component Analysis—Based on Principal Curves and Neural Networks." *Computers & Chemical Engineering* 20 (1): 65–78.

Hlavac, Marek. 2018. *Stargazer: Well-Formatted Regression and Summary Statistics Tables*. Bratislava, Slovakia: Central European Labour Studies Institute (CELSI). https://CRAN.R-project.org/package=stargazer.

Horikoshi, Masaaki, and Yuan Tang. 2018. *Ggfortify: Data Visualization Tools for Statistical Analysis Results*. https://CRAN.R-project.org/package=ggfortify.

Jolliffe, Ian T, and Jorge Cadima. 2016. "Principal Component Analysis: A Review and Recent Developments." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374 (2065): 20150202.

Kassambara, Alboukadel, and Fabian Mundt. 2020. *Factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. https://CRAN.R-project.org/package=factoextra.

Pedersen, Thomas Lin. 2020. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

Richard A. Becker, Original S code by, Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka, and Alex Deckmyn. 2018. *Maps: Draw Geographical Maps*. https://CRAN.R-project.org/package=maps.

Robinson, David, Alex Hayes, and Simon Couch. 2021. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Schloerke, Barret, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. 2021. *GGally: Extension to 'Ggplot2'*. https://CRAN.R-project.org/package=GGally.

UCLA. 2019. "How do I standardize variables in Stata? | Stata FAQ." https://stats.idre.ucla.edu/stata/faq/how-do-i-standardize-variables-in-stata/.

"UCLA-LoPucki Bankruptcy Research Database." 2020. https://lopucki.law.ucla.edu/glossary.php%20https://lopucki.law.ucla.edu/.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2019. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.