Assignment 2 Report

Aarathy Babu, Dewi Amaliah, Priya Dingorkar, Rahul Bharadwaj

1 Introduction

The UCLA-LoPucki Bankruptcy Research Database (BRD) is a UCLA School of Law data gathering, data linking, and data distribution initiative. The goal of the BRD is to encourage bankruptcy research by making bankruptcy data available to academic investigators worldwide. All of the data was gathered when the companies declared bankruptcy. 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.

2 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.

3 Data Description

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. Let's take a closer look at these variables and what they mean. Let's go further into the dataset to see if we can find any examples of data cleaning or wrangling.

The dataset has 436 observations and 7 variables 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

Let us further examine the bankruptcy statistics. We will undertake some preliminary data analysis. We will also go through the numerous approaches used for this high-dimensional data in detail later.

4 Preliminary Data Analysis

Before carrying out further analysis of the data, let us conduct some preliminary data analysis. From the summary shown below, we can see that the data is a high dimensional dataset with 21 variables, out of which 6 are character variables and 15 are 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. It was found that there are observations where the `EmplUnion` values are more than `Employees` which was removed from the data and that certain companies have 1 Employee 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.

```
## # A tibble: 3 x 3
##   Name                                            Employees EmplUnion
##   <chr>                                               <int>     <int>
## 1 Residential Resources Mortgage Investments Corp.        1         1
## 2 Mortgage & Realty Trust (1990)                          1         1
## 3 Promus Companies Inc. (Harrahs Jazz Co. only)           1      3000
```

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

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

The missing values in the data has been visualized as shown in 1. Throughout our strategy, we have tried to retain the data as much as possible while maintaining high data quality and credibility.

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 existance , 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 as the variable Employees explains similar aspect.

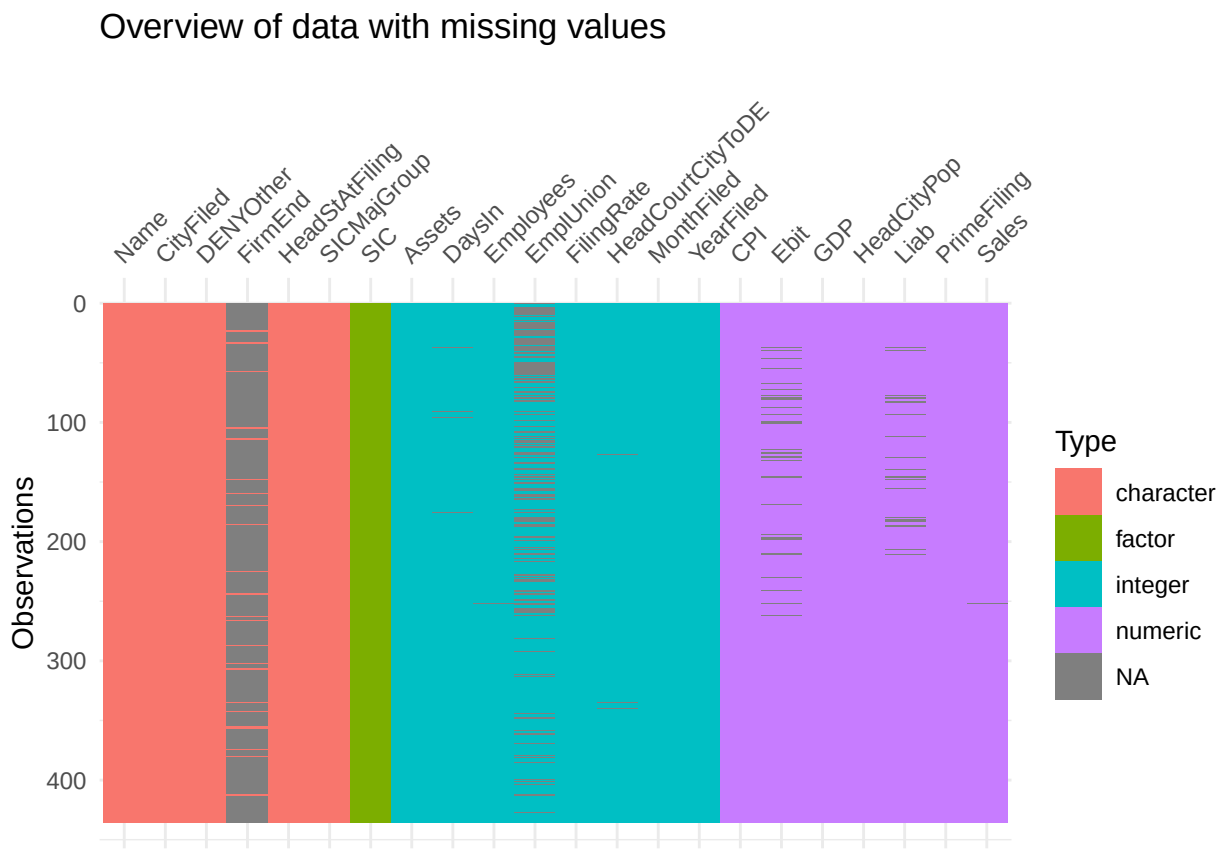


Figure 1: Overview of missing values in the data

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.
```

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

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

```
##    31.0   248.0   509.0   635.5   867.5  3730.0
```

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 publicaly available data on headquarter address and the `CityFiled`, the distances between these cities were found and imputed into the data accordingly. See Appendix for more information.

```
##    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 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.

```
## # A tibble: 1 x 4
##   Name              Sales Employees  Ebit
##   <chr>             <dbl>     <int> <dbl>
## 1 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

The data was then checked for outliers, even though we haven't found suspicious outliers in majority of the variables (see Appendix for more information), 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.

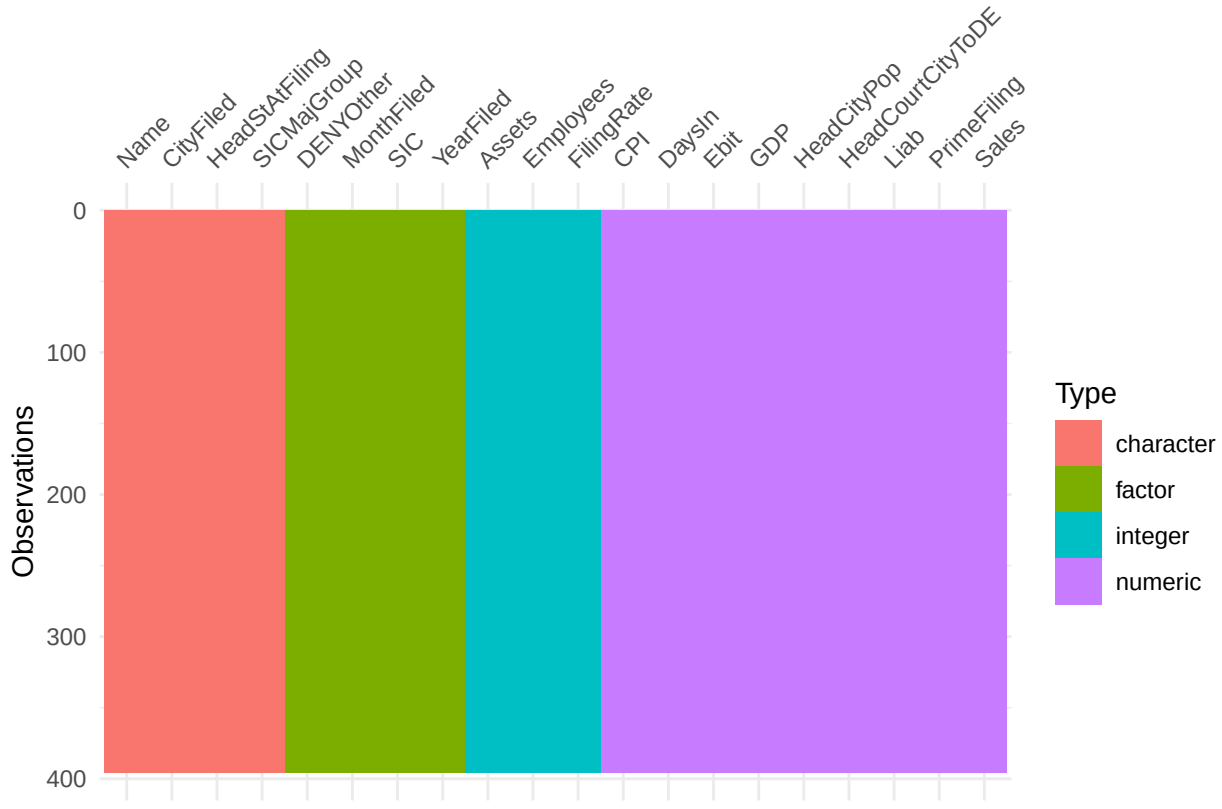


Figure 2: Overview of Cleaned Data

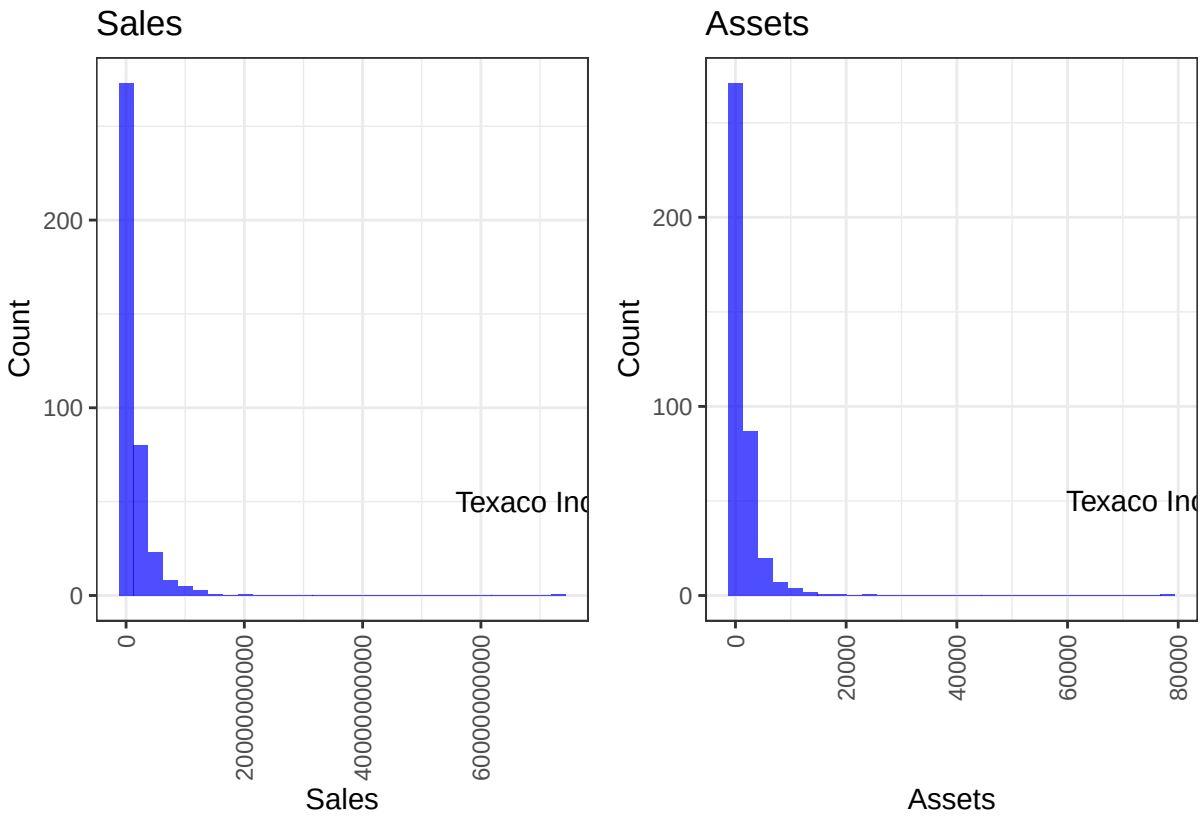


Figure 3: Presence of Outliers in Sales and Assests

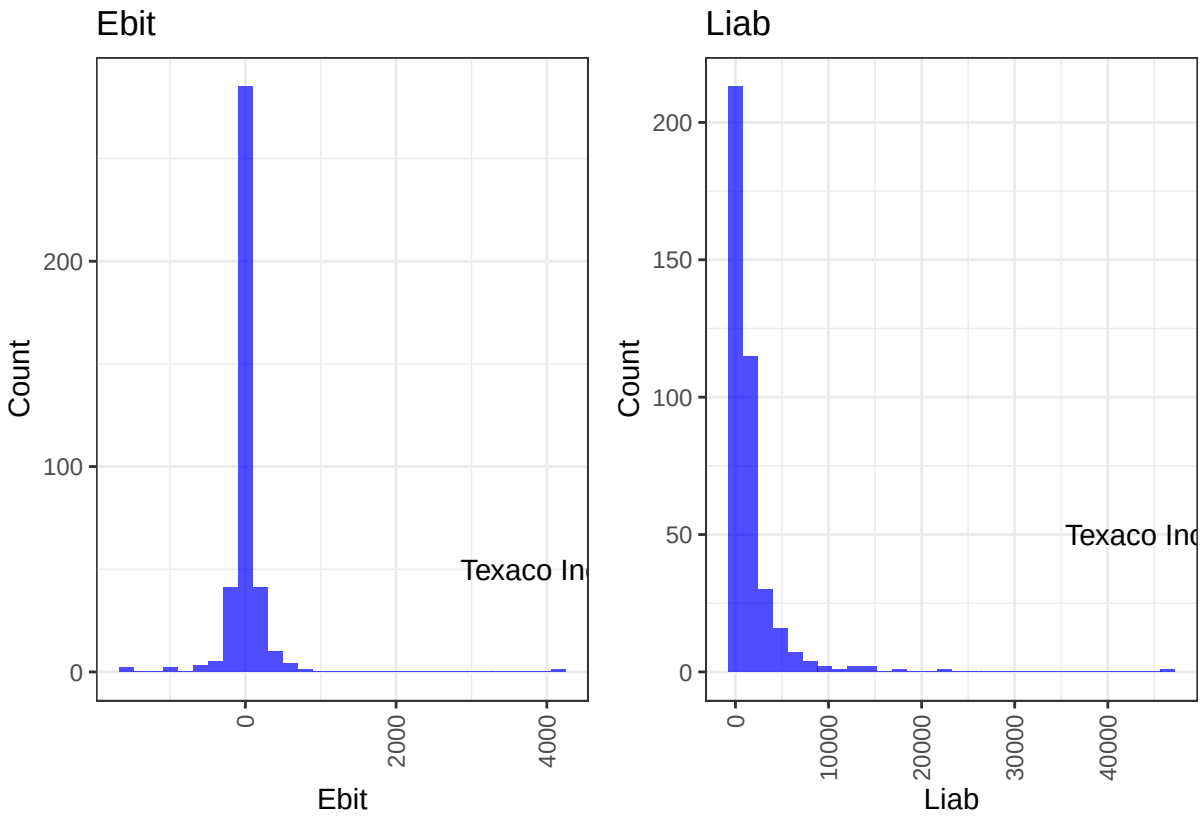
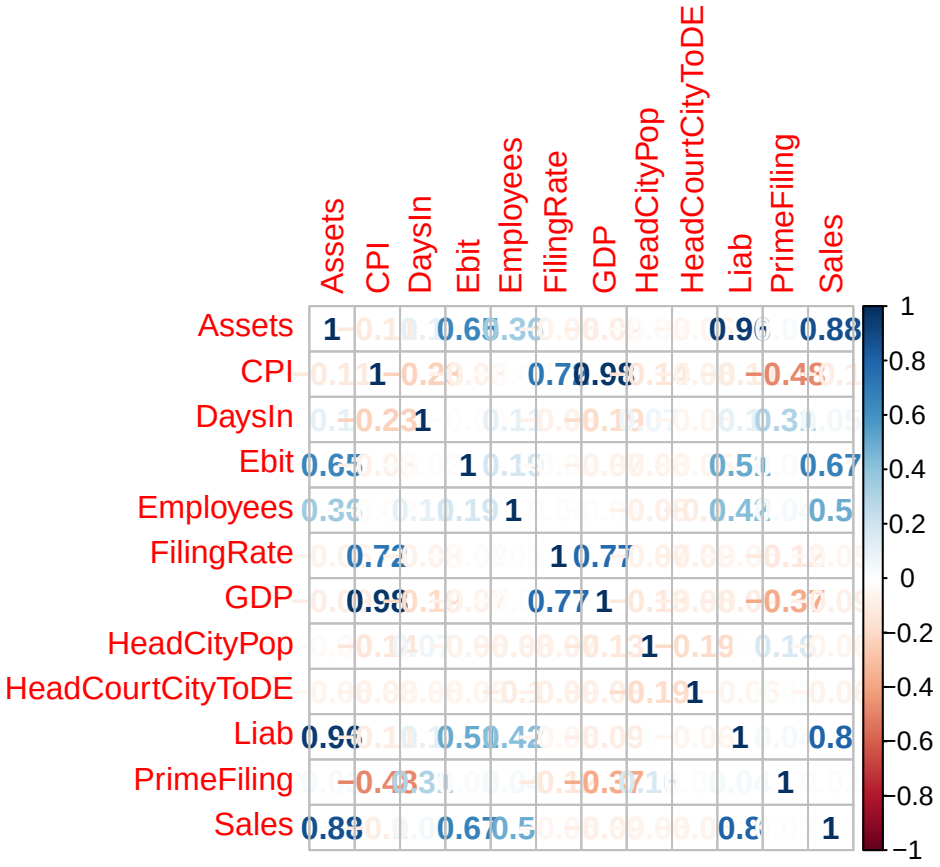


Figure 4: Presence of Outliers in Ebit and Liab



5 Multidimensional Scaling (MDS)

MDS is a statistical method to represent multidimensional data into lower-dimensional (2D) data. The `bankruptcy` data has 21 variables which can be considered as data with high dimensions. Thus, MDS is relevant to represent this 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.

5.1 Classical MDS

Figure 5 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 5 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 6 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

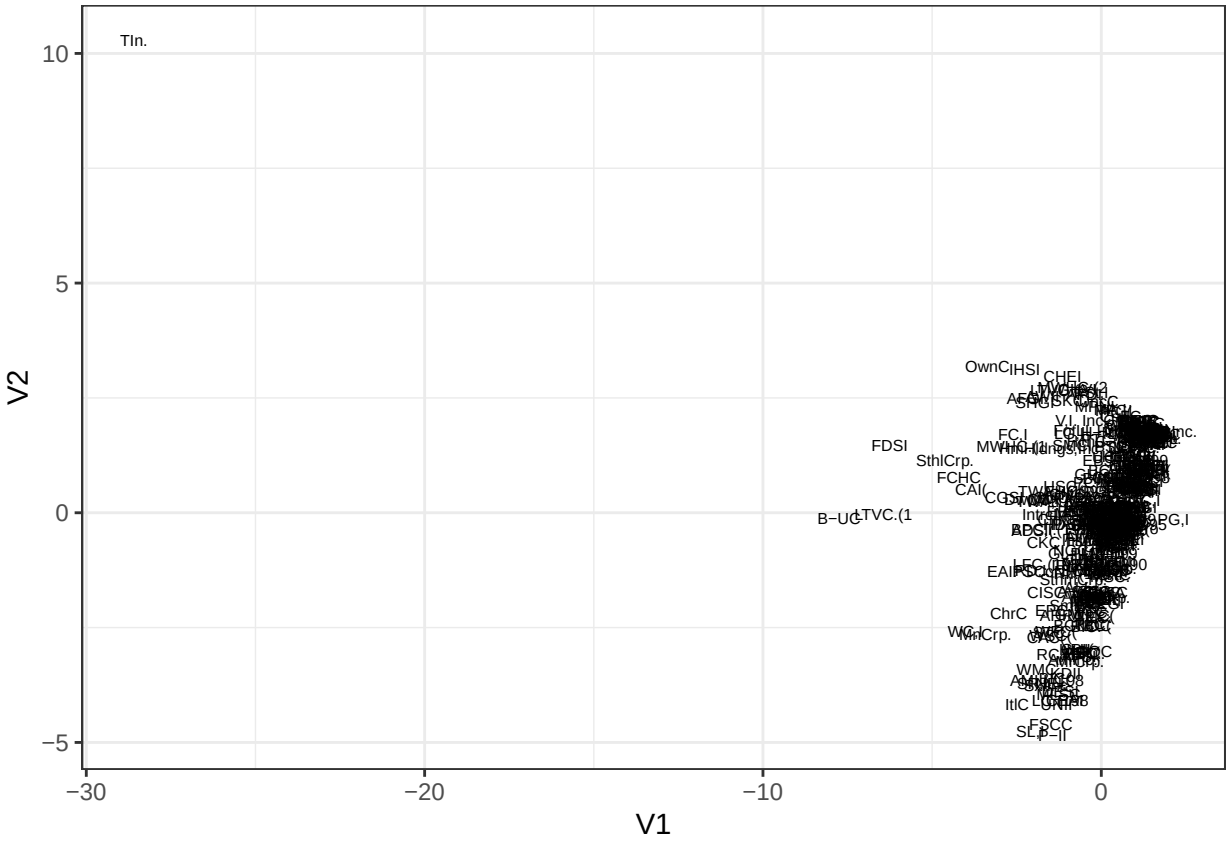


Figure 5: Classical MDS solution for bankruptcy data. The x and y-axis represent the new variables as the result of MDS. Some outliers observed in thedata

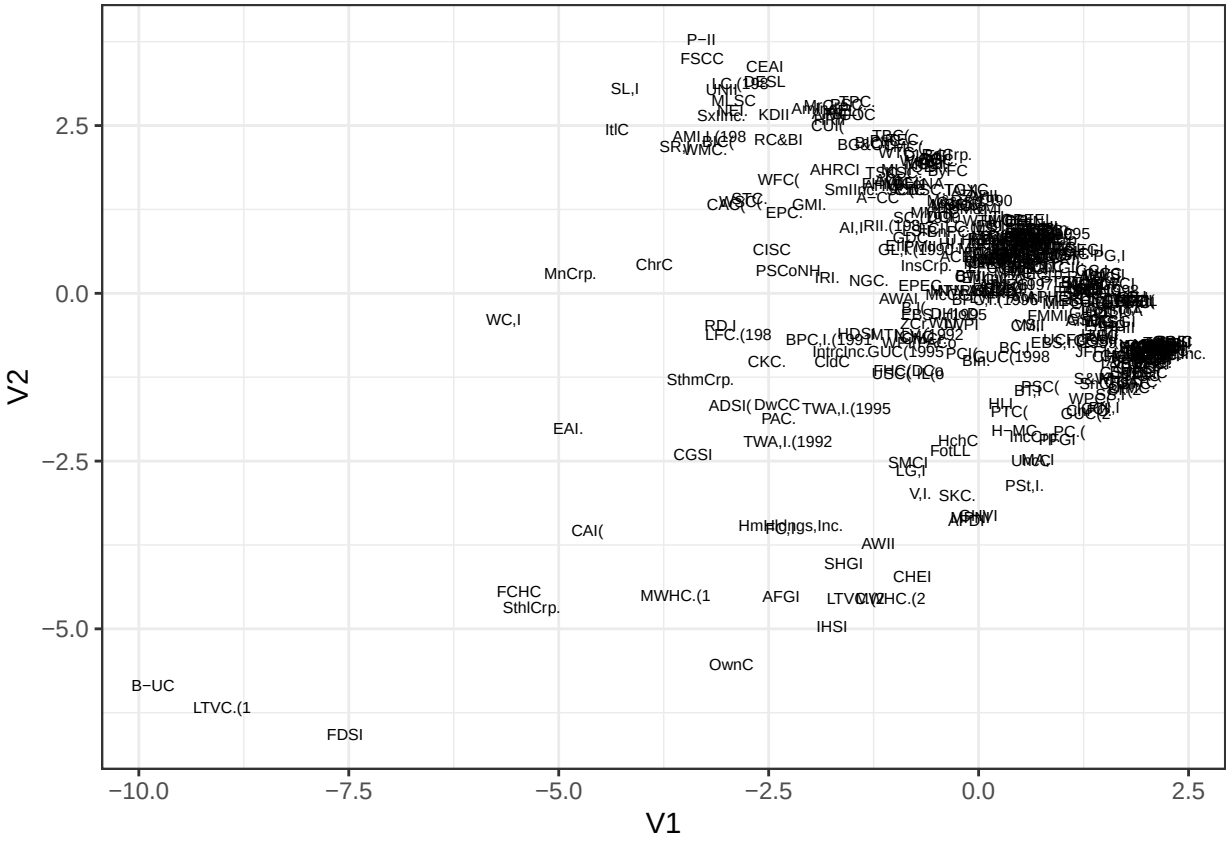


Figure 6: Classical MDS solution for bankruptcy data after excluding Texaco Inc. The x and y-axis represent the new variables as the result of MDS. We get a clearer visualisation compared to the previous MDS result

means that these firms have different profile.

5.2 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:

```
##  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).

5.3 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.

5.4 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.

The classical MDS solution plotted by year as shown in Figure 7 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.

6 Principal Component Analysis (PCA)

- Now that we've seen how to input this high-dimensional data into Multidimensional scaling (MDS) to obtain a low (typically 2) dimensional representation. Let us now perform a Principal Component Analysis (PCA), which is a dimensional-reduction method that is frequently used to reduce the dimensional of large data sets by transforming a large set of variables into a smaller one that still contains the majority of the information in the large set.
- On performing a PCA on the clean dataset during the data investigation process we found that, just like MDS sees Texaco Inc. which comes under Petroleum SIC and Refining And Related Industries as the outlier we get similar results when we feed the complete dataset to carry out PCA.
- Keeping in mind, the word limit we have constrained to show only the PCA using the data without any outliers. Note the complete analysis of PCA for the data can be found in the Appendix section at the end of this report.
- Let's carry out PCA on our bankruptcy data. Lets investigate if our data is a good fit for PCA. Let's us further investigate how the variables in our data are correlated and how many PC's explain the variation of our data.

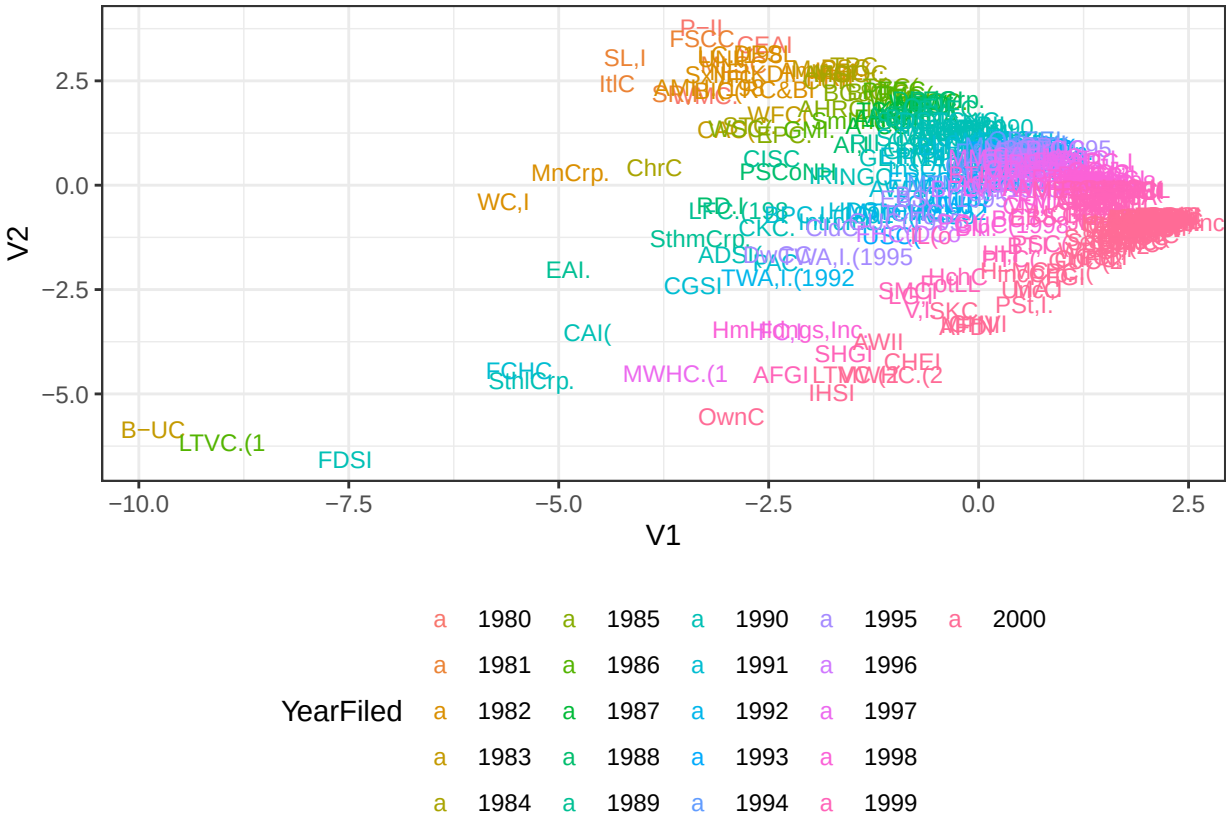


Figure 7: Classical MDS solution plotted by year when the bankruptcy filed.

- Before we apply this principle to our data, it is very important that we standardized our variables as this ensures that results are not sensitive to the units of measurement. Thus giving us more accurate analysis.

6.1 PCA without outlier

```
## 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
```

- Using the `summary` function we can infer the following:
  - 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
  - Using kaisers rule, choose those PC's whose variance and standard deviation greater than 1, we will choose 5 PC's that explains the most variation in our data.
- Let us now use the scree plot to find the total number of PC's that best explain our data.

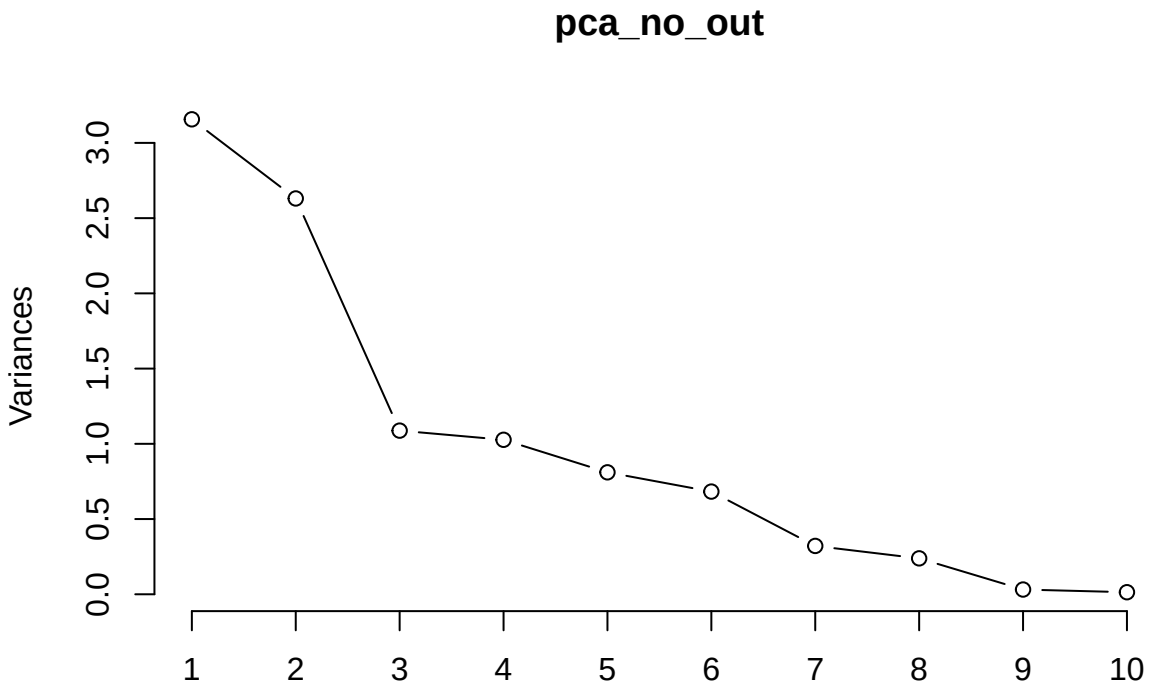


Figure 8: Screeplot

- Interestingly the scree plot and the kaiser's rule do not agree with each other. But since the scree plot is a more accurate measure in helping us confirm that most of the variations is captured by the first three principal components from our dataset.
- As we can see from the summary of our PCA that our third PC explains approx 11% of variation in our data but we cannot visualize this with a biplot. Nevertheless we still plot the biplot for our first two PC's as they explain around 50.00% of variation in our data.

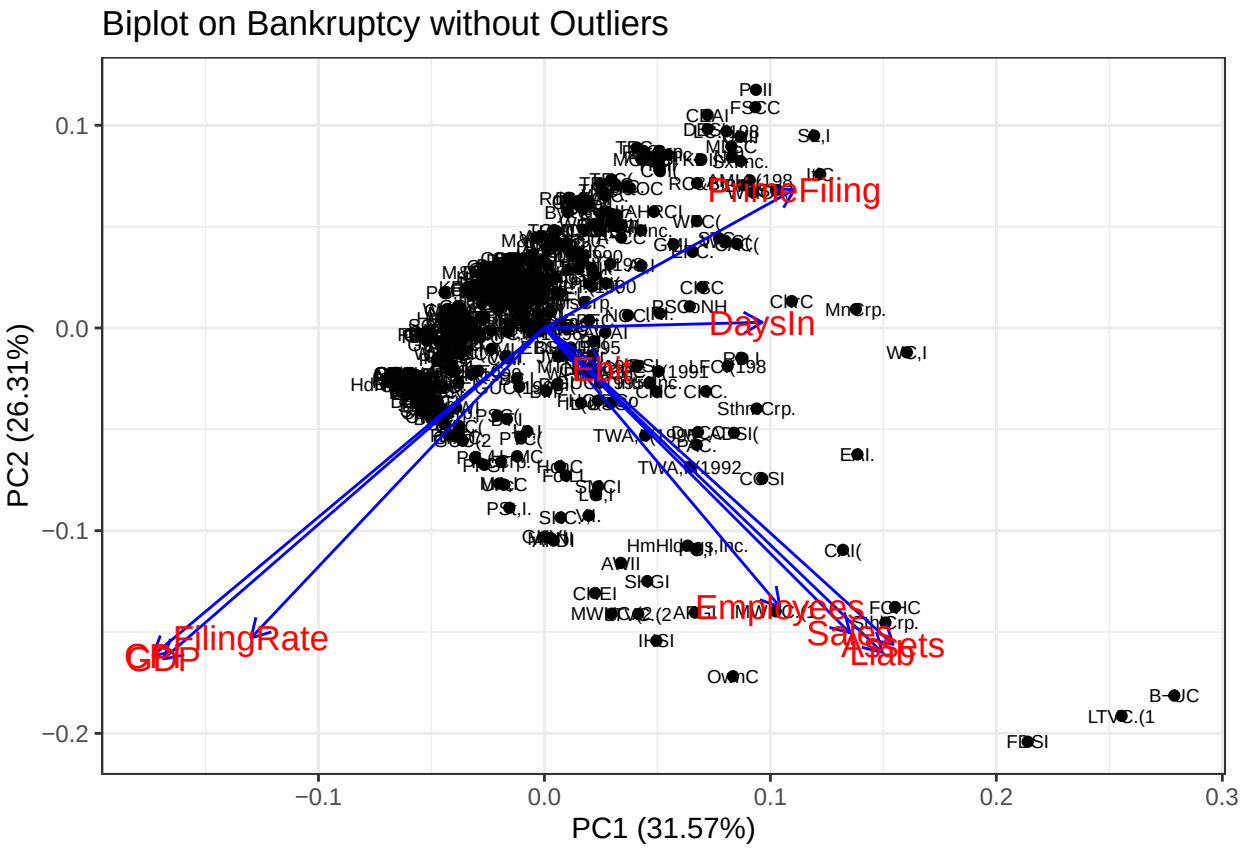


Figure 9: PCA Biplot without Outliers

- Looking at the figure 9 we can infer the following:
  - We can see the various spread of the different companies while filing for bankruptcy. We can see how these companies are been 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.
  - We also know that variables at an angle of 90° indicates no correlation between them, In our data we can infer that Employees with GDP/CPI shows an angle of almost 90° thus showing no correlation between these variables.
  - We also infer that variables with 180° angle indicates negative correlation. In our case we can say that PrimeFiling is negative correlated with GDP, CPI and FilingRate making an almost 180° angle.

– Similarly,variables with angle close to 0° indicates positive correlation. In our dataset we can see that CPI and GDP are highly positively correlated. Assest, Liab, Ebit and Sales are all also positively correlated as the angle between all them is nearly zero.

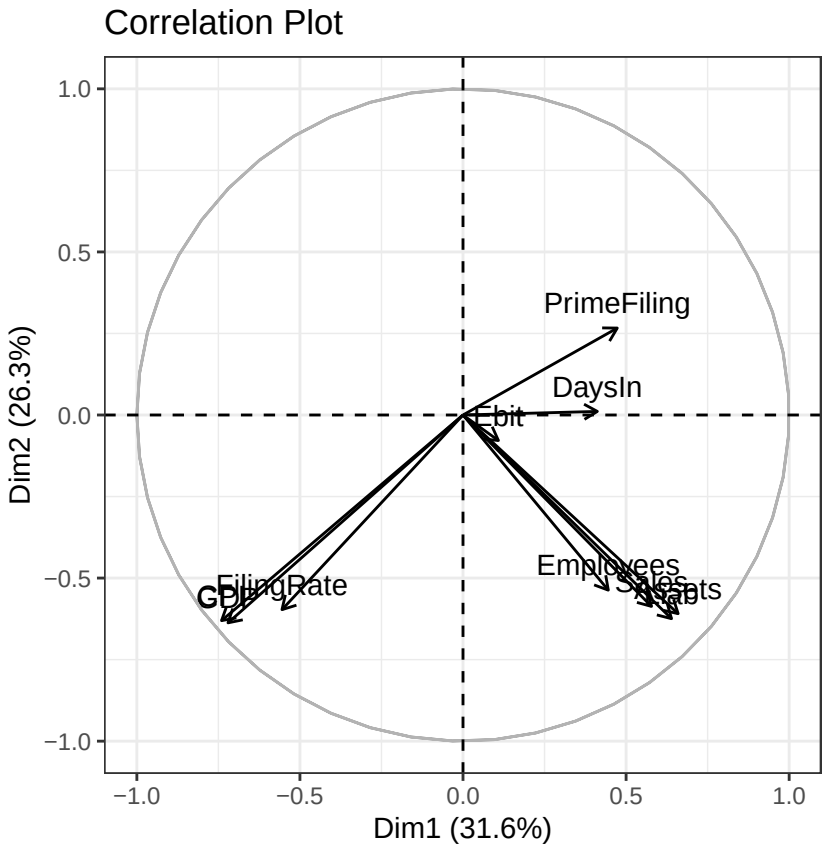


Figure 10: PCA Correlation Plot

- We can also refer to figure 10 for more clear visuals of the angles between variables making it clearly for the audience to distinguish between the variables that are positively, negatively or not correlated at all.

7 Cluster Analysis

8 Limitations

- We only use numerical data in the analysis due to the complexity of incorporating non-numeric data. However, we tried to also display that categorical variable when visualising the MDS result.
- One of the most important limitation in our PCA analysis is that, as seen in the screeplot at figure we inferred that our data was most explained by the the first three PCs. But 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.

Name	Sales
County Seat, Inc.	NA

Name	Employees
County Seat, Inc.	NA

9 Conclusions

10 Appendix

10.1 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`
- Checking Outliers

10.2 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

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

Figure 13 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 14 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 14 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.

10.3 PCA

- PCA performed on the complete data including outliers

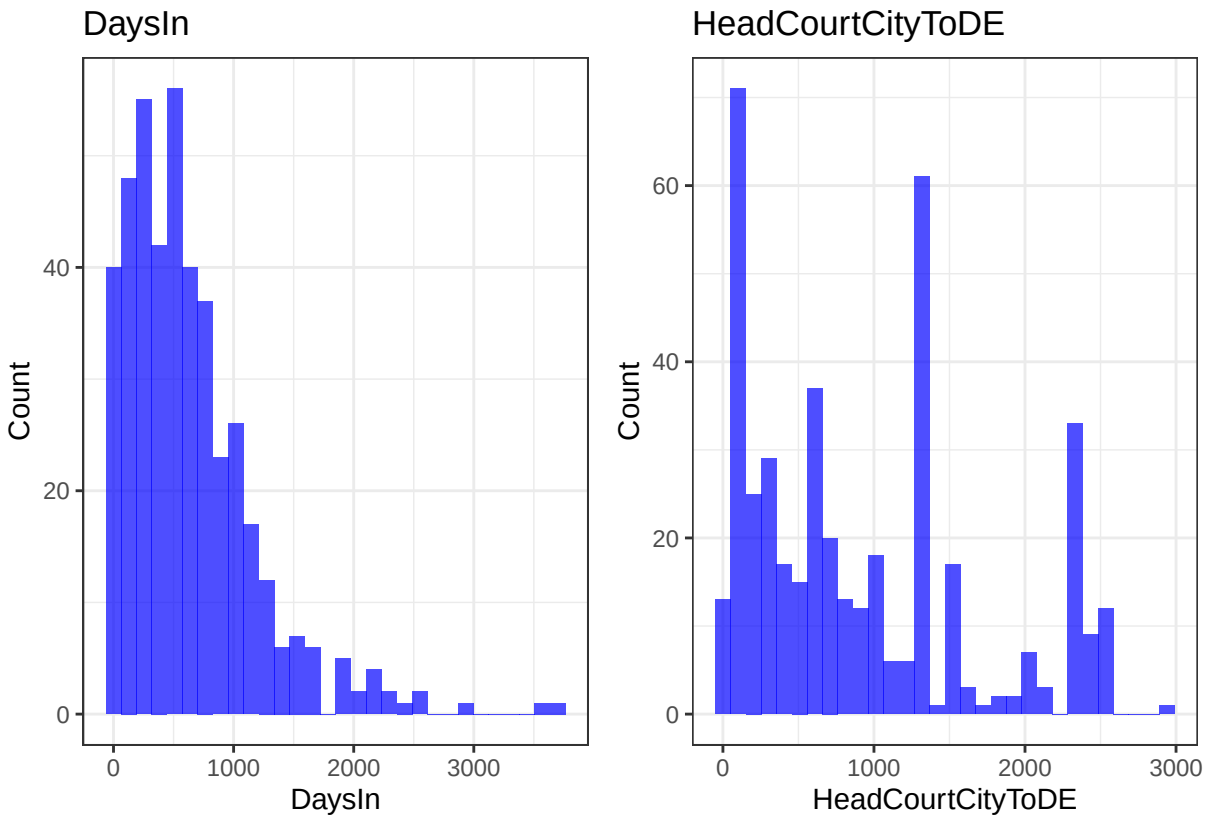


Figure 11: The figure indicates that there isnt any outliers in the variables DaysIn and HeadCourtCitytoDE

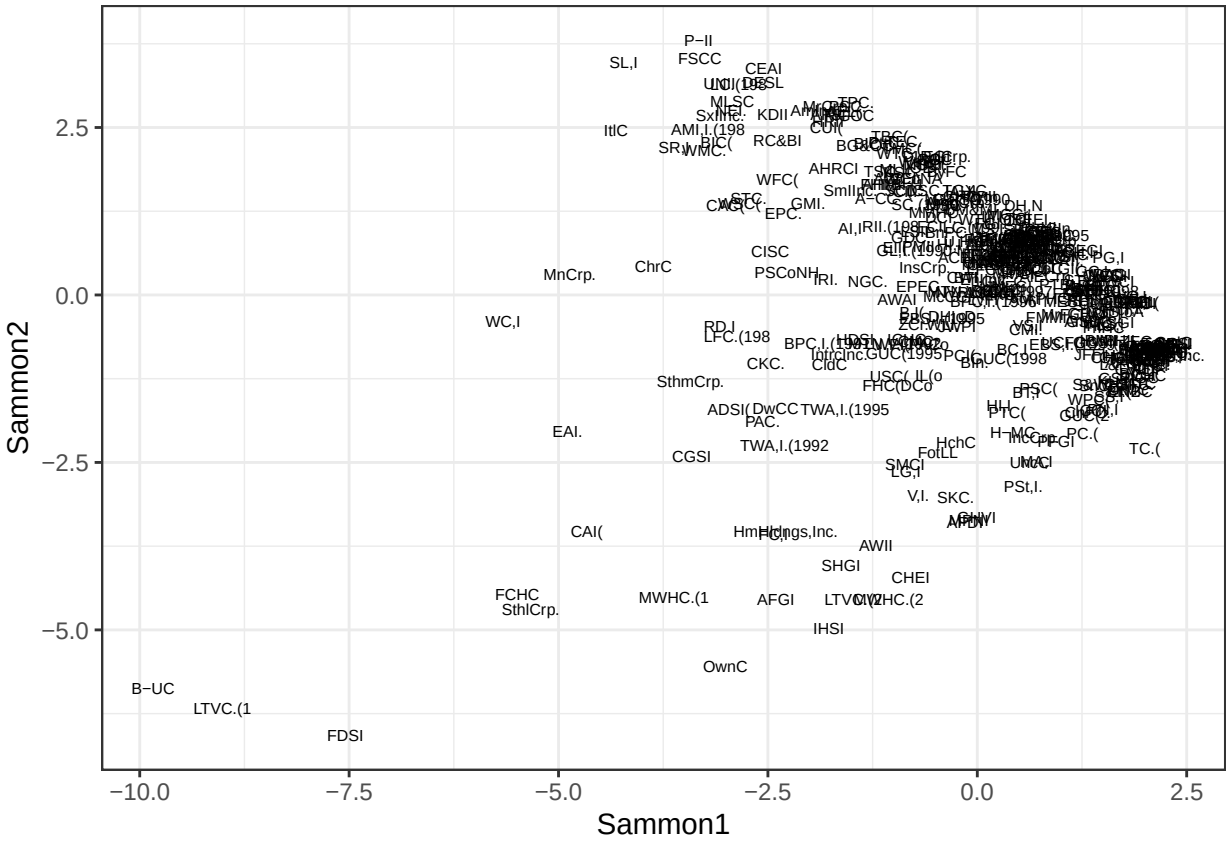


Figure 12: MDS solution using Sammon mapping

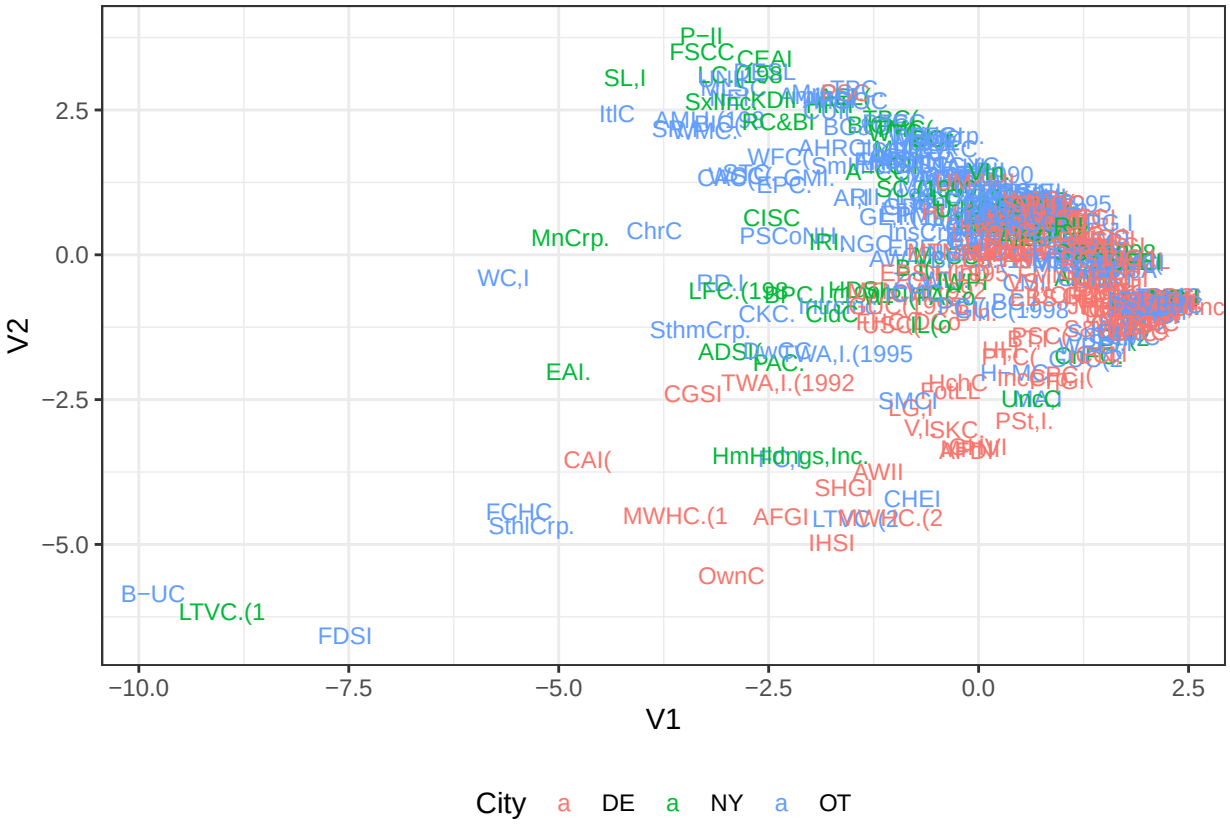


Figure 13: Classical MDS solution plotted by city where the bankruptcy filed.

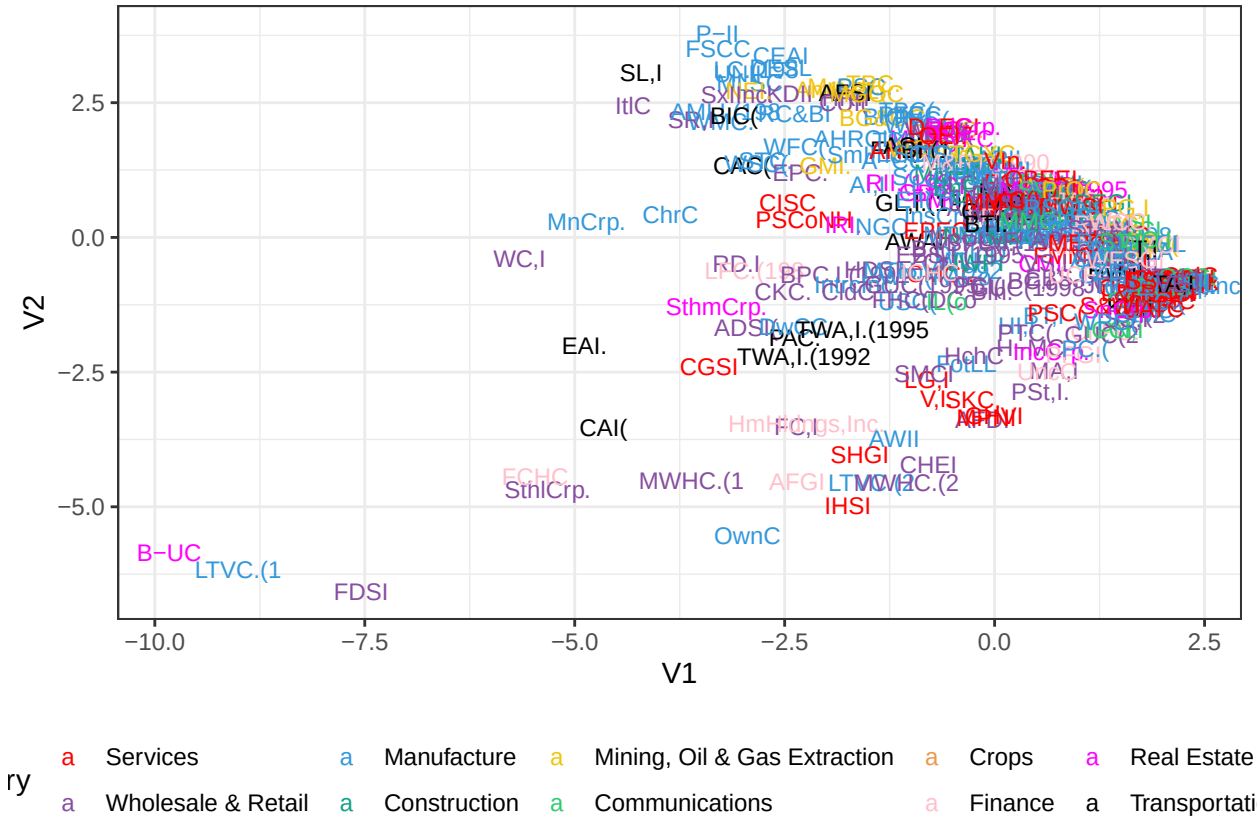


Figure 14: Classical MDS solution plotted by industry.

```
## 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
## 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.

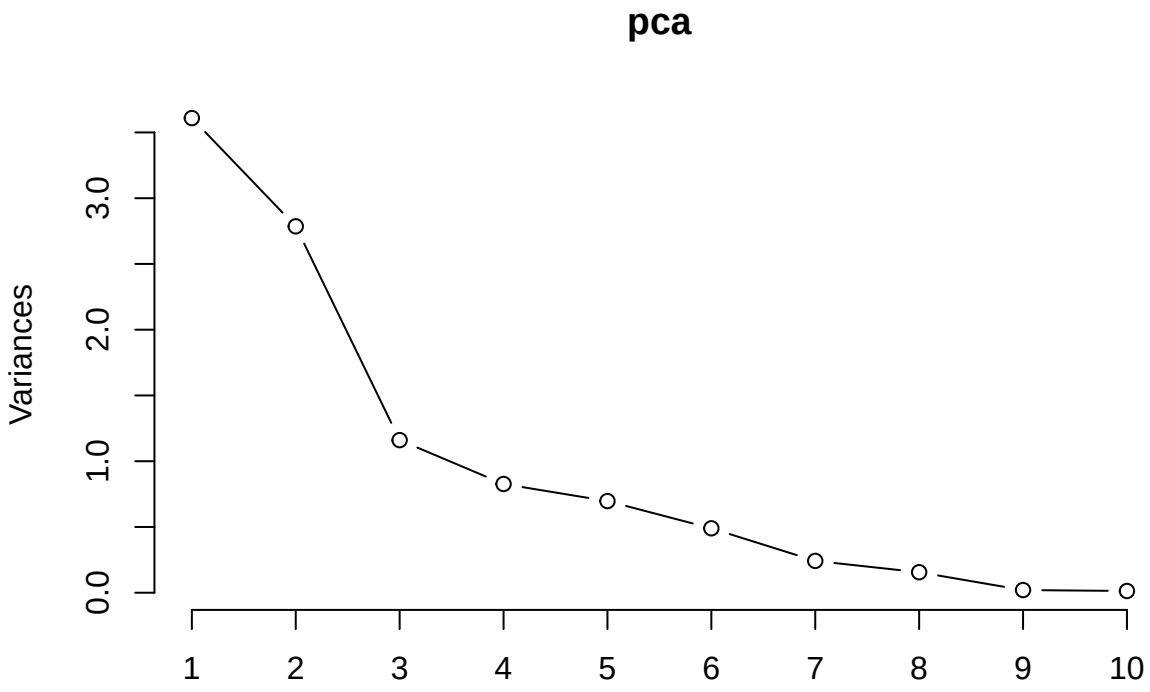


Figure 15: Scree Plot

- 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).

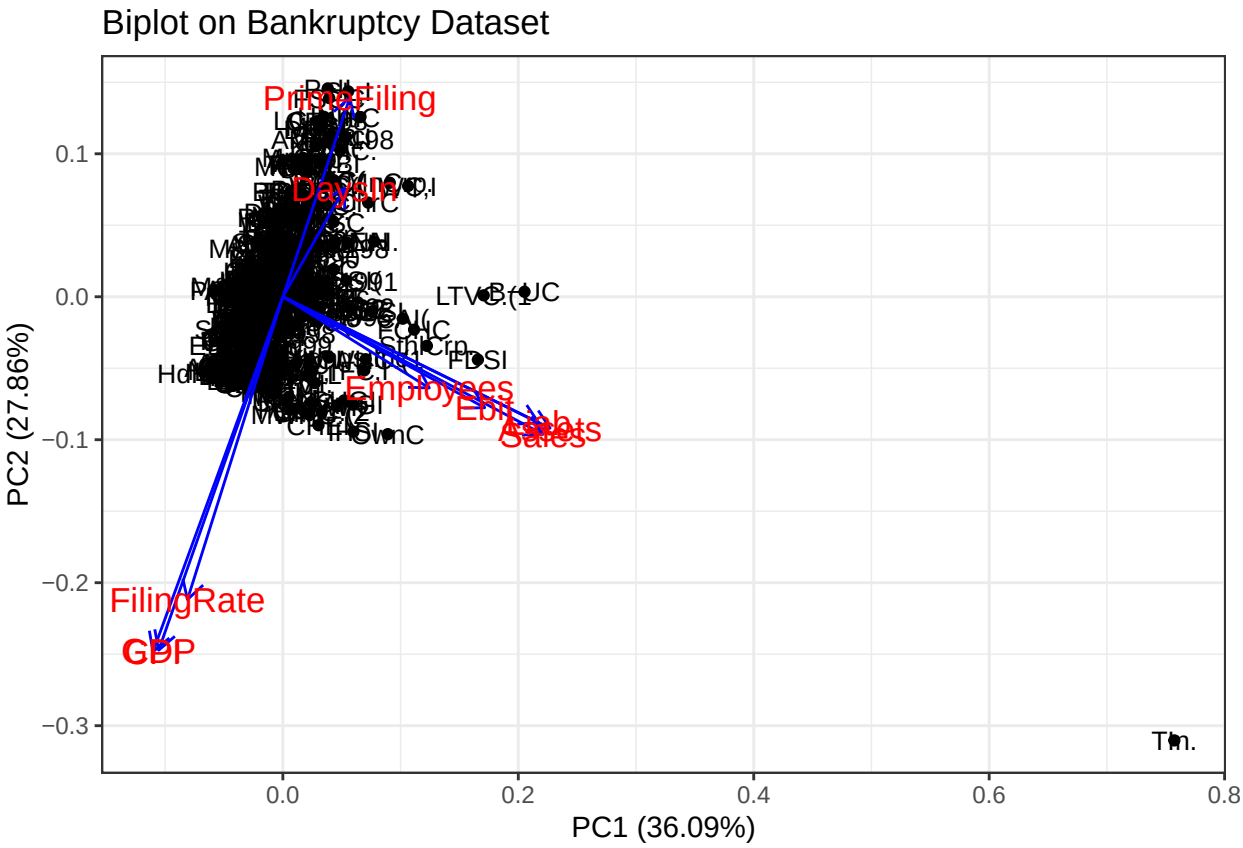


Figure 16: PCA Biplot