Assignment 2 Report

Aarathy Babu, Dewi Amaliah, Priya Dingorkar, Rahul Bharadwaj

1 Introduction

2 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.

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.

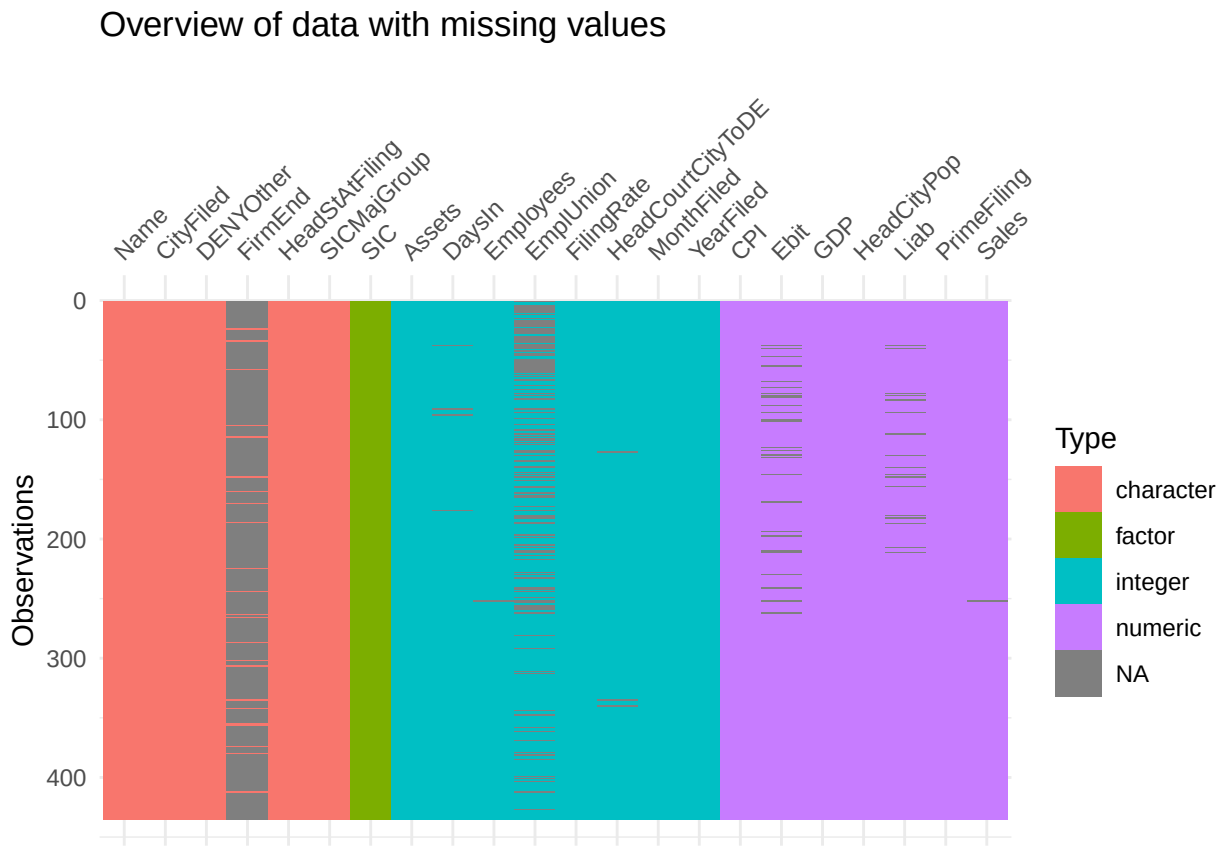


Figure 1: Overview of missing values in the data

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

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

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

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.

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

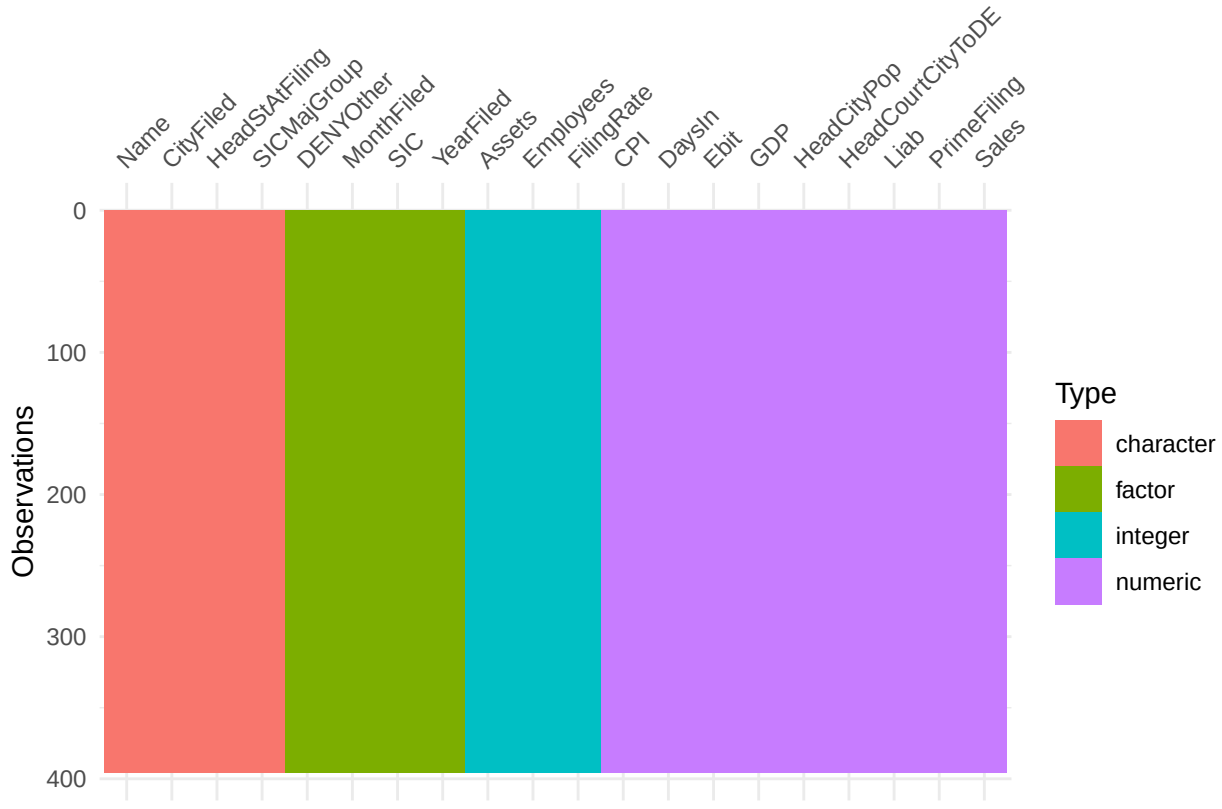


Figure 2: Overview of Cleaned Data

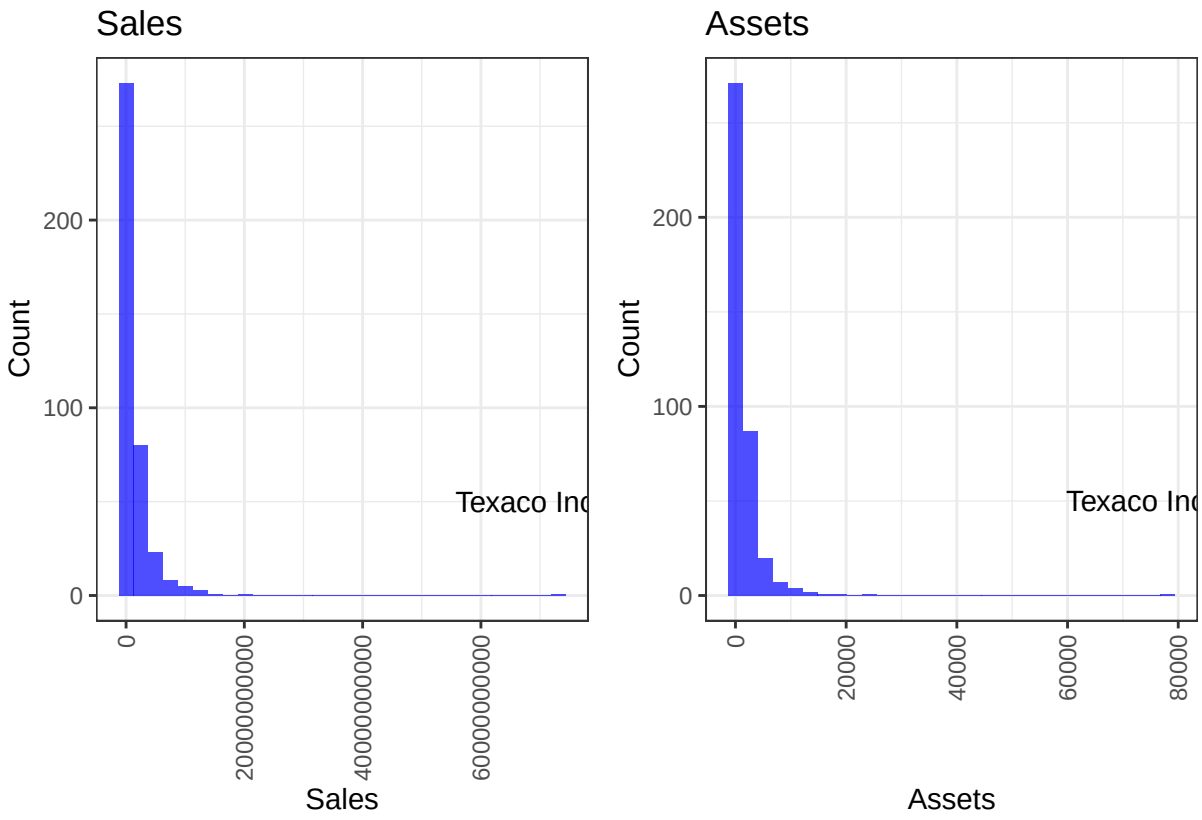


Figure 3: Presence of Outliers in Sales and Assests

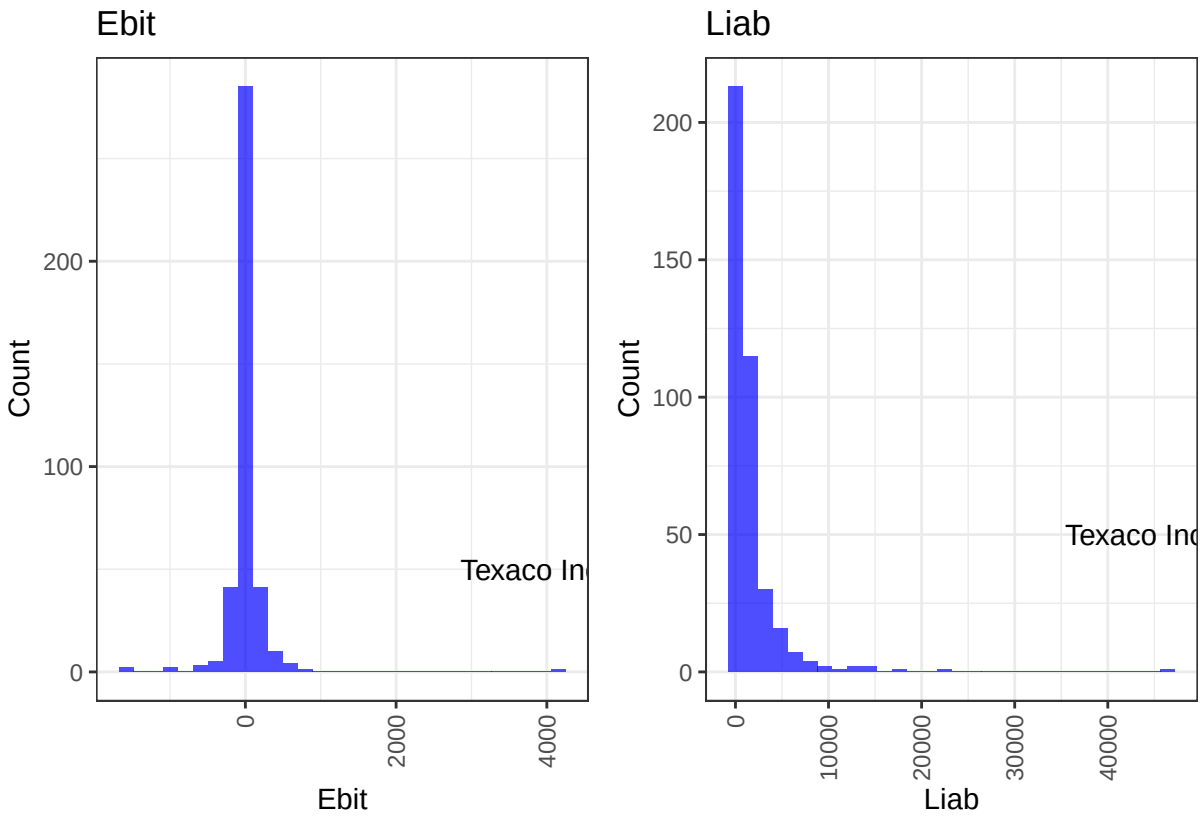
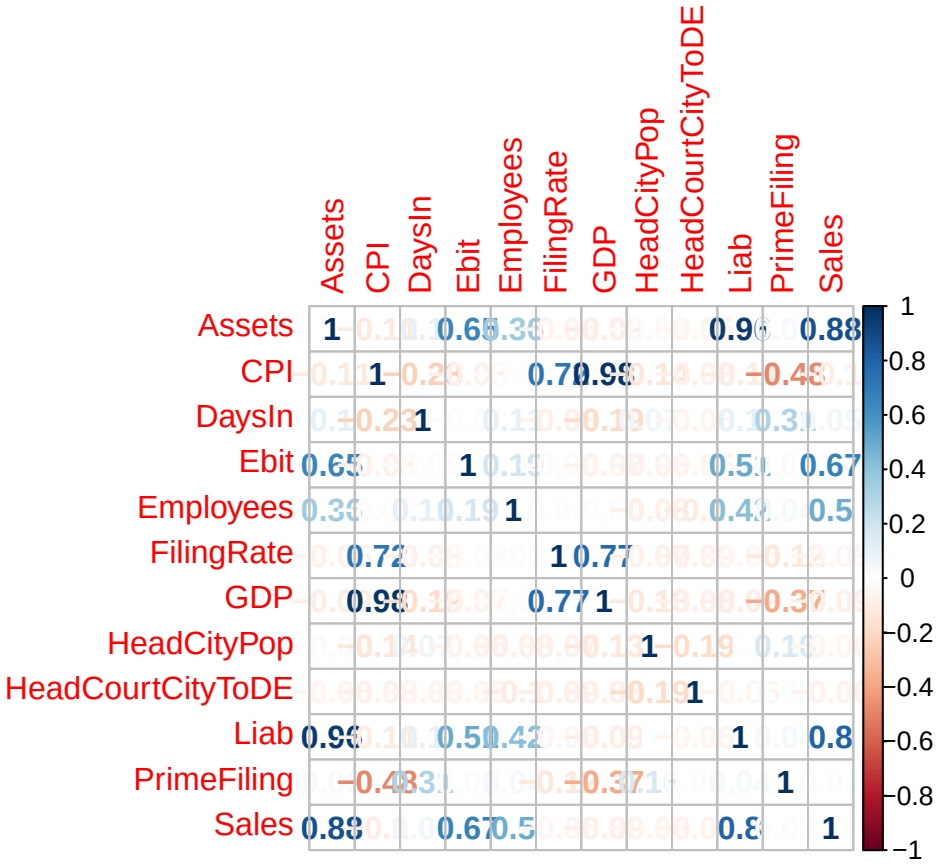


Figure 4: Presence of Outliers in Ebit and Liab



3 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.

3.1 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

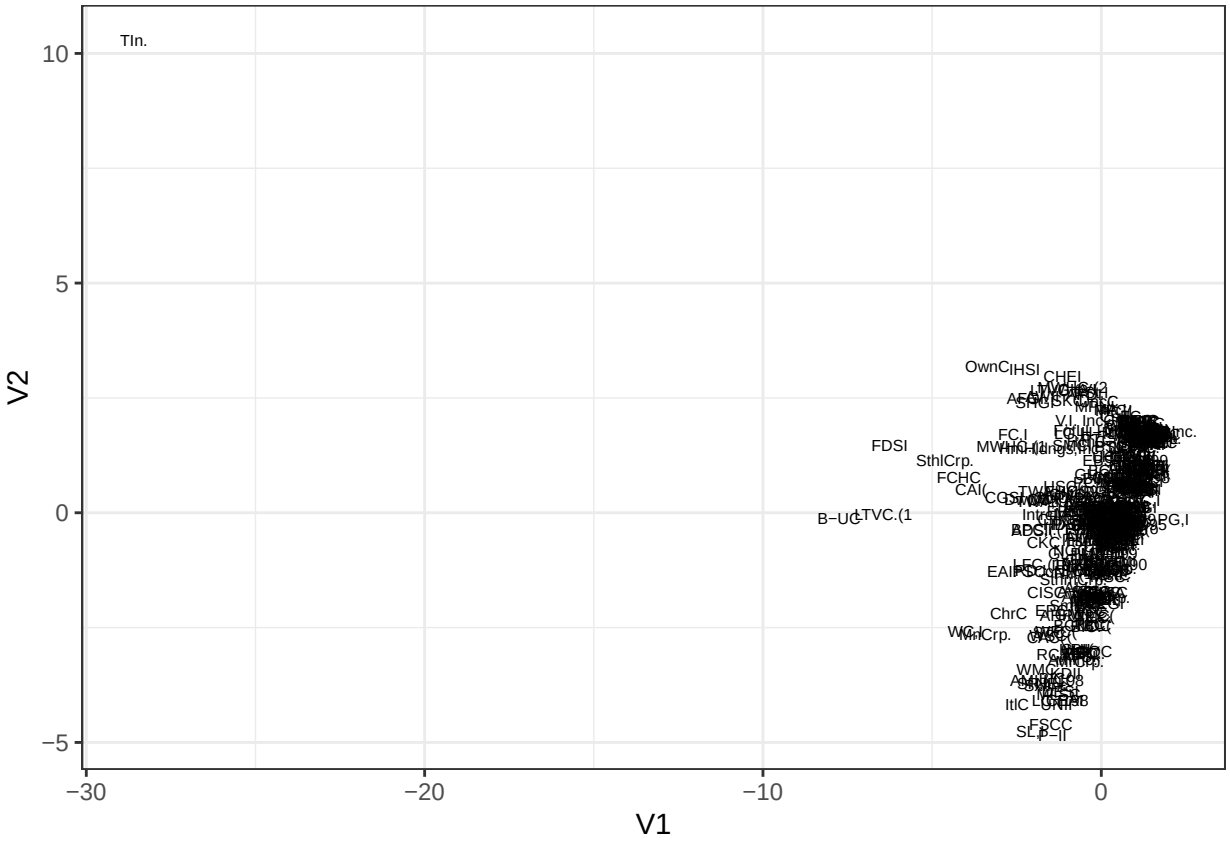


Figure 5: Classical MDS solution for bankruptcy data. The x and y-axis represent the new variables as the result of MDS. Some outliers observed in thedata

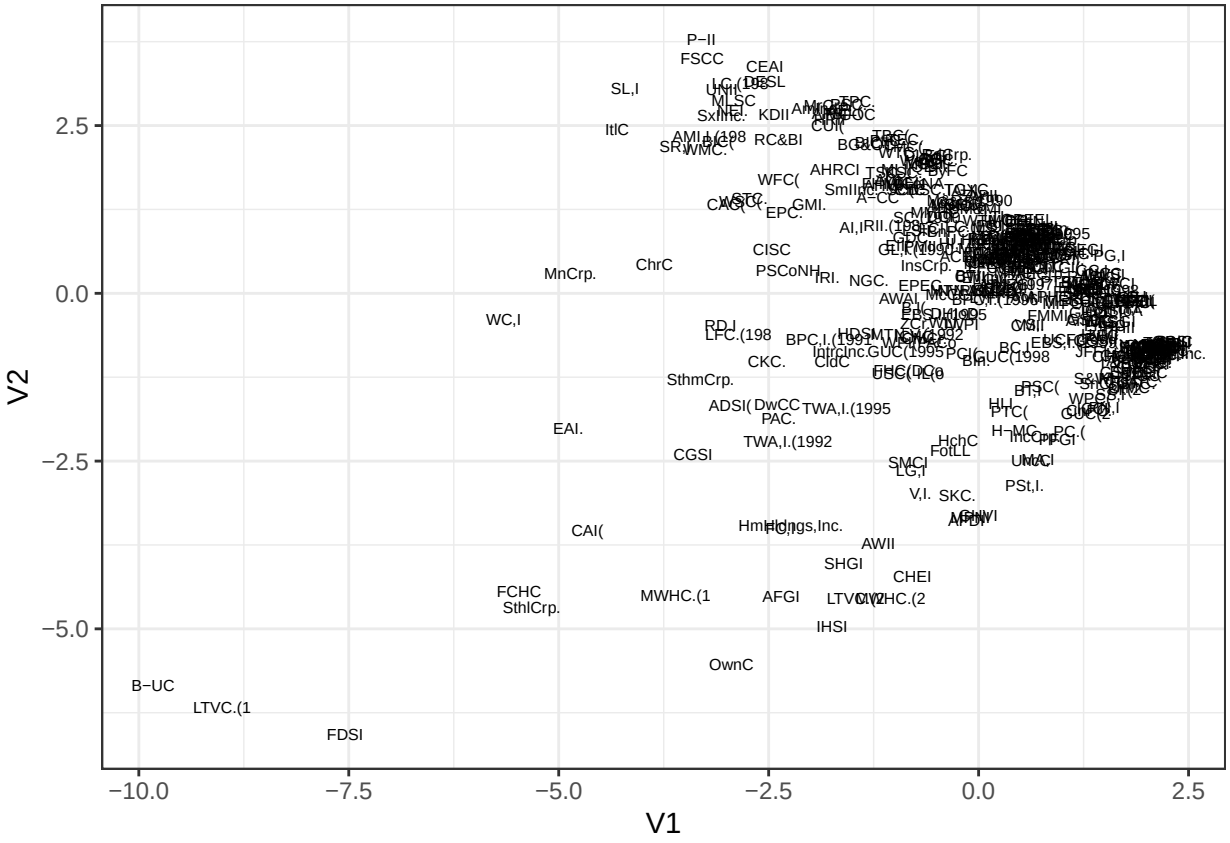


Figure 6: Classical MDS solution for bankruptcy data after excluding Texaco Inc. The x and y-axis represent the new variables as the result of MDS. We get a clearer visualisation compared to the previous MDS result

means that these firms have different profile.

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:

```
##  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.

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.

4 Principal Component Analysis (PCA)

5 Cluster Analysis

6 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.

7 Conclusions

8 Appendix

8.1 Data Cleaning

- Imputation of variable `HeadCourtCityToDE`

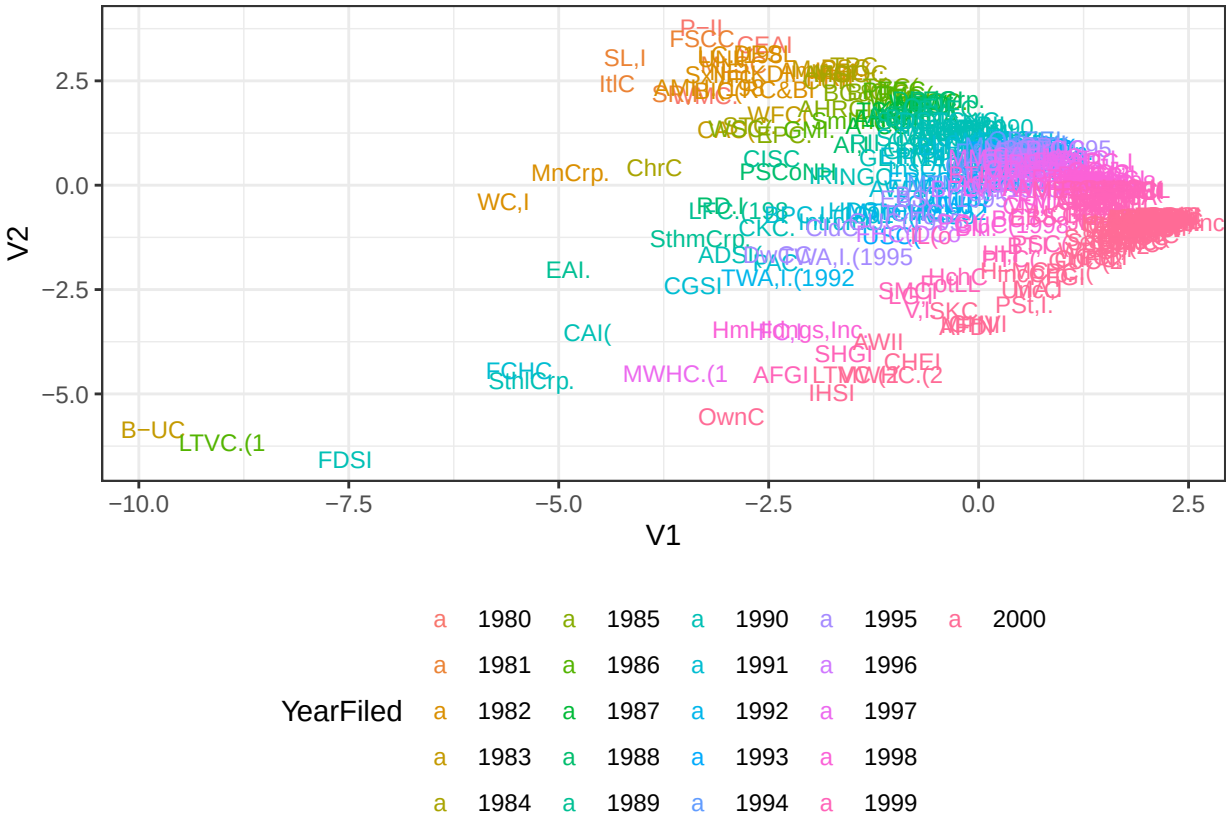


Figure 7: Classical MDS solution plotted by year when the bankruptcy filed.

Name	Sales
County Seat, Inc.	NA

Name	Employees
County Seat, Inc.	NA

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

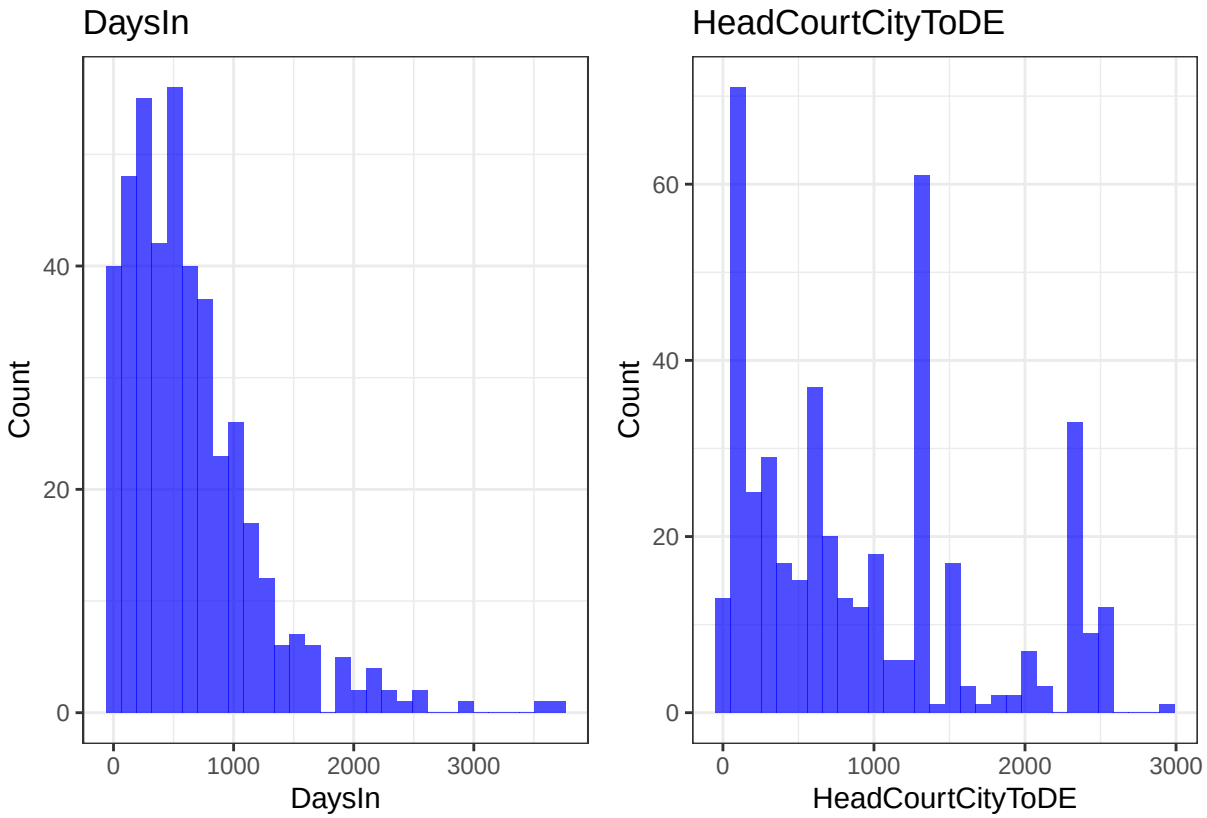


Figure 8: The figure indicates that there isnt any outliers in the variables DaysIn and HeadCourtCitytoDE

8.2 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

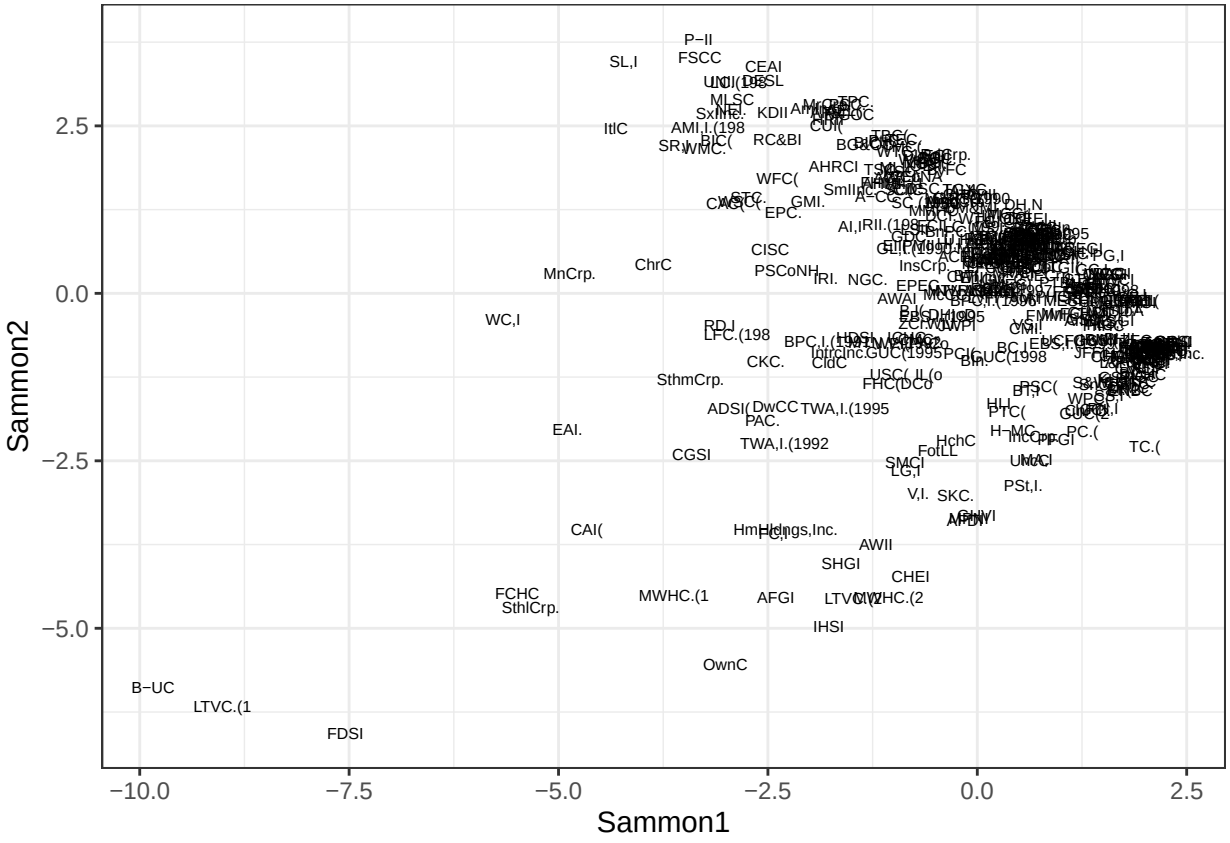


Figure 9: MDS solution using Sammon mapping

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

Figure 10 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 11 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 11 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.

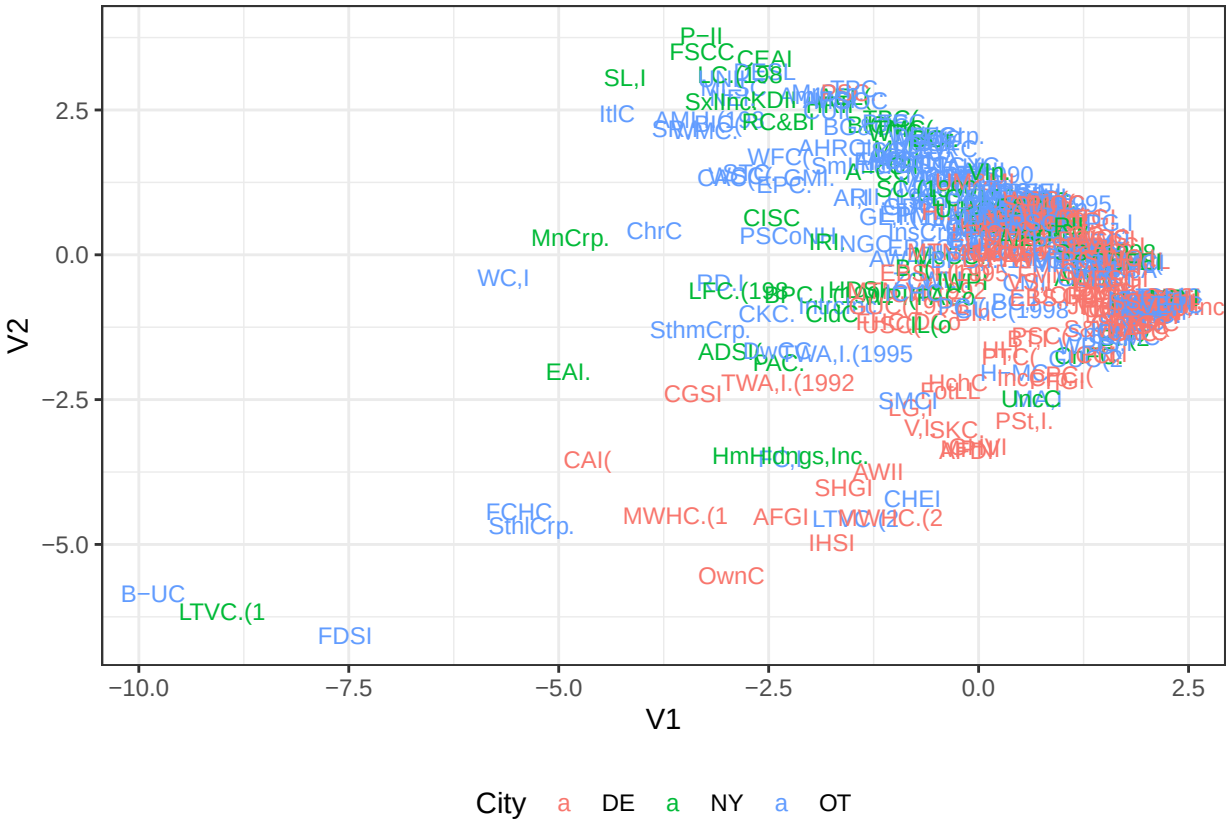


Figure 10: Classical MDS solution plotted by city where the bankruptcy filed.

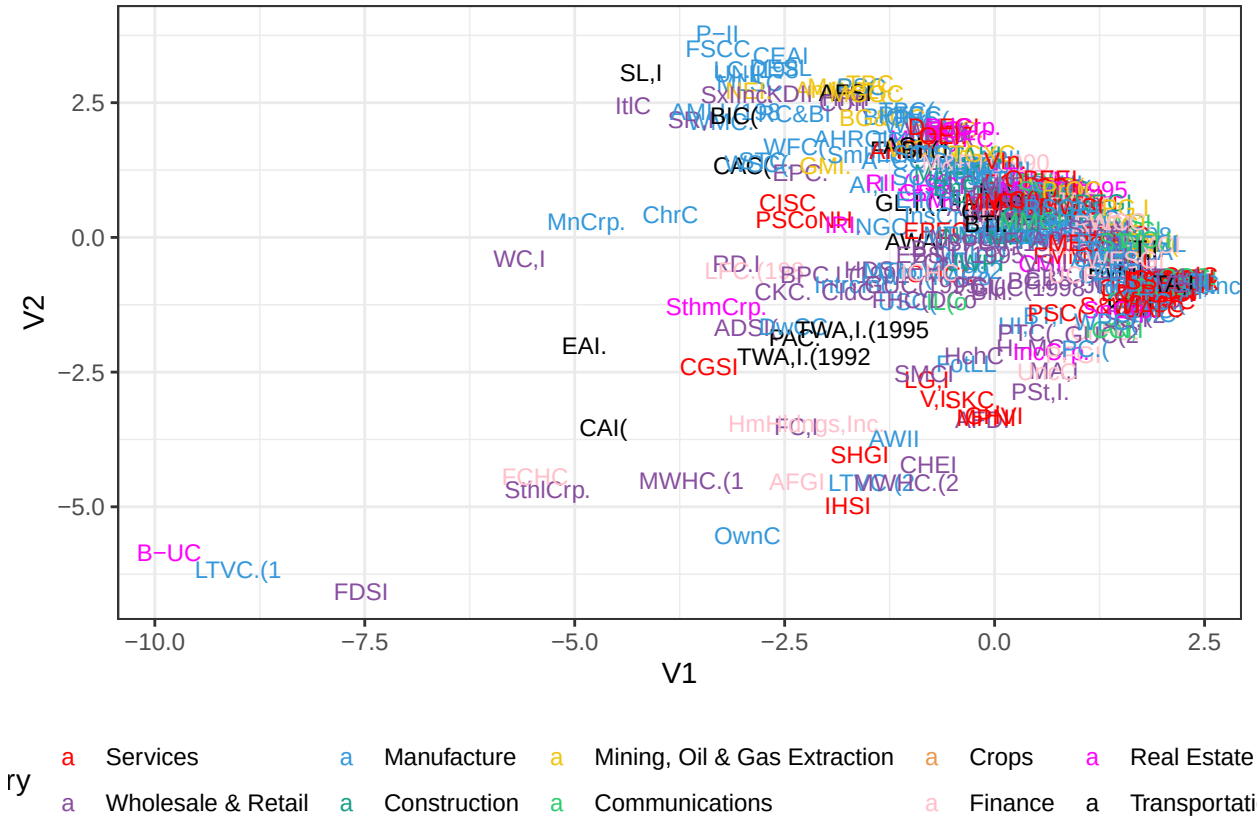


Figure 11: Classical MDS solution plotted by industry.