Multidimensional Reduction Analysis for Bankrupt Companies

Group 10

Yuheng Cui Chenjie Gong Peimin Lin Panagiotis Stylianos

Contents

1 Data description

Data are collected on 21 variables each representing different measures of status of 436 bankrupt companies in the US. Table 1 has the detailed variable description.

Table 1: Variable Description

Variable	Decription
Name	Name of the firm
Assets	Total assets (in millions of dollars)
CityFiled	City where filing took place
CPI	U.S CIP 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

Among these variables, `Assets`, `Ebit`, `GDP`, `Liab`, `Employees` and `Sales` are the measures of the status of the companies. `CPI`, `PrimeFiling` and `CityFiled` describe the external environment of the companies. `FirmEnd` tells three different endings of the companies: merged with others, bankruptcy, and continuing the operations.

Figure 1 shows the missing values in the data set. The most missing values are in `EmplUnion`; fortunately, this variable is not important.

Table 2: Suspicious Observations

Name	Assets	Employees
Residential Resources Mortgage Investments Corp.	513	1
Mortgage & Realty Trust (1990)	1022	1
EUA Power Corp.	686	1
NACO Finance Corp.	328	1
Commonwealth Equity Trust	489	1
Promus Companies Inc. (Harrahs Jazz Co. only)	1095	1

The data set has some suspicious observations. Table 2 shows some companies which only have one employee having millions of assets. Therefore, the data set might not be so trustworthy. Further investigation is required.

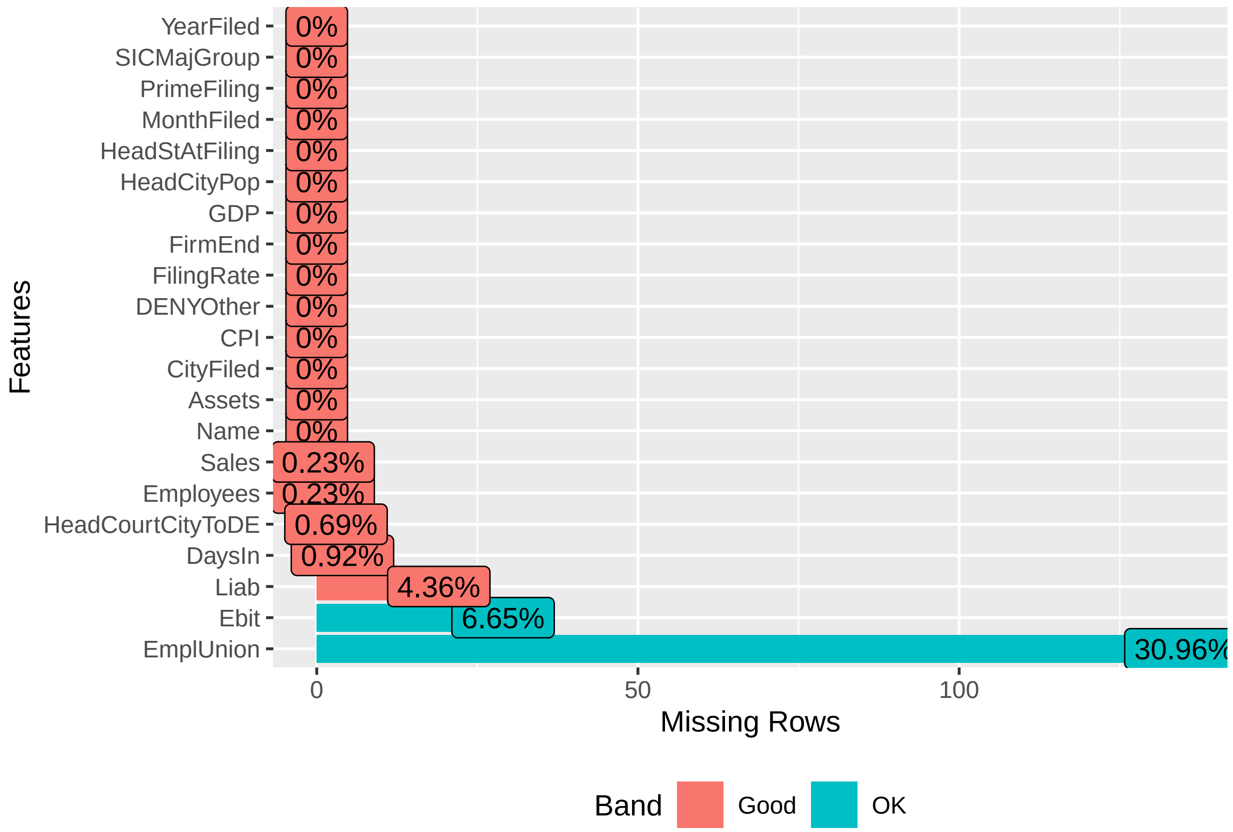


Figure 1: Missing values in the data set

2 Data cleaning

Figure 2 shows the relations between any two of the numeric variables in the data set. We can clearly see some outliers in Figure 2. In addition, we can tell some linear relationship between `Assets` and `Ebit`. For `Sales`, it is difficult to tell any clear relationship with any one of the other variables. We assume that the firms which have similar amounts of assets, EBIT, and liability would have similar sales in the same industry. Therefore, we use `impute_knn()` to impute missing values in `Sales`. Following the same logic, we can impute missing values in `Employees` as well.

2.1 Imputation

```
bankruptcy_imp <- impute_lm(bankruptcy_clean, Liab ~ Assets) %>% # impute ‘Liab‘

  impute_lm(Ebit ~ Assets) %>% # impute ‘Ebit‘

  impute_knn(Sales ~ Assets + Ebit + Liab + group_code,
  pool = "univariate", k = 5) %>% # impute ‘Sales‘

  impute_knn(Employees ~ Assets + Ebit + Sales + group_code,
  pool = "univariate", k = 5) # impute ‘Employees‘
```

`bankruptcy_imp` is the data set after imputation. In Figure 3, we can see that all important numeric variables have no missing values.

3 Factor Analayasis

Table 3: Check correlation between factors

	Factor1	Factor2	Factor3
Factor1	1.0000000	-0.0115084	-0.0317567
Factor2	-0.0115084	1.0000000	0.0165982
Factor3	-0.0317567	0.0165982	1.0000000

We use `varimax` rotation and `Bartlett` score methods for **Factor Analysis**. We tried different numbers of factors, and found 3 factors were the most reasonable. The correlation between factors are all very small (Table 3).

```
##
## Call:
## factanal(x = ., factors = 3, scores = "Bartlett", rotation = "varimax",     lower = 0.01)
##
## Uniquenesses:
##      Assets         CPI        Ebit   Employees        Liab PrimeFiling
##       0.010       0.573       0.440       0.742       0.031       0.478
##       Sales HeadCityPop
##       0.010       0.958
##
## Loadings:
```

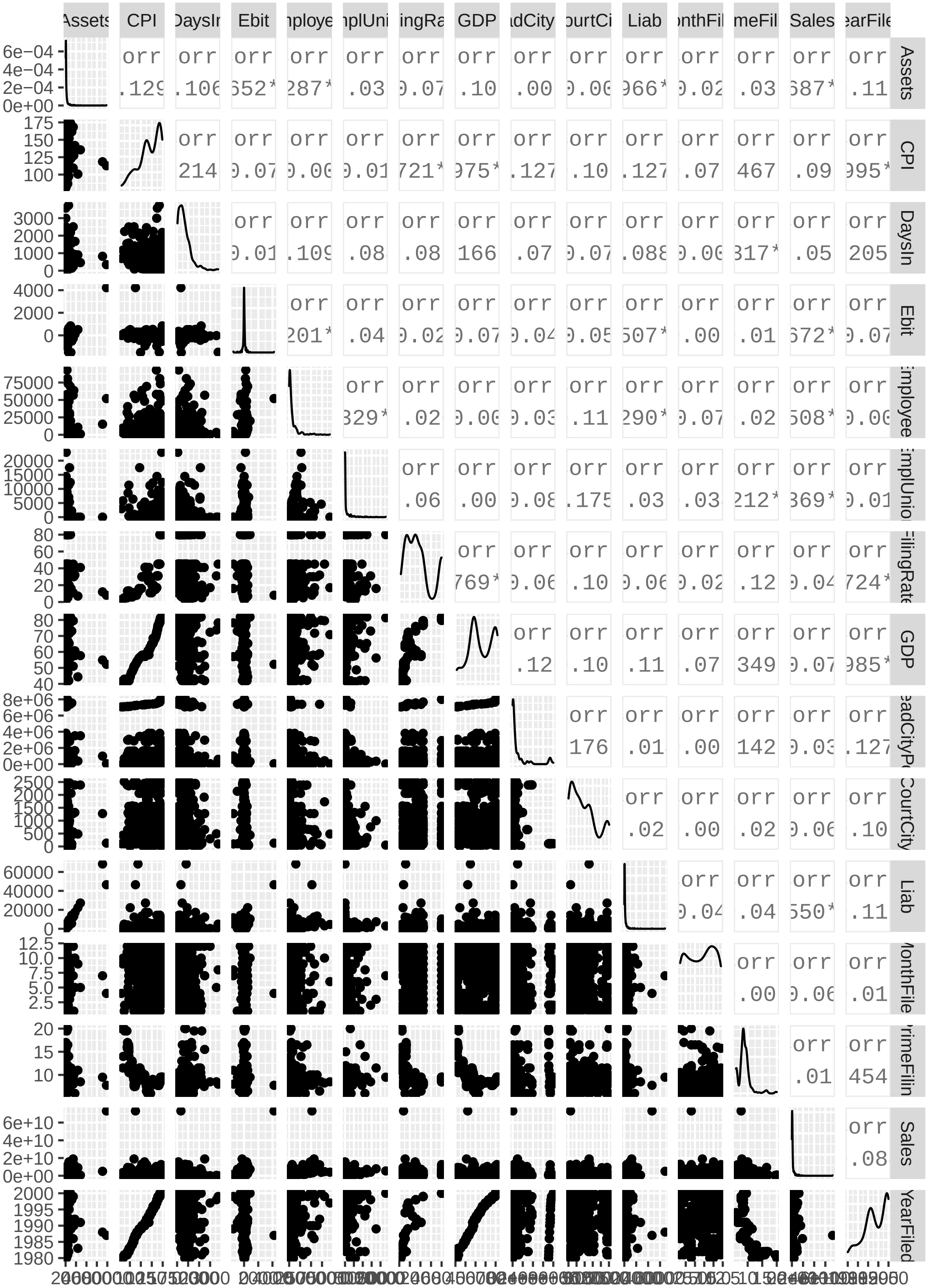


Figure 2: Overview of all numeric variables in the data set

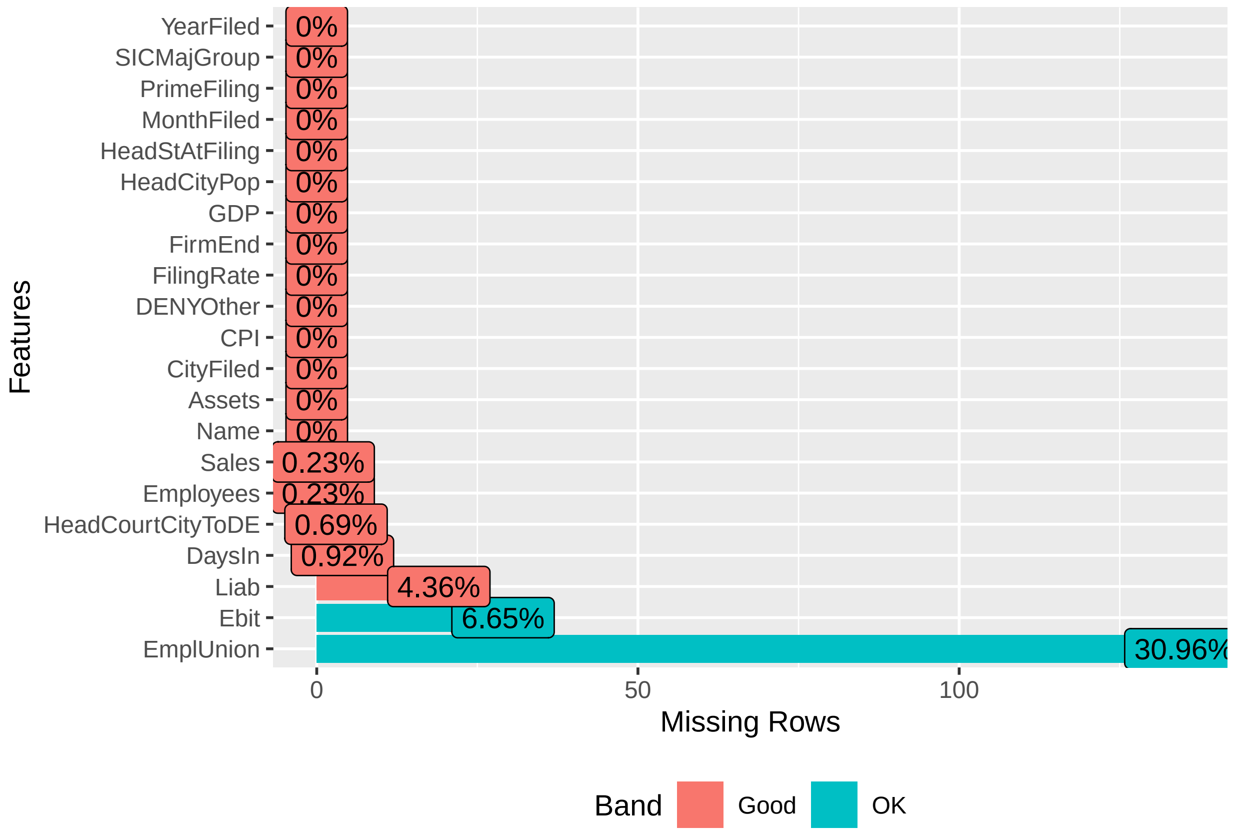


Figure 3: Check missing values after imputation

```
##             Factor1 Factor2 Factor3
## Assets       0.936   0.334        
## CPI                          -0.646
## Ebit         0.649   0.372        
## Employees    0.154   0.484        
## Liab         0.967   0.172        
## PrimeFiling                   0.722
## Sales        0.410   0.906        
## HeadCityPop                   0.200
##
##                Factor1 Factor2 Factor3
## SS loadings      2.429   1.344   0.986
## Proportion Var   0.304   0.168   0.123
## Cumulative Var   0.304   0.472   0.595
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 184.84 on 7 degrees of freedom.
## The p-value is 1.85e-36
```

`Factor 1` has high loadings for `Assets` and `Liab`; it is a company's economies of scale factor with higher scores associated with larger scale companies. `Factor 2` has high loadings for `Sales`; it is a sales factor with higher scores associated with bigger sales. `Factor 3` has high loadings for `PrimeFiling`; it is the interest rate of borrowing factor with higher score associated with higher borrowing rate.

3.1 Limitation of FA

According to the ***Factor Analysis*** output, `HeadCityPop` and `Employees` have very high value of `Uniquenesses` – 95.9% of `HeadCityPop` cannot be explained by the ***Factor Analysis*** while 73.8% of `Employees` cannot be explained.

4 Principal Components Analysis and Biplot

Principal components analysis finds a small number of linear combinations of the orginal variables that explain a large proportion of overall variation in the data. Since the variables in the dataset under investigation are measured in different units we standardise the data by dividing by the standard deviation before conducting the analysis. By selecting two principal components we are able visualise the data using a biplot which is included below

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5    PC6     PC7
## Standard deviation     1.7746 1.2311 0.9657 0.72613 0.70425 0.6039 0.12061
## Proportion of Variance 0.4499 0.2165 0.1332 0.07532 0.07085 0.0521 0.00208
## Cumulative Proportion  0.4499 0.6664 0.7996 0.87497 0.94583 0.9979 1.00000
```

Proportion of variance explained by the first five PCs together is 77.62%. Proportion of variance explained by the three PC alone is 15.5%. By Kaisers rule select 3 PCs

The elbow appears at the three PC, therefore 3 or 4 PCs should be used. Figure **??**)

The biplot 5) can be interpreted as follows. The first principal component is a measure of overall company situation since it is positively correlated with variables that indicate a company's bankruptcy, such as cpi, Ebit and GDP . Some values with low values of the first principal component are CDSI, CHVI and AWII.

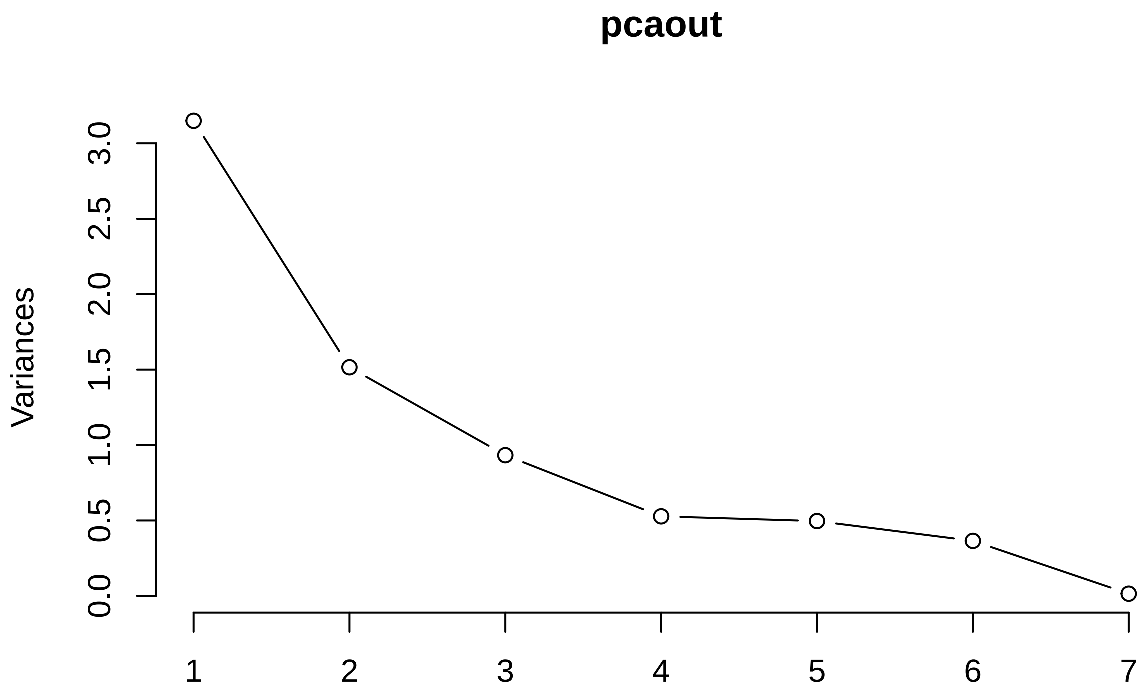


Figure 4: Screeplot of PCA

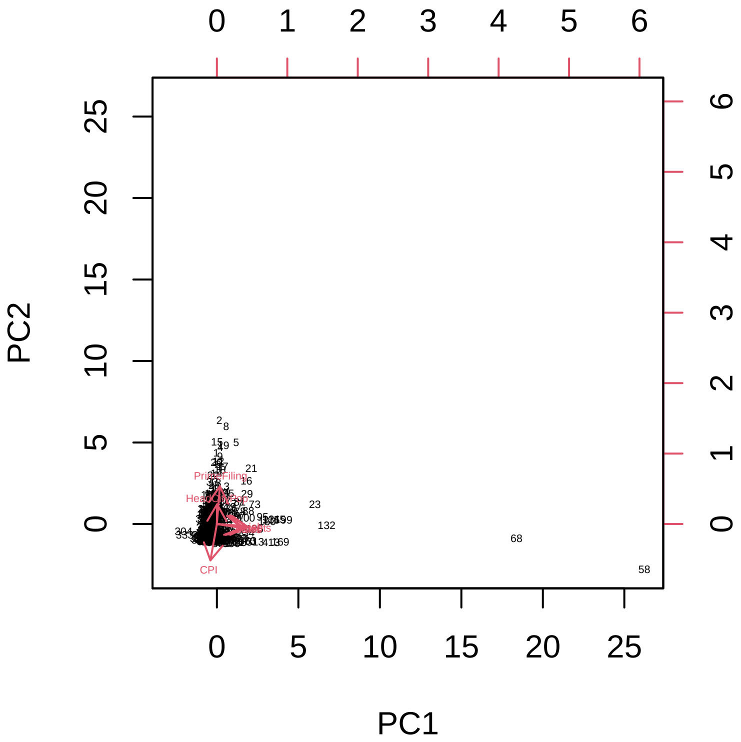


Figure 5: Biplot produced by PCA

It’s all pharmaceutical companies, and the response to the cpi numbers is not that big, because they’re all essential elements of life

The biplot also highlights that the B-UC and SthpCrp are outliers, particularly on the first principal component. The first principal component has a high weight of 0.33. Texaco Corporation Texaco Corporation is one of the largest oil companies in the United States and an international oil multinational company with more than 120 subsidiaries and branches. It explores oil in 32 countries, extracts oil in 18 countries, and sells in more than 130 countries. and FRBC is a bank. So these two companies are resource-based companies and banks, both of which are very competitive, so their data may not be very reliable and cannot help us analyze.

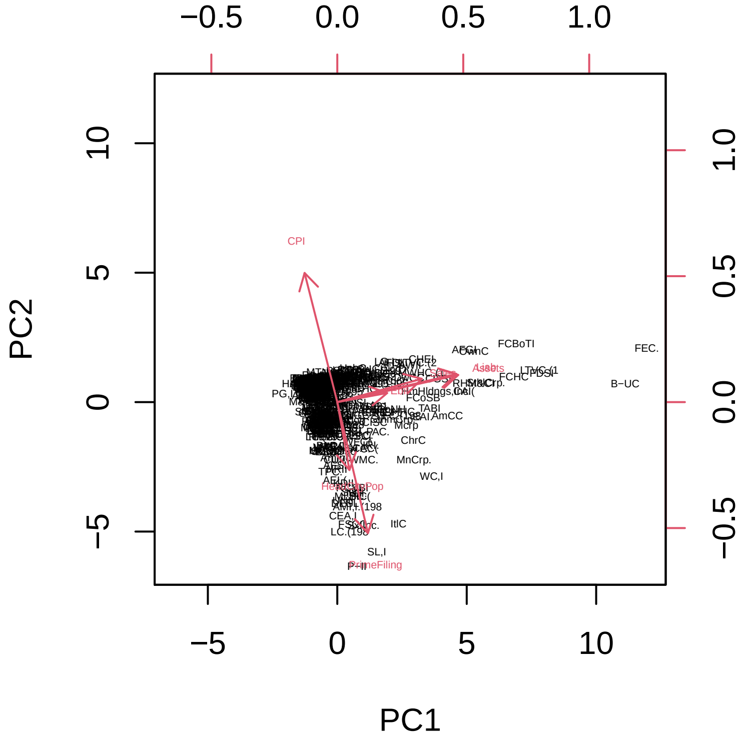


Figure 6: Biplot without Outliers

6) From this plot we can more viivd to see the result, because we remove these two outliers, which are First RepublicBank Corp and Texaco Inc. and we can see primeFliling has longest line, which means this factor has biggest influence with company. and Liabs and Sales are the same way, and the CPI is Completely opposite direction.

5 Limitations of the Analysis

Any dimension reduction technique such as principal components analysis represents a loss of information. In this example 0.7762 of the overall variation is explained by the first two principal components and therefore accurately depicted in the biplot. Finally there is some concern that the outliers of FRBC lead to a misleading analysis.

6 Cluster

Why does a company go bankrupt? The biggest reason will be related to their financial position. According to the data, the company's relevant information are Assets, EBIT, Liabilities and Sales. For Assets is a company owned and can provide future economic benefit. Liabilities represent money owed for other parties.

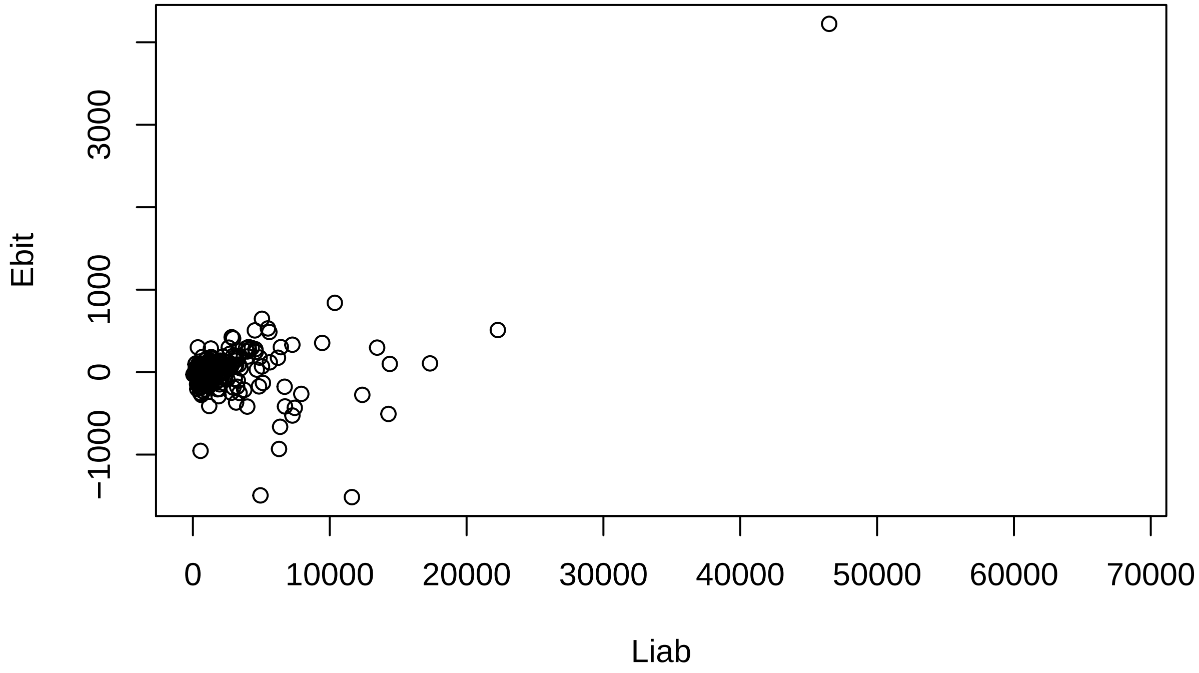


Figure 7: Overview companies financial positions

EBIT is an significant index to evaluated the company's operating efficiency. Sales reflect the company's transaction between other parties. Thus, the cluster analysis will focus on company's financial position. Figure 7 shows the EBIT and liabilities of companies, most companies' EBIT are less than liabilities and even in negativewhich means they did not have ability to pay the debt which caused bankruptcy in the end. The company `First RepublicBank Corp` and `Texaco Inc.` have large amount of liabilities than other companies and for better clustering, we will consider them as outliers and remove them. As variable `Sales` amounts are larger than other financial positions, we have to normalize them before calculating the distance.

As variable `Sales` amounts are larger than other financial positions, we have to normalize them before calculating the distance.

```
## ****************************************
## *** INPUT:
## ****************************************
## * nbCluster =  4 5 6 7 8 9 10
## * criterion =  BIC
## ****************************************
## *** MIXMOD Models:
```

```
## * list =  Gaussian_pk_Lk_C
## * This list includes only models with free proportions.
## ****************************************
## * data (limited to a 10x10 matrix) =
##       Assets Ebit   Liab  Sales     PrimeFiling
##  [1,] 531    13.83  309.7 3.575e+08 14
##  [2,] 552    -13.52 377.9 9.001e+08 20
##  [3,] 1897   102.6  1202  3.662e+09 11.5
##  [4,] 821    71.5   751.4 4.239e+08 19.5
##  [5,] 4097   176.4  4872  6.017e+08 20
##  [6,] 1200   -90.19 845.2 1.652e+09 15.75
##  [7,] 1141   35.63  995.8 1.32e+09  16
##  [8,] 2628   50.46  2659  2.144e+08 19.5
##  [9,] 1456   -68.62 1419  1.739e+09 16.5
## [10,] 1031   53.97  899.9 1.371e+08 11.5
## * ... ...
## ****************************************
## *** MIXMOD Strategy:
## * algorithm            =  EM
## * number of tries      =  1
## * number of iterations =  200
## * epsilon              =  0.001
## *** Initialization strategy:
## * algorithm            =  smallEM
## * number of tries      =  10
## * number of iterations =  5
## * epsilon              =  0.001
## * seed                 =  NULL
## ****************************************
##
##
## ****************************************
## *** BEST MODEL OUTPUT:
## *** According to the BIC criterion
## ****************************************
## * nbCluster   =  8
## * model name  =  Gaussian_pk_Lk_C
## * criterion   =  BIC(39531.5719)
## * likelihood  =  -19556.2659
## ****************************************
## *** Cluster 1
## * proportion =  0.0323
## * means      =  869.2312 6.6517 727.9773 995795213.5832 15.9917
## * variances  = | 246658.0521 16745.1884 191927.2926 62415560468.5982   -25.4744 |
##                | 16745.1884  9844.1123  9697.3443 11997820101.3069     9.4961 |
##                | 191927.2926  9697.3443 245839.0403 52436526074.8874   -40.6777 |
##                | 62415560468.5982 11997820101.3069 52436526074.8874 738912081977142784.0000 -43172100
##                |   -25.4744     9.4961   -40.6777 -43172100.3318     1.2670 |
## *** Cluster 2
## * proportion =  0.0979
## * means      =  569.0556 -7.1329 563.0046 671694443.8937 6.1349
## * variances  = | 47475.0245  3222.9973 36940.8289 12013312494.3033    -4.9031 |
##                |  3222.9973  1894.7262  1866.4773 2309256875.1337     1.8277 |
##                | 36940.8289  1866.4773 47317.3868 10092617435.8433    -7.8294 |
##                | 12013312494.3033 2309256875.1337 10092617435.8433 142220652669996368.0000 -8309465.
##                |    -4.9031     1.8277    -7.8294 -8309465.2748     0.2439 |
## *** Cluster 3
## * proportion =  0.1963
## * means      =  466.7333 -2.9590 432.8462 458560196.5762 8.3001
## * variances  = | 30347.8241  2060.2613 23613.9695 7679361918.1239    -3.1343 |
##                |  2060.2613  1211.1804  1193.1226 1476163990.1132     1.1684 |
##                | 23613.9695  1193.1226 30247.0562 6451581279.3317    -5.0048 |
##                | 7679361918.1239 1476163990.1132 6451581279.3317 90912798997162448.0000 -5311723.242
##                |    -3.1343     1.1684    -5.0048 -5311723.2422     0.1559 |
## *** Cluster 4
## * proportion =  0.1172
## * means      =  5000.4507 42.5632 5031.1677 4637559791.2439 9.9316
## * variances  = | 4221006.9329 286556.8571 3284411.0529 1068104248650.0616  -435.9375 |
##                | 286556.8571 168460.2065 165948.5967 205316150788.9096   162.5039 |
##                | 3284411.0529 165948.5967 4206991.3572 897335149513.1565  -696.1096 |
##                | 1068104248650.0616 205316150788.9096 897335149513.1565 12644845743806466048.0000 -73
##                |  -435.9375   162.5039  -696.1096 -738794996.6531    21.6814 |
## *** Cluster 5
## * proportion =  0.0247
## * means      =  15948.1349 35.0452 14990.8710 4843532176.6741 8.8815
## * variances  = | 13641937.5940 926127.5388 10614938.8827 3452022641886.6963 -1408.9133 |
##                | 926127.5388 544449.1470 536331.8365 663564443418.4304   525.1989 |
##                | 10614938.8827 536331.8365 13596640.4380 2900111349051.5293 -2249.7673 |
##                | 3452022641886.6963 663564443418.4304 2900111349051.5293 40867072540861603840.0000 -
##                | -1408.9133   525.1989 -2249.7673 -2387722976.8373    70.0724 |
## *** Cluster 6
## * proportion =  0.2827
## * means      =  839.2159 -11.3615 818.4152 806735036.4847 8.9987
## * variances  = | 142116.2862  9648.0288 110582.2162 35961800462.8503   -14.6775 |
##                |  9648.0288  5671.8549  5587.2920 6912750750.5031     5.4713 |
##                | 110582.2162  5587.2920 141644.3984 30212207877.5351   -23.4372 |
##                | 35961800462.8503 6912750750.5031 30212207877.5351 425736926050999616.0000 -24874349
##                |   -14.6775     5.4713   -23.4372 -24874349.3776     0.7300 |
## *** Cluster 7
## * proportion =  0.0698
## * means      =  444.8920 1.7975 388.5113 424807332.9175 9.5304
## * variances  = | 16398.3657  1113.2567 12759.7454 4149522708.3971    -1.6936 |
##                |  1113.2567   654.4581   644.7006 797641270.6681     0.6313 |
##                | 12759.7454   644.7006 16343.9160 3486094718.4264    -2.7043 |
##                | 4149522708.3971 797641270.6681 3486094718.4264 49124488198992248.0000 -2870175.4715
##                |    -1.6936     0.6313    -2.7043 -2870175.4715     0.0842 |
## *** Cluster 8
## * proportion =  0.1791
## * means      =  1978.5280 17.4309 1868.3704 2245657821.2240 8.5987
## * variances  = | 639055.2719 43384.3567 497255.8045 161709673041.5073   -66.0004 |
##                | 43384.3567 25504.6686 25124.4140 31084613375.6866    24.6029 |
##                | 497255.8045 25124.4140 636933.3309 135855440908.3393  -105.3901 |
##                | 161709673041.5073 31084613375.6866 135855440908.3393 1914414134646107392.0000 -1118
##                |   -66.0004    24.6029  -105.3901 -111852656.2407     3.2825 |
## ***************************************
```

As the data is multidimensional, we use Gaussian Mixture model tofit the data. Comparing the BIC, we can find that clusters of 8 are the best cluster group numbers.

Then, we are also using Ward method, average method, centroid method and complete method to check the rand index. The rand index of ward's method with average method is 0.15, with centroid method is 0.1 and with complete method is 0.29. Though, the rand index is low, complete method has a relatively high level of agreement with Ward's method.

Table 4: Mean of each clusters

Group.1	Assets	Ebit	Liab	Sales	PrimeFiling
1	-0.3375478	-0.1239119	-0.3404528	-0.3450447	0.0307723
2	-0.1163503	0.0203941	-0.1232866	-0.2144911	3.8960084
3	1.0119377	1.1989002	0.8769185	0.3896130	0.4244365
4	5.8259844	1.6195397	5.9730896	0.2471584	-0.1783107
5	0.2521142	0.2437523	0.2749703	1.9206714	-0.0125887
6	2.9789713	1.2034281	2.8946256	5.6483557	0.0098640
7	-0.2387320	-0.0646959	-0.1893929	-0.0918781	-1.1583875
8	1.6705682	-3.9272793	1.9098296	1.8675638	0.1745169

Table 4 shows the mean of each cluster, cluster 4 has the highest average assets and liabilities, however, their EBIT are not the best and prime rate of interest's average are negative. Cluster 3 has the highest average EBIT, but their average sales are low. Overview, the cluster can find that the bankruptcy companies always arise some issues in their financial position that caused their companies went into filling.

7 Multidimensional Scaling

For this part the focus was on finding a 2D representation of the data that accurately depicts the distance of the observations in the higher dimensions. The main concerns was how to combine both the numeric and the categorical variables in the dataset when calculating the distance metric, along with which variables to select that provide a high goodness of fit measure for the multidimensional scaling providing at the same time that these subset of variables doesn't alter the structure of the observations in the higher dimensional space.

Furthermore, we were also curious to see if the clusters found in the previous section could also be identified in the 2D representation of the multidimensional scaling.

Classical multidimensional scaling was used, as there was no indication from the initial exploratory analysis that the relationship between some observations was non-linear.

The distance metric used in the analysis was the sum of the Euclidean distances between the numeric variables and the Jaccard distance between the non-numeric variables for each observation. To calculate the latter, all character variables were transformed to dummy variables (0, 1).

Table 5: Goodnes of Fit Measures

Value
0.6819489
0.6819489

Table 5 suggests that selecting a subset of the numeric variables that represent the financial health status of the company alongside with the character variables, such as location and industry yields a respectable value for the metrics, indicating a trustworthy multidimensional scaling.

Figure 8 indicates that the result of the multidimensional scaling is consistent with the clustering results, as the 2D representation of the companies follows a clear structure with easily identifiable clusters, especially for those with a lot of observations.

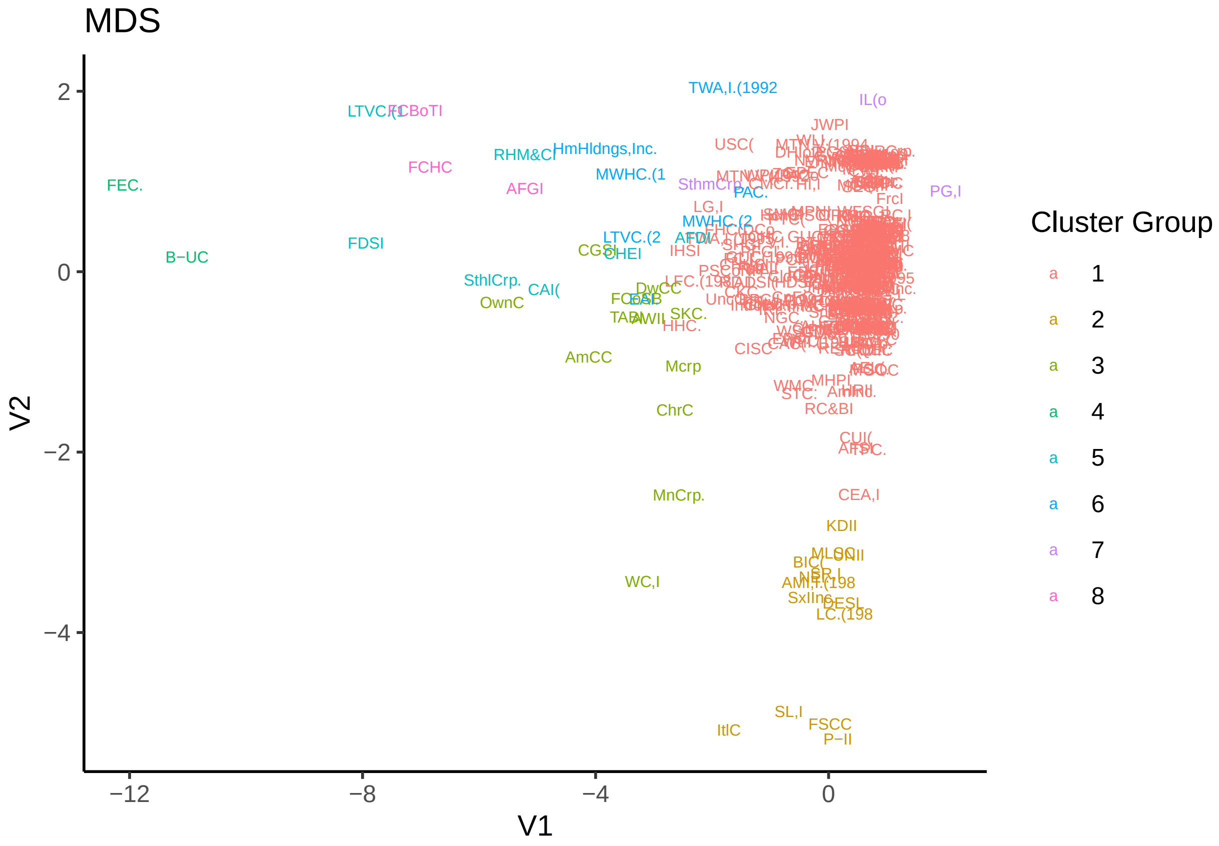
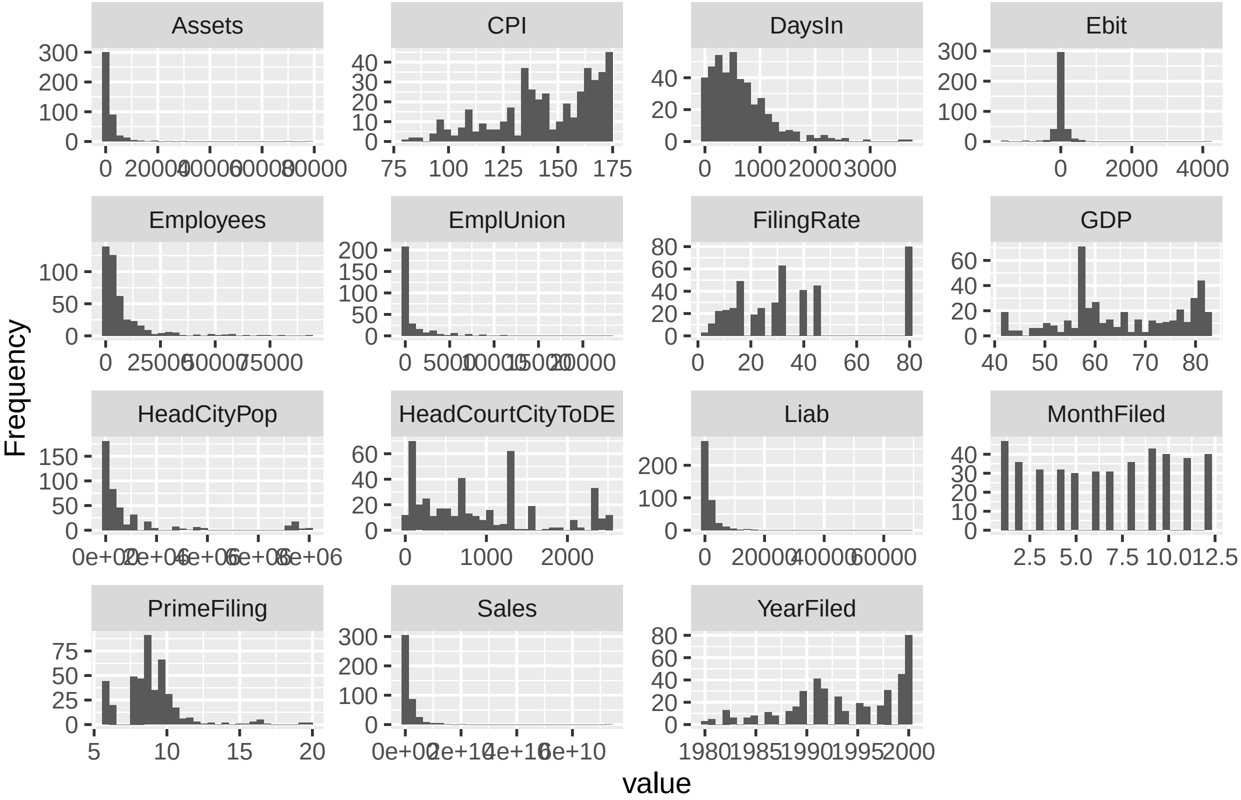


Figure 8: Multidimensional Scaling 2D representation

8 APPENDIX



In the above plot we see that there are no as such obvious outliers. Thus, there is no need to cap outliers from this data set.

Also, an observation regarding the data is that, most of the variables showing histogram are **left Skewed** this is because the mean is greater than the median. In this case because skewed-right or left data have a few large values that drive the mean upward but do not affect where the exact middle of the data is (that is, the median).

```
##      Name               Assets          CityFiled              CPI
##  Length:436         Min.   :  258.0   Length:436         Min.   : 81.0
##  Class :character   1st Qu.:  460.2   Class :character   1st Qu.:133.5
##  Mode  :character   Median :  723.0   Mode  :character   Median :146.4
##                     Mean   : 2112.9                      Mean   :145.4
##                     3rd Qu.: 1827.8                      3rd Qu.:166.2
##                     Max.   :78401.0                      Max.   :174.1
##
##      DaysIn         DENYOther               Ebit            Employees
##  Min.   :  31.0   Length:436         Min.   :-1515.602   Min.   :    1
##  1st Qu.: 248.0   Class :character   1st Qu.:  -40.105   1st Qu.: 1086
##  Median : 510.0   Mode  :character   Median :    4.464   Median : 3330
##  Mean   : 635.3                      Mean   :    9.437   Mean   : 7592
##  3rd Qu.: 870.0                      3rd Qu.:   56.121   3rd Qu.: 8000
##  Max.   :3730.0                      Max.   : 4222.978   Max.   :93000
##  NA's   :4                           NA's   :29          NA's   :1
##    EmplUnion       FilingRate      FirmEnd               GDP
##  Min.   :    1   Min.   : 3.00   Length:436         Min.   :41.30
##  1st Qu.:    1   1st Qu.:16.00   Class :character   1st Qu.:57.50
##  Median :    1   Median :31.00   Mode  :character   Median :62.91
##  Mean   : 1126   Mean   :35.75                      Mean   :65.46
##  3rd Qu.:  750   3rd Qu.:45.00                      3rd Qu.:77.17
##  Max.   :22950   Max.   :80.00                      Max.   :81.87
##  NA’s   :135
##   HeadCityPop      HeadCourtCityToDE HeadStAtFiling          Liab
##  Min.   :    144   Min.   :   1      Length:436         Min.   :   38.53
##  1st Qu.:  40968   1st Qu.: 241      Class :character   1st Qu.:  420.96
##  Median : 215333   Median : 685      Mode  :character   Median :  730.53
##  Mean   :1082146   Mean   : 920                         Mean   : 1996.24
##  3rd Qu.: 996558   3rd Qu.:1318                         3rd Qu.: 1672.18
##  Max.   :8008278   Max.   :2514                         Max.   :68403.04
##                    NA’s   :3                            NA’s   :19
##    MonthFiled     PrimeFiling         Sales            SICMajGroup
##  Min.   : 1.00   Min.   : 6.000   Min.   :2.815e+05   Length:436
##  1st Qu.: 3.00   1st Qu.: 7.750   1st Qu.:3.726e+08   Class :character
##  Median : 7.00   Median : 8.500   Median :7.648e+08   Mode  :character
##  Mean   : 6.58   Mean   : 8.878   Mean   :1.684e+09
##  3rd Qu.:10.00   3rd Qu.: 9.500   3rd Qu.:1.519e+09
##  Max.   :12.00   Max.   :20.000   Max.   :7.313e+10
##                                   NA’s   :1
##    YearFiled
##  Min.   :1980
##  1st Qu.:1990
##  Median :1994
##  Mean   :1994
##  3rd Qu.:1999
##  Max.   :2000
##
```

- Skimming function is use to check the data quality. It gives an overview of the data frame including number of rows and column, column types and if data frame is grouped. Initially we have:

1. **Data Summary**: There are **436** number of rows and **21** number of columns.
2. **Column type frequency**: In this data there total three types of variable viz. **Character**, **Interger**, and **Numeric**. The maximum missing values are seen in **NUmeric**
3. **Central tendency for all the variables**: The central tendency provides us with the information of **Mean, Standard deviation, Percentile, and Range.**