Multidimensional Reduction Analysis for Bankrupt Companies

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.

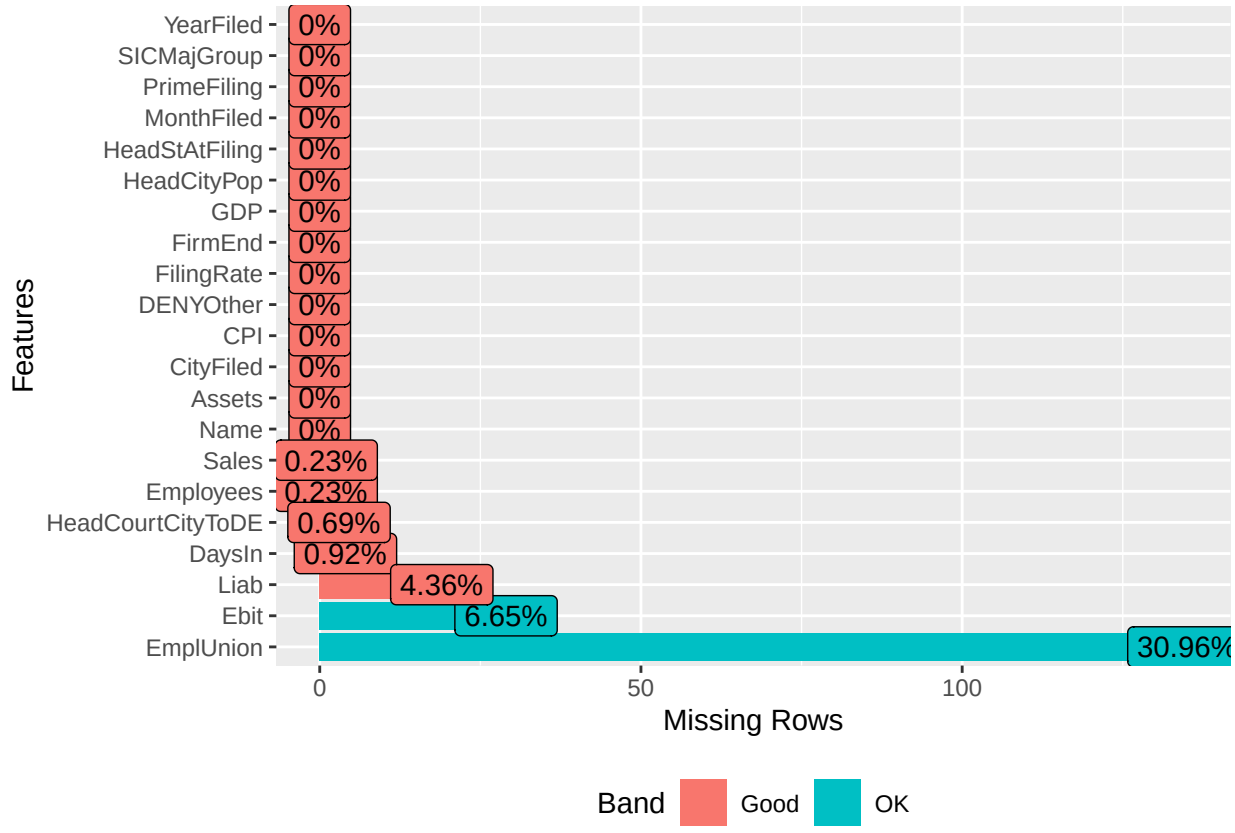


Figure 1: Missing values in the data set

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 Analysis

Table 3: Check correlation between factors

	Factor1	Factor2	Factor3
Factor1	1.0000000	-0.0115341	-0.0317576
Factor2	-0.0115341	1.0000000	0.0168070
Factor3	-0.0317576	0.0168070	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       0.575       0.440       0.741       0.031       0.475
##       Sales HeadCityPop
##       0.010       0.958
##
## Loadings:
##             Factor1 Factor2 Factor3
## Assets       0.936   0.335
## CPI                         -0.644
## Ebit         0.649   0.372
```

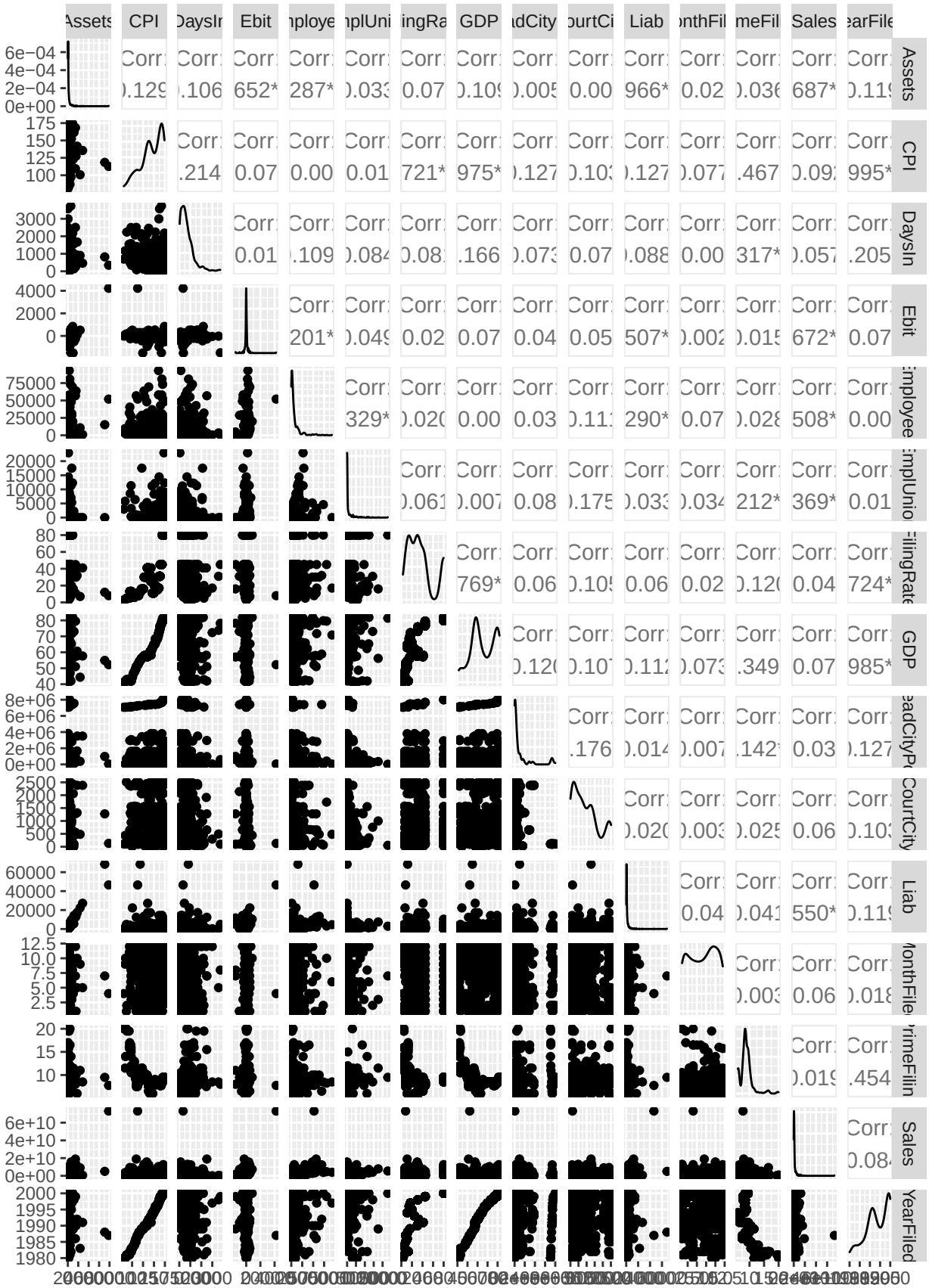


Figure 2: Overview of all numeric variables in the data set

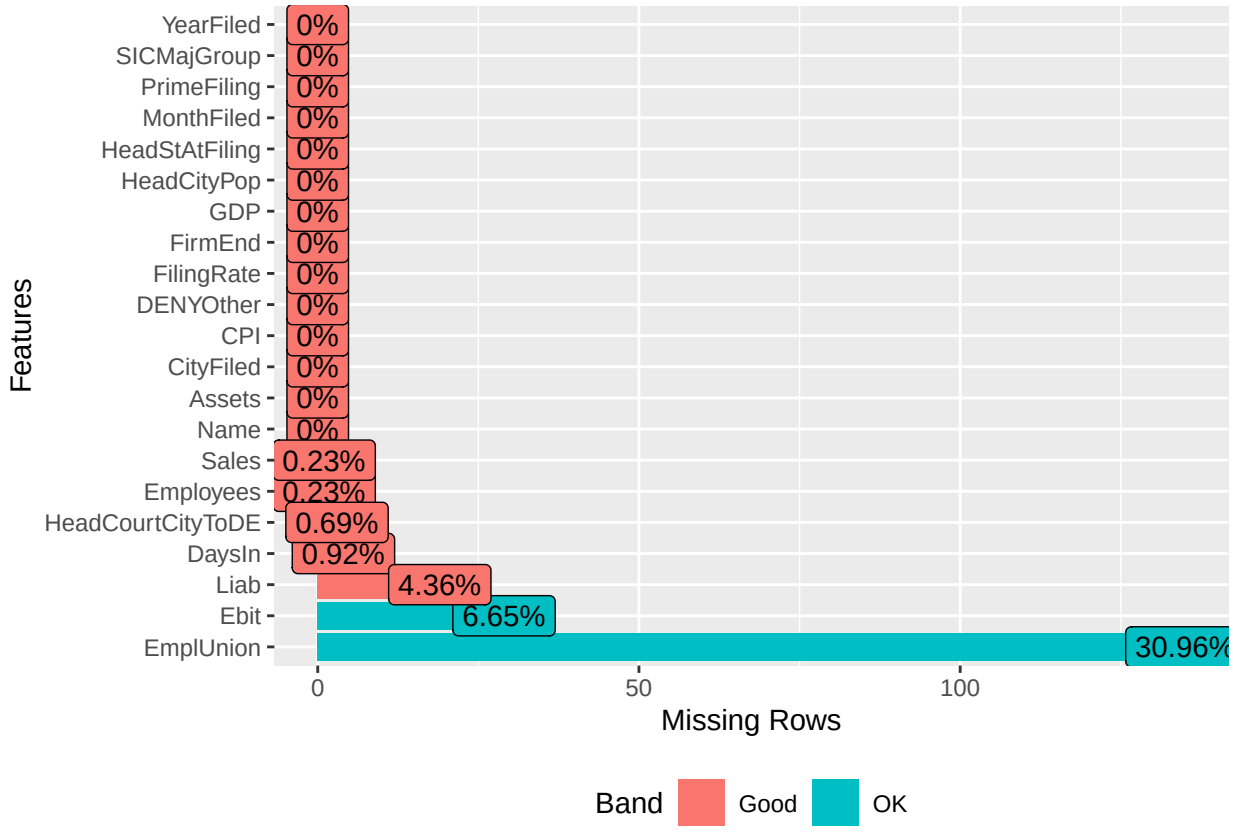


Figure 3: Check missing values after imputation

```
## Employees    0.155   0.485
## Liab         0.967   0.172
## PrimeFiling                  0.723
## Sales        0.410   0.907
## HeadCityPop                  0.200
##
##                Factor1 Factor2 Factor3
## SS loadings      2.428   1.346   0.987
## 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 185.06 on 7 degrees of freedom.
## The p-value is 1.67e-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     1.7747 1.2311 0.9657 0.72613 0.70423 0.6039 0.12060
## Proportion of Variance 0.4499 0.2165 0.1332 0.07532 0.07085 0.0521 0.00208
## Cumulative Proportion  0.4499 0.6664 0.7997 0.87498 0.94583 0.9979 1.00000
```

4.0.1 Proportion of variance explained by the first five PCs together is 77.62%

4.0.2 Proportion of variance explained by the three PC alone is 15.5%

4.0.3 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. 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

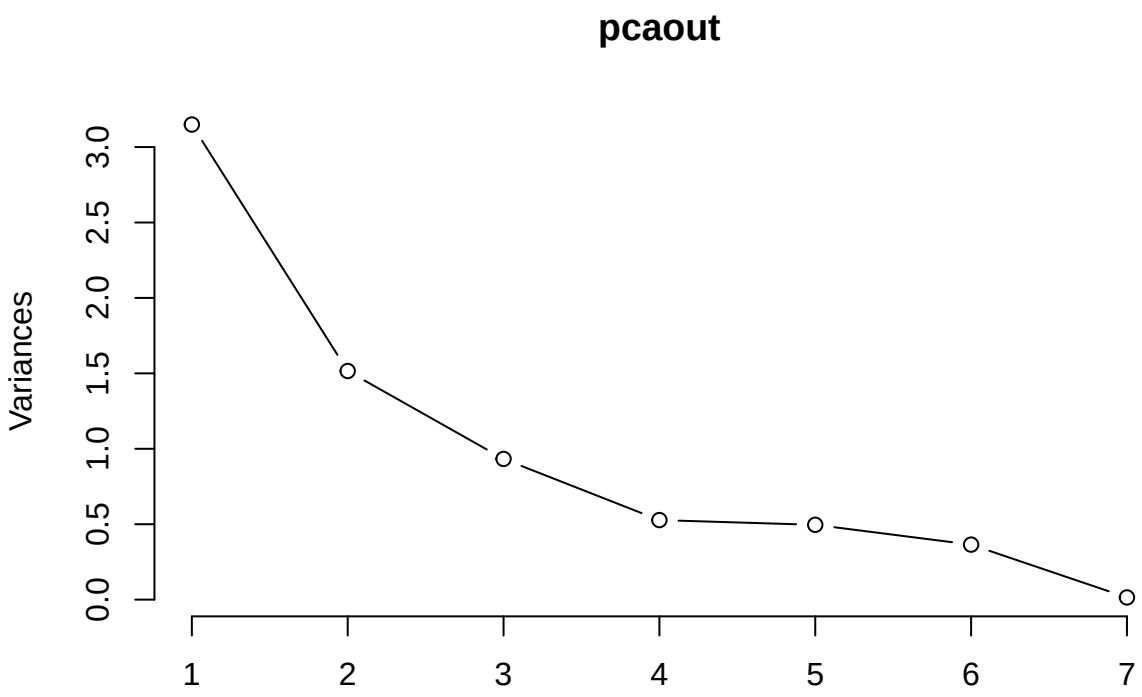


Figure 4: Screeplot of PCA

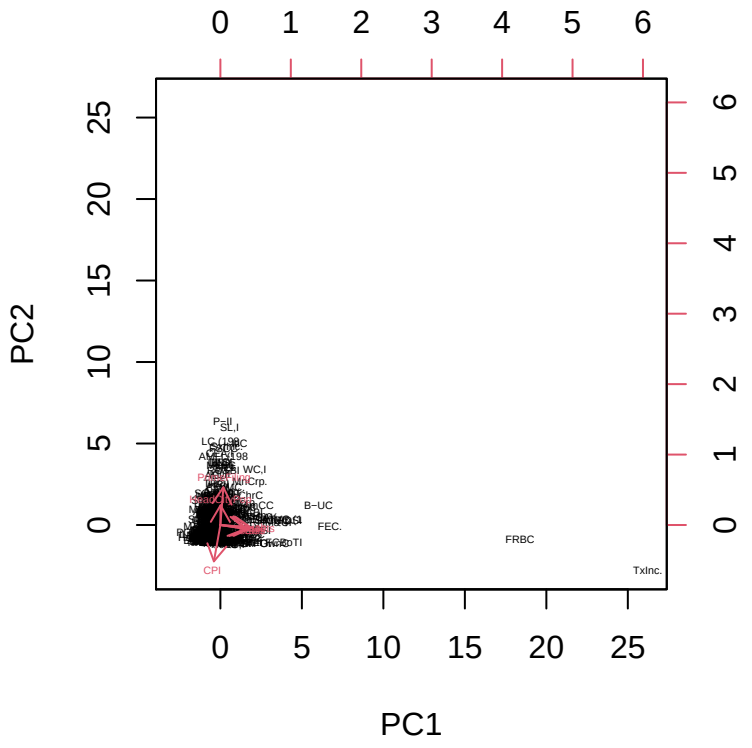


Figure 5: Biplot produced by PCA

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.

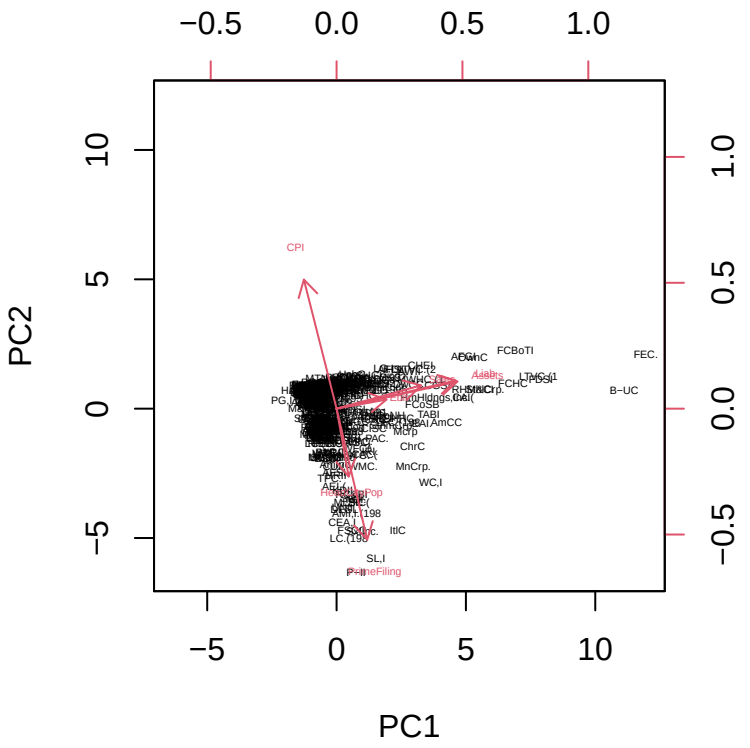


Figure 6: Biplot without Outliers

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.

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.

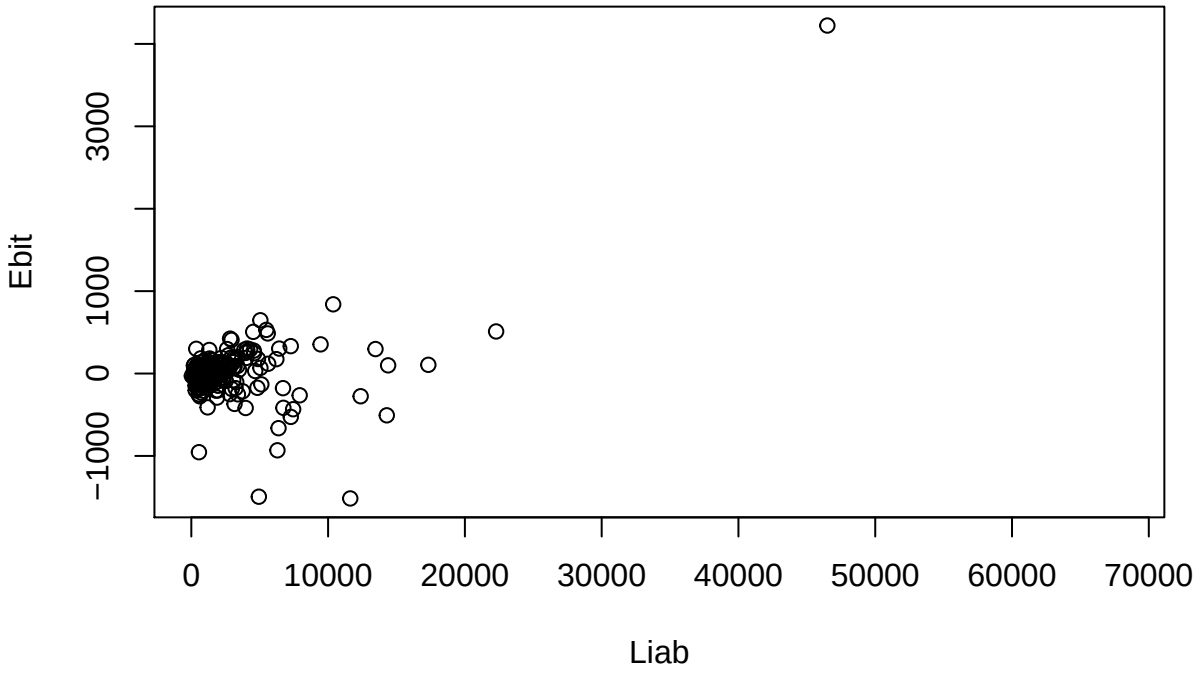


Figure 7: Overview companies financial positions

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(39531.0148)
## * likelihood  =  -19555.9873
```

```
## ****************************************
## *** Cluster 1
## * proportion =  0.0697
## * means      =  444.8634 1.8028 388.4712 424731737.8204 9.5304
## * variances  = | 16393.7088  1113.3886 12756.9950 4169710428.6214    -1.6973 |
##                |  1113.3886   654.3441   644.8033 798529166.3608     0.6306 |
##                | 12756.9950   644.8033 16341.5149 3504299296.0257    -2.7065 |
##                | 4169710428.6214 798529166.3608 3504299296.0257 49080840042678176.0000 -2834598.9282
##                |    -1.6973     0.6306    -2.7065 -2834598.9282     0.0842 |
## *** Cluster 2
## * proportion =  0.1792
## * means      =  1977.9531 17.3978 1867.7600 2244920889.8439 8.5987
## * variances  = | 638960.3818 43395.3806 497216.0070 162518427147.7534   -66.1556 |
##                | 43395.3806 25503.6841 25131.8221 31123433238.3782    24.5786 |
##                | 497216.0070 25131.8221 636926.0762 136583395800.3068  -105.4868 |
##                | 162518427147.7534 31123433238.3782 136583395800.3068 1912972390618420736.0000 -1104
##                |   -66.1556    24.5786  -105.4868 -110481187.4332     3.2816 |
## *** Cluster 3
## * proportion =  0.0979
## * means      =  569.0658 -7.1388 563.0262 671779147.8209 6.1349
## * variances  = | 47493.6040  3225.5568 36957.8159 12079913007.6295    -4.9173 |
##                |  3225.5568  1895.6760  1868.0357 2313389149.8751     1.8269 |
##                | 36957.8159  1868.0357 47342.3951 10152175162.6006    -7.8408 |
##                | 12079913007.6295 2313389149.8751 10152175162.6006 142190276328840064.0000 -8212011.
##                |    -4.9173     1.8269    -7.8408 -8212011.1337     0.2439 |
## *** Cluster 4
## * proportion =  0.2825
## * means      =  839.0797 -11.3512 818.3080 805317004.8214 8.9993
## * variances  = | 141916.1303  9638.3198 110434.0327 36096113229.5075   -14.6935 |
##                |  9638.3198  5664.4892  5581.8969 6912662089.9558     5.4590 |
##                | 110434.0327  5581.8969 141464.3014 30335819799.6606   -23.4291 |
##                | 36096113229.5075 6912662089.9558 30335819799.6606 424880238066214336.0000 -24538395
##                |   -14.6935     5.4590   -23.4291 -24538395.5611     0.7289 |
## *** Cluster 5
## * proportion =  0.1964
## * means      =  466.7569 -2.9649 432.9227 458933144.1245 8.3001
## * variances  = | 30380.8935  2063.3367 23641.3195 7727325767.5128    -3.1455 |
##                |  2063.3367  1212.6334  1194.9524 1479837775.0588     1.1686 |
##                | 23641.3195  1194.9524 30284.1676 6494182920.0855    -5.0156 |
##                | 7727325767.5128 1479837775.0588 6494182920.0855 90956829363891280.0000 -5253091.243
##                |    -3.1455     1.1686    -5.0156 -5253091.2430     0.1560 |
## *** Cluster 6
## * proportion =  0.0247
## * means      =  15947.1558 34.9478 14989.8107 4842553435.5214 8.8816
## * variances  = | 13645161.8133 926719.4138 10618174.5619 3470622434769.7920 -1412.7701 |
##                | 926719.4138 544637.6735 536696.4672 664648849609.9646   524.8819 |
##                | 10618174.5619 536696.4672 13601718.6975 2916773229971.1699 -2252.6973 |
##                | 3470622434769.7920 664648849609.9646 2916773229971.1699 40852012983976706048.0000 -
##                | -1412.7701   524.8819 -2252.6973 -2359353917.3082    70.0790 |
## *** Cluster 7
## * proportion =  0.0323
## * means      =  869.2573 6.6538 727.9918 995773837.6932 15.9920
## * variances  = | 246666.1600 16752.4814 191946.7413 62739095391.4692   -25.5389 |
##                | 16752.4814  9845.5178  9701.9631 12014982430.7454     9.4884 |
```

```
##                | 191946.7413  9701.9631 245880.8306 52727059007.3804   -40.7224 |
##                | 62739095391.4692 12014982430.7454 52727059007.3804 738489532557078272.0000 -4265048
##                |   -25.5389     9.4884   -40.7224 -42650485.1111     1.2668 |
## *** Cluster 8
## * proportion =  0.1172
## * means      =  5000.1154 42.5828 5030.8512 4637464642.8177 9.9316
## * variances  = | 4221559.1065 286709.7389 3285065.5881 1073745987434.9291  -437.0848 |
##                | 286709.7389 168500.7596 166043.8982 205629983881.9196   162.3887 |
##                | 3285065.5881 166043.8982 4208118.6151 902395351497.4901  -696.9426 |
##                | 1073745987434.9291 205629983881.9196 902395351497.4901 12638852495372109824.0000 -7
##                |  -437.0848   162.3887  -696.9426 -729940190.6323    21.6811 |
## ****************************************
```

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 and use Average linkage has a relatively high level of agreement with Ward's method.

```
##   Group.1     Assets        Ebit       Liab       Sales  PrimeFiling
## 1       1 -0.3375478 -0.12391186 -0.3404528 -0.34512348  0.030772303
## 2       2 -0.1163503  0.02039407 -0.1232866 -0.21436442  3.896008438
## 3       3  1.0119377  1.19890019  0.8769185  0.38974202  0.424436476
## 4       4  5.8259844  1.61953970  5.9730896  0.24728684 -0.178310743
## 5       5  0.2521142  0.24375227  0.2749703  1.92080637 -0.012588669
## 6       6  2.9789713  1.20342811  2.8946256  5.64850530  0.009863999
## 7       7 -0.2387320 -0.06469590 -0.1893929 -0.09175094 -1.158387522
## 8       8  1.6705682 -3.92727930  1.9098296  1.86769864  0.174516897
```

7 Multidimensional Scaling

For this part the focus was on findinf a 2D representation of the data that accurately depicts the distanse of the observations in the higher dimensions. The main concerns was how to combineboth 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 multidemsnsional scaling.

Classical multidemsnsional scaling was used, as their was no intication 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 for each observation. To calculate the later, all character variables were transformded to dummy variables (0, 1).

Table 4: Goodnes of Fit Measures

Value
0.6819628
0.6819628

Table 4 suggests that selecting a subest 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.

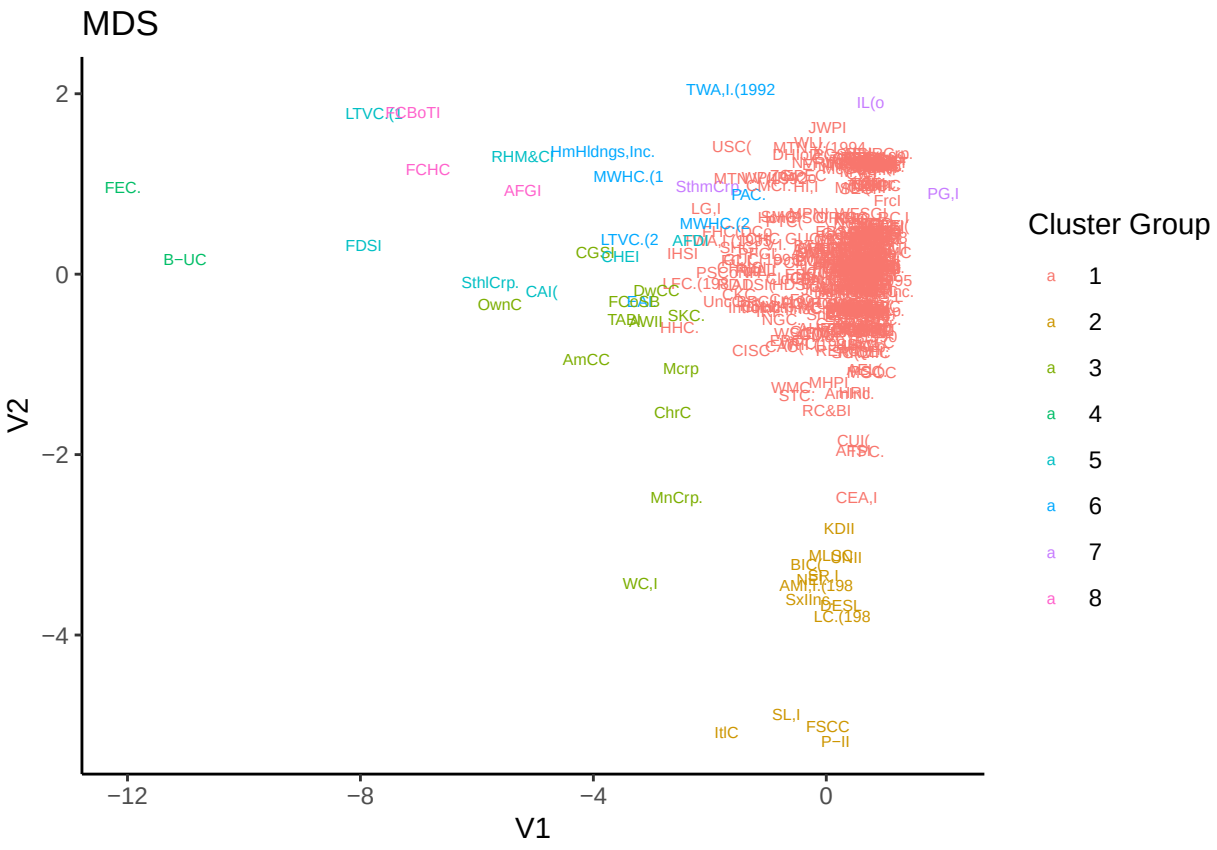
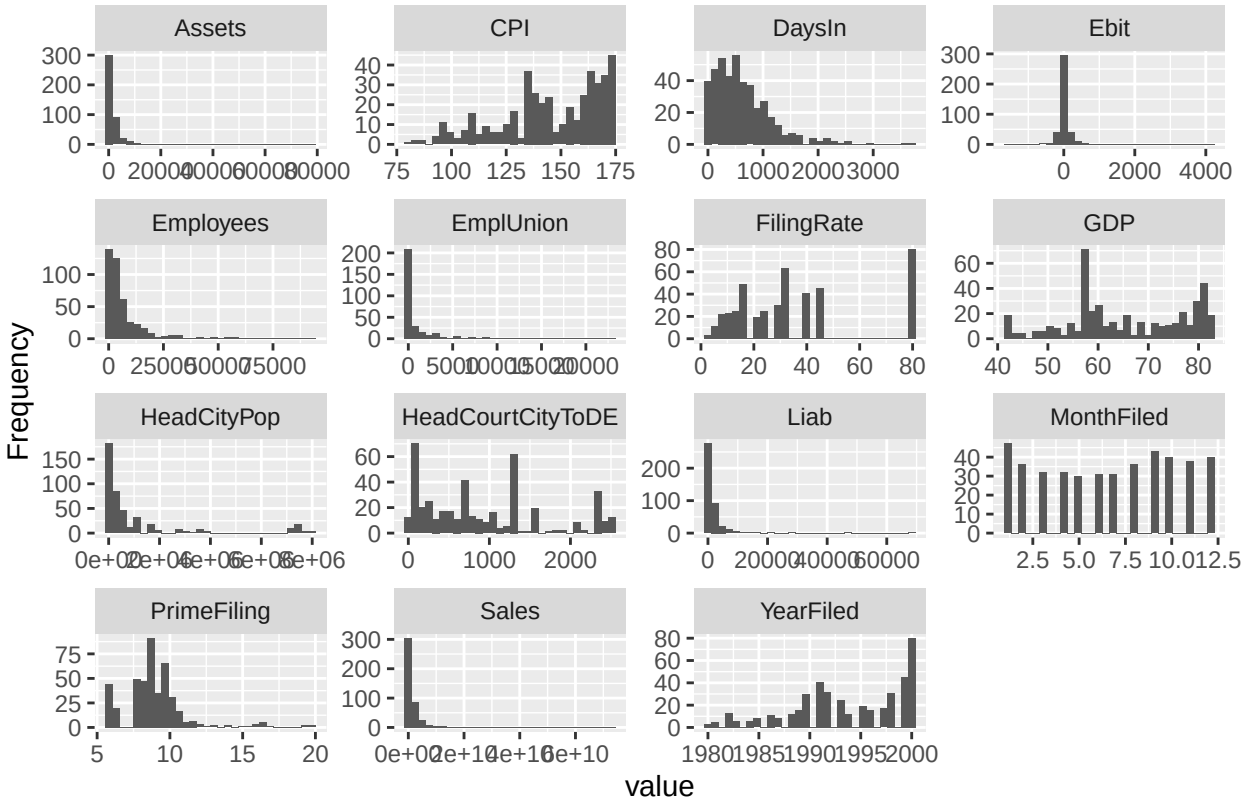


Figure 8: Multidimensional Scaling 2D representation

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.

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.**