A2

Group 10

9/12/2021

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.

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 **??**fig:vismiss) shows the missing values in the data set. The most missing values are in `EmplUnion`; fortunately, this variable is not important.

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.

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 amount of assets, EBIT, and liability would have similar sales in same industry. Therefore, we use `impute_knn()` to impute missing values in `Sales`. Following same logic, we can impute missing values in `Employees` as well.

2.1 Imputation

```
impute_lm(bankruptcy_clean, Liab ~ Assets) %>% # impute ‘Liab‘
  impute_lm(Ebit ~ Assets) %>% # impute ‘Ebit‘
```

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

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

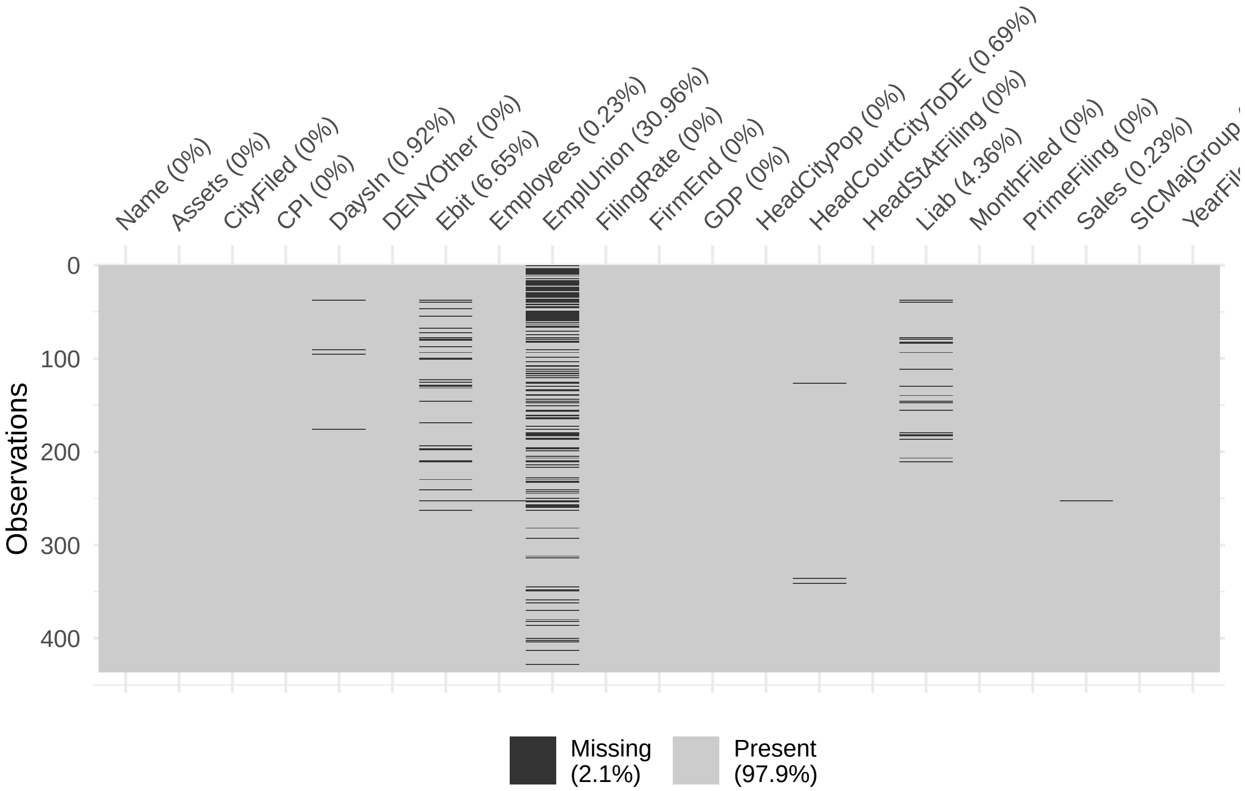


Figure 1: Missing values in the data set

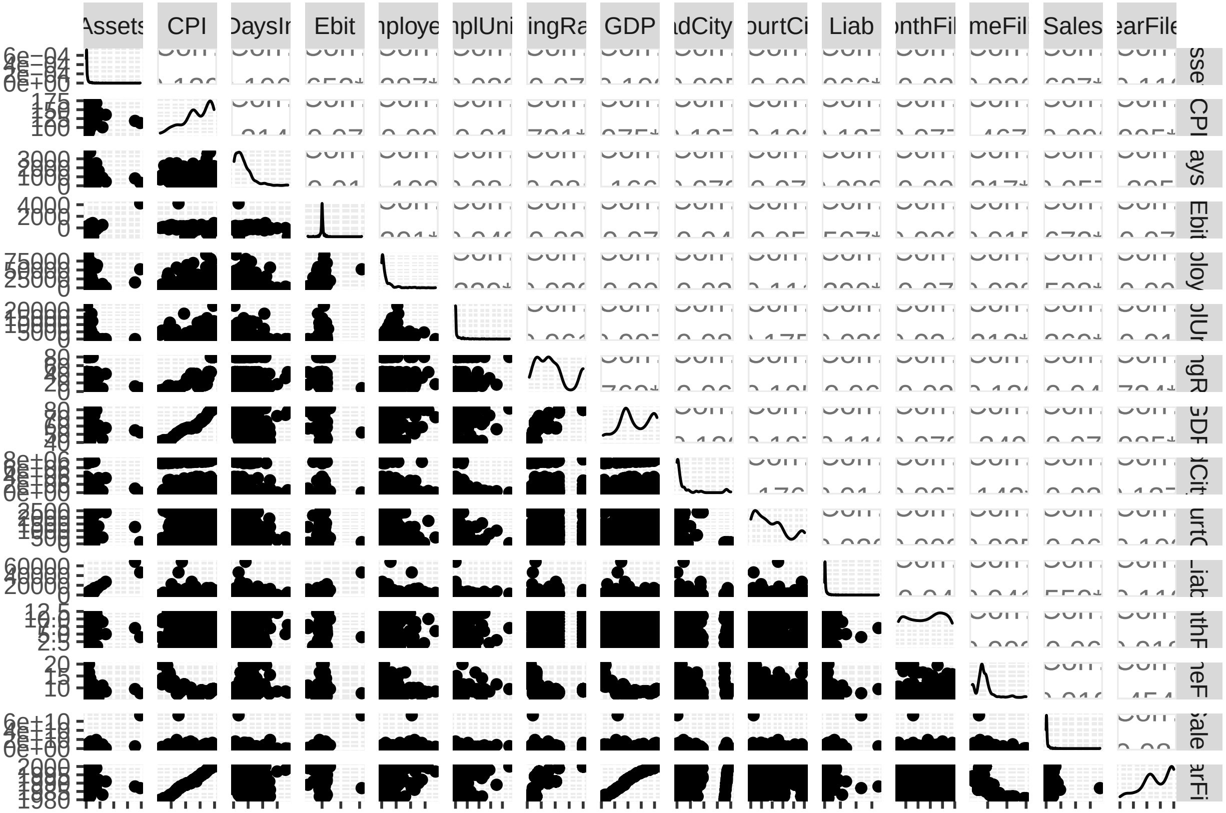


Figure 2: Overview of all numeric variables in the data set

Table 3: Check correlation between factors

	Factor1	Factor2	Factor3
Factor1	1.0000000	-0.0115122	-0.0317599
Factor2	-0.0115122	1.0000000	0.0166920
Factor3	-0.0317599	0.0166920	1.0000000

```
  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 ‘Emplo
  -> bankruptcy_imp
```

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

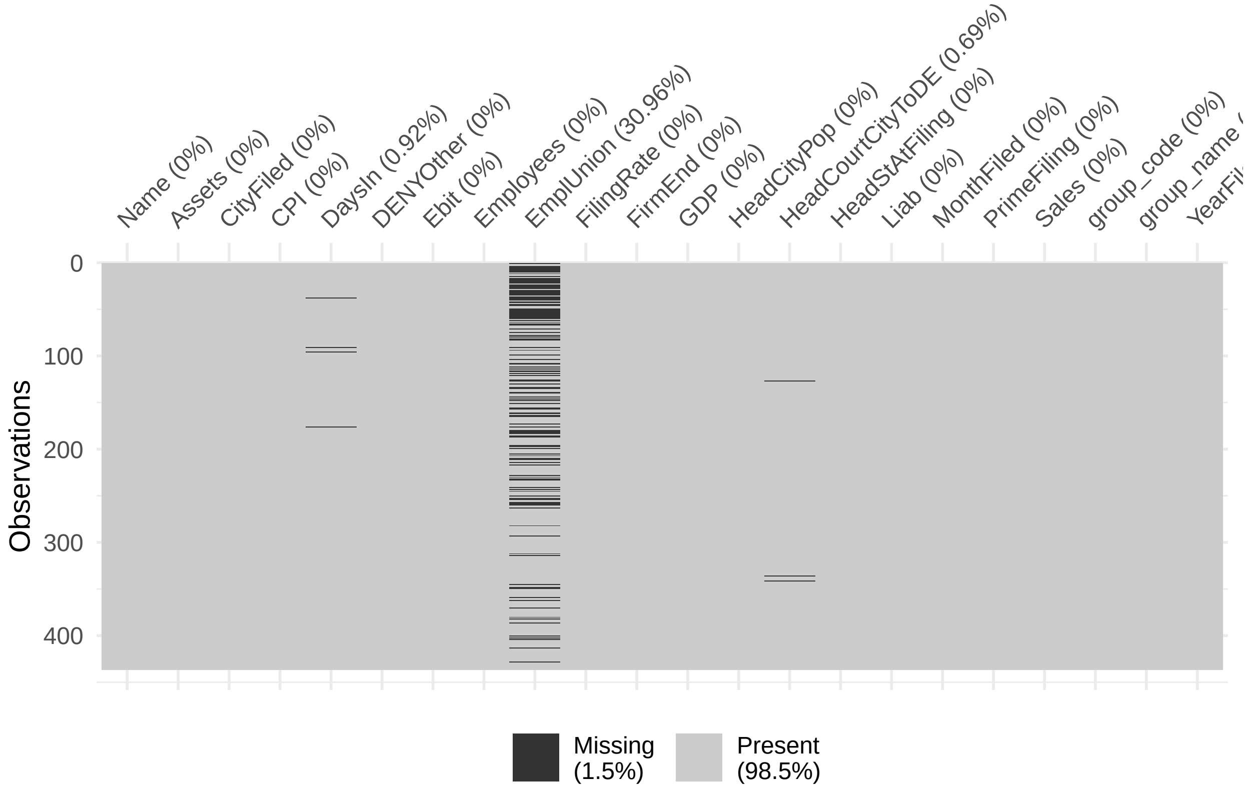


Figure 3: Check missing values after imputation

3 Factor Analaysis

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       0.741       0.031       0.477
##       Sales HeadCityPop
##       0.010       0.959
##
## Loadings:
##             Factor1 Factor2 Factor3
## Assets       0.936   0.334
## CPI                          -0.645
## Ebit         0.649   0.372
## Employees    0.154   0.485
## 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 185.21 on 7 degrees of freedom.
## The p-value is 1.55e-36
```

`Factor 1` has high loadings for `Assets` and `Liab`; it is a company's economies of scale factor with higher score associated with larger scale companies.

`Factor 2` has high loadings for `Sales`; it is sales factor with higher score associated with bigger sales.

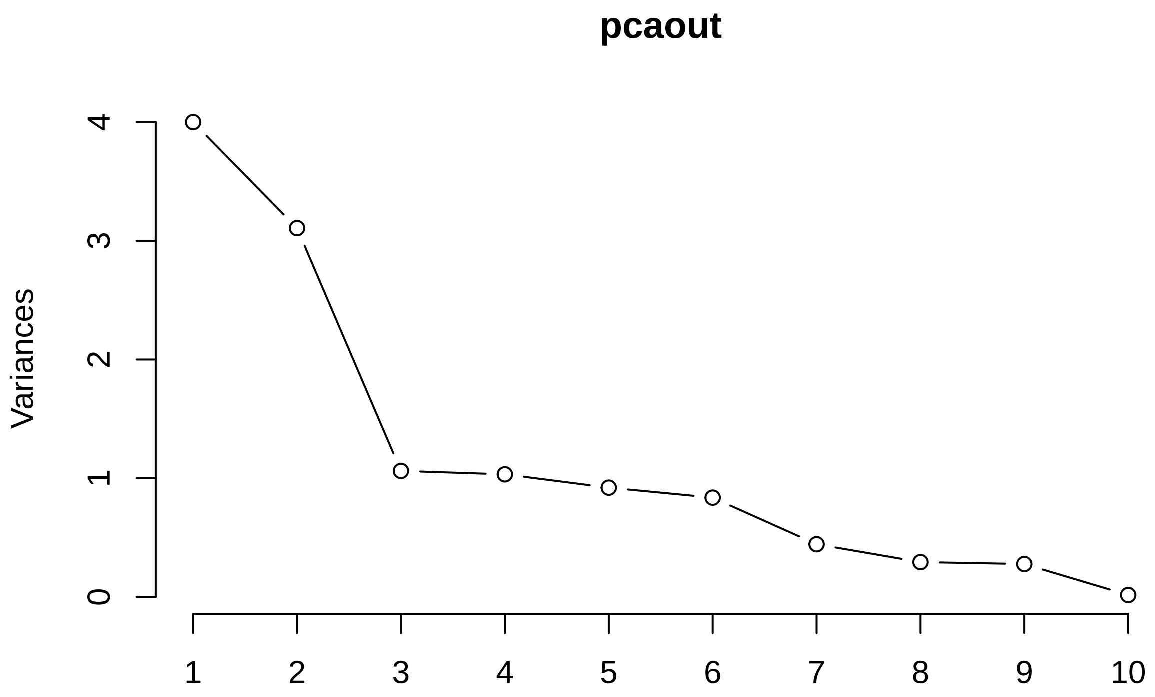
`Factor 3` has high loadings for `PrimeFiling`; it is interest rate of borrowing factor with higher score associated with higher borrowing rate.

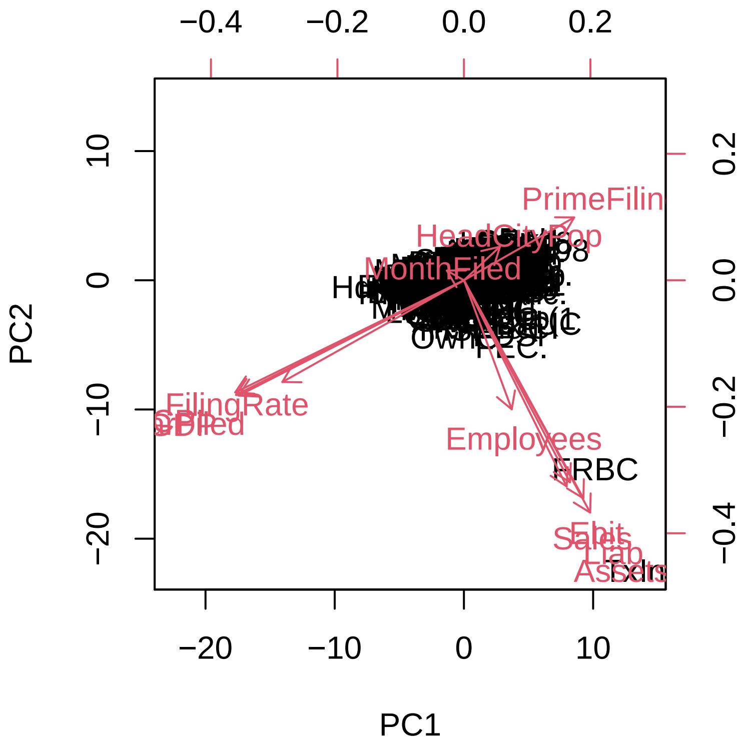
3.1 Limitation of FA

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 PCA

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5    PC6     PC7
## Standard deviation     1.9999 1.7626 1.03034 1.01629 0.95951 0.9145 0.66645
## Proportion of Variance 0.3333 0.2589 0.08847 0.08607 0.07672 0.0697 0.03701
## Cumulative Proportion  0.3333 0.5922 0.68063 0.76670 0.84342 0.9131 0.95013
##                            PC8     PC9    PC10   PC11    PC12
## Standard deviation     0.54191 0.52663 0.12746 0.1038 0.01912
## Proportion of Variance 0.02447 0.02311 0.00135 0.0009 0.00003
## Cumulative Proportion  0.97461 0.99772 0.99907 1.0000 1.00000
```


pcaout
Variances
0
1
2
3
4
1
2
3
4
5
6
7
8
9
10



outliers:

5 Cluster

Why a company fo bankrupt? The most reason will related to their financial position. Accroding to the data, we have company's relevant information are Assets, EBIT, Liabilities and Sales. For Assets is a company owns 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 4 shows the EBIT and liabilities of companies, most companies' EBIT are less than liabilities and even in negative which means they did not have ability to pay the debt which casued bankrupt at the end. There is one company have large amount of liabilities than other companies and for better clustering, we will consider it as an outlier and not use in cluster analysis.

As variable `Sales` amount are larger than other financial positions, we have to normalize them before calculate 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.
## ****************************************
```

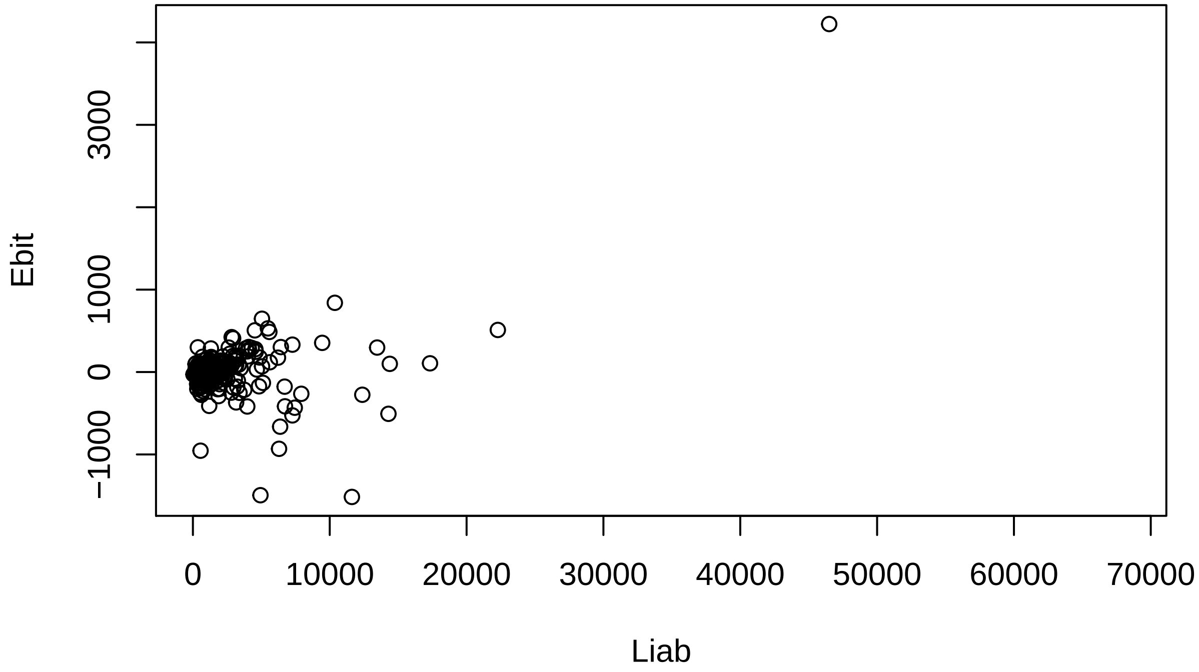


Figure 4: Overview companies financial positions

```
## * 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   =  7
## * model name  =  Gaussian_pk_Lk_C
## * criterion   =  BIC(38930.4015)
## * likelihood  =  -19277.4404
## ****************************************
## *** Cluster 1
## * proportion =  0.1153
## * means      =  4933.6023 50.6938 5011.4526 4726662892.6248 9.9751
## * variances  = | 4098649.6090 283759.8313 3189317.7681 964327863593.3093  -347.5381 |
##                | 283759.8313 156736.5960 173341.2955 177093594606.2131   210.4207 |
##                | 3189317.7681 173341.2955 4035285.2870 785827789719.4180  -821.7513 |
##                | 964327863593.3093 177093594606.2131 785827789719.4180 1231843823125641012.0000 -85
##                |  -347.5381   210.4207  -821.7513 -857843230.8262    33.0191 |
## *** Cluster 2
## * proportion =  0.0914
## * means      =  543.1688 -6.7375 532.1440 631132561.2351 6.1827
## * variances  = | 40549.7183  2807.3591 31553.3040 9540513822.8099    -3.4383 |
##                |  2807.3591  1550.6631  1714.9406 1752063744.1461     2.0818 |
##                | 31553.3040  1714.9406 39922.8276 7774535169.2213    -8.1299 |
##                | 9540513822.8099 1752063744.1461 7774535169.2213 121871652430335168.0000 -8487015.16
##                |    -3.4383     2.0818    -8.1299 -8487015.1641     0.3267 |
## *** Cluster 3
## * proportion =  0.2707
## * means      =  466.1975 -0.0065 427.6266 471814504.5192 8.6490
## * variances  = | 32381.5941  2241.8593 25197.3707 7618722376.0787    -2.7457 |
##                |  2241.8593  1238.3056  1369.4919 1399137142.8997     1.6624 |
##                | 25197.3707  1369.4919 31880.9810 6208473270.6683    -6.4923 |
##                | 7618722376.0787 1399137142.8997 6208473270.6683 97322461098558672.0000 -6777435.0038
##                |    -2.7457     1.6624    -6.4923 -6777435.0038     0.2609 |
## *** Cluster 4
## * proportion =  0.2993
## * means      =  862.4880 -11.5908 836.8442 828079800.3448 8.8640
## * variances  = | 150710.4252 10434.0622 117273.6103 35459059987.3779   -12.7792 |
##                | 10434.0622  5763.3224  6373.8897 6511864513.7173     7.7373 |
##                | 117273.6103  6373.8897 148380.4715 28895478174.3825   -30.2164 |
##                | 35459059987.3779 6511864513.7173 28895478174.3825 452958227884560768.0000 -31543540
##                |   -12.7792     7.7373   -30.2164 -31543540.0449     1.2141 |
## *** Cluster 5
## * proportion =  0.0265
## * means      =  15504.9578 -7.9347 14555.2729 4725921485.4662 8.8731
## * variances  = | 14234918.7911 985519.2660 11076740.8196 3349183300826.2437 -1207.0262 |
##                | 985519.2660 544358.0027 602027.3749 615059392277.9777   730.8069 |
##                | 11076740.8196 602027.3749 14014849.7284 2729239100119.4048 -2854.0042 |
##                | 3349183300826.2437 615059392277.9777 2729239100119.4048 42782863768606121984.0000 -2
##                | -1207.0262   730.8069 -2854.0042 -2979354151.0916   114.6777 |
## *** Cluster 6
## * proportion =  0.1654
## * means      =  2050.8140 22.8470 1927.2376 2307589438.6094 8.6482
## * variances  = | 628824.8175 43535.1253 489312.9091 147949532320.2007   -53.3201 |
##                | 43535.1253 24046.9107 26594.4442 27170131122.4813    32.2833 |
##                | 489312.9091 26594.4442 619103.3087 120563615718.8101  -126.0751 |
##                | 147949532320.2007 27170131122.4813 120563615718.8101 1889924831621666048.0000 -1316
##                |   -53.3201    32.2833  -126.0751 -131612400.3011     5.0659 |
## *** Cluster 7
## * proportion =  0.0315
## * means      =  883.4839 8.2483 734.3618 1014712339.0093 16.1223
## * variances  = | 185915.4533 12871.3949 144668.0042 43742078249.6300   -15.7644 |
##                | 12871.3949  7109.5990  7862.7911 8032996001.9077     9.5447 |
##                | 144668.0042  7862.7911 183041.2367 35645284105.5075   -37.2748 |
##                | 43742078249.6300 8032996001.9077 35645284105.5075 558766483234271552.0000 -38911916
##                |   -15.7644     9.5447   -37.2748 -38911916.9376     1.4978 |
## ****************************************

## **************************************************************
## * Number of samples    =  427
## * Problem dimension    =  5
## **************************************************************
## *       Number of cluster =  7
## *              Model Type =  Gaussian_pk_Lk_C
## *               Criterion =  BIC(38930.4015)
## *              Parameters =  list by cluster
## *                  Cluster  1 :
##                         Proportion =  0.1153
##                              Means =  4933.6023 50.6938 5011.4526 4726662892.6248 9.9751
##                          Variances = | 4098649.6090 283759.8313 3189317.7681 964327863593.3093  -347
##                                      | 283759.8313 156736.5960 173341.2955 177093594606.2131   210.4
##                                       | 3189317.7681 173341.2955 4035285.2870 785827789719.4180  -82
##                                       | 964327863593.3093 177093594606.2131 785827789719.4180 123184
##                                       |  -347.5381   210.4207  -821.7513 -857843230.8262    33.0191
## *                  Cluster  2 :
##                          Proportion =  0.0914
##                               Means =  543.1688 -6.7375 532.1440 631132561.2351 6.1827
##                           Variances = | 40549.7183  2807.3591 31553.3040 9540513822.8099    -3.4383
##                                       |  2807.3591  1550.6631  1714.9406 1752063744.1461     2.0818
##                                       | 31553.3040  1714.9406 39922.8276 7774535169.2213    -8.1299
##                                       | 9540513822.8099 1752063744.1461 7774535169.2213 121871652430
##                                       |    -3.4383     2.0818    -8.1299 -8487015.1641     0.3267 |
## *                  Cluster  3 :
##                          Proportion =  0.2707
##                               Means =  466.1975 -0.0065 427.6266 471814504.5192 8.6490
##                           Variances = | 32381.5941  2241.8593 25197.3707 7618722376.0787    -2.7457
##                                       |  2241.8593  1238.3056  1369.4919 1399137142.8997     1.6624
##                                       | 25197.3707  1369.4919 31880.9810 6208473270.6683    -6.4923
##                                       | 7618722376.0787 1399137142.8997 6208473270.6683 973224610985
##                                       |    -2.7457     1.6624    -6.4923 -6777435.0038     0.2609 |
## *                  Cluster  4 :
##                          Proportion =  0.2993
##                               Means =  862.4880 -11.5908 836.8442 828079800.3448 8.8640
##                           Variances = | 150710.4252 10434.0622 117273.6103 35459059987.3779   -12.779
##                                       | 10434.0622  5763.3224  6373.8897 6511864513.7173     7.7373
##                                       | 117273.6103  6373.8897 148380.4715 28895478174.3825   -30.21
##                                       | 35459059987.3779 6511864513.7173 28895478174.3825 4529582278
##                                       |   -12.7792     7.7373   -30.2164 -31543540.0449     1.2141 |
## *                  Cluster  5 :
##                          Proportion =  0.0265
##                               Means =  15504.9578 -7.9347 14555.2729 4725921485.4662 8.8731
##                           Variances = | 14234918.7911 985519.2660 11076740.8196 3349183300826.2437 -
##                                       | 985519.2660 544358.0027 602027.3749 615059392277.9777   730.
##                                       | 11076740.8196 602027.3749 14014849.7284 2729239100119.4048 -
##                                       | 3349183300826.2437 615059392277.9777 2729239100119.4048 4278
##                                       | -1207.0262   730.8069 -2854.0042 -2979354151.0916   114.6777
## *                  Cluster  6 :
##                          Proportion =  0.1654
##                               Means =  2050.8140 22.8470 1927.2376 2307589438.6094 8.6482
##                           Variances = | 628824.8175 43535.1253 489312.9091 147949532320.2007   -53.3
##                                       | 43535.1253 24046.9107 26594.4442 27170131122.4813    32.2833
##                                       | 489312.9091 26594.4442 619103.3087 120563615718.8101  -126.0
##                                       | 147949532320.2007 27170131122.4813 120563615718.8101 1889924
##                                       |   -53.3201    32.2833  -126.0751 -131612400.3011     5.0659
## *                  Cluster  7 :
##                          Proportion =  0.0315
##                               Means =  883.4839 8.2483 734.3618 1014712339.0093 16.1223
##                           Variances = | 185915.4533 12871.3949 144668.0042 43742078249.6300   -15.76
##                                       | 12871.3949  7109.5990  7862.7911 8032996001.9077     9.5447
##                                       | 144668.0042  7862.7911 183041.2367 35645284105.5075   -37.27
##                                       | 43742078249.6300 8032996001.9077 35645284105.5075 5587664832
##                                       |   -15.7644     9.5447   -37.2748 -38911916.9376     1.4978 |
## *          Log-likelihood =  -19277.4404
## **************************************************************
```

Comparing the BIC, we can found that cluster of 5 is the best cluster group numbers. Then, we are 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.

```
## [1] 0.127923
```

```
## [1] 0.1297993
```

```
## [1] 0.2686493
```

```
##   Group.1     Assets          Ebit       Liab      Sales  PrimeFiling
## 1       1 -0.3162745 -0.0878648860 -0.3255413 -0.3193949  0.022349523
## 2       2 -0.1164325  0.0164230588 -0.1235302 -0.2156145  3.876521739
## 3       3  0.8572370  0.7446031141  0.7944376  1.2544405  0.306521990
## 4       4  6.2874287  2.4429732620  6.2551044  0.4035978 -0.178741358
## 5       5  2.9683878  1.1919011669  2.8793712  5.6094677  0.008553267
## 6       6  1.5128771 -4.3464417256  1.9819587  0.5447019  0.055376924
## 7       7 -0.2557204  0.0002705943 -0.2056853 -0.1828001 -1.282999254
```