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

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.

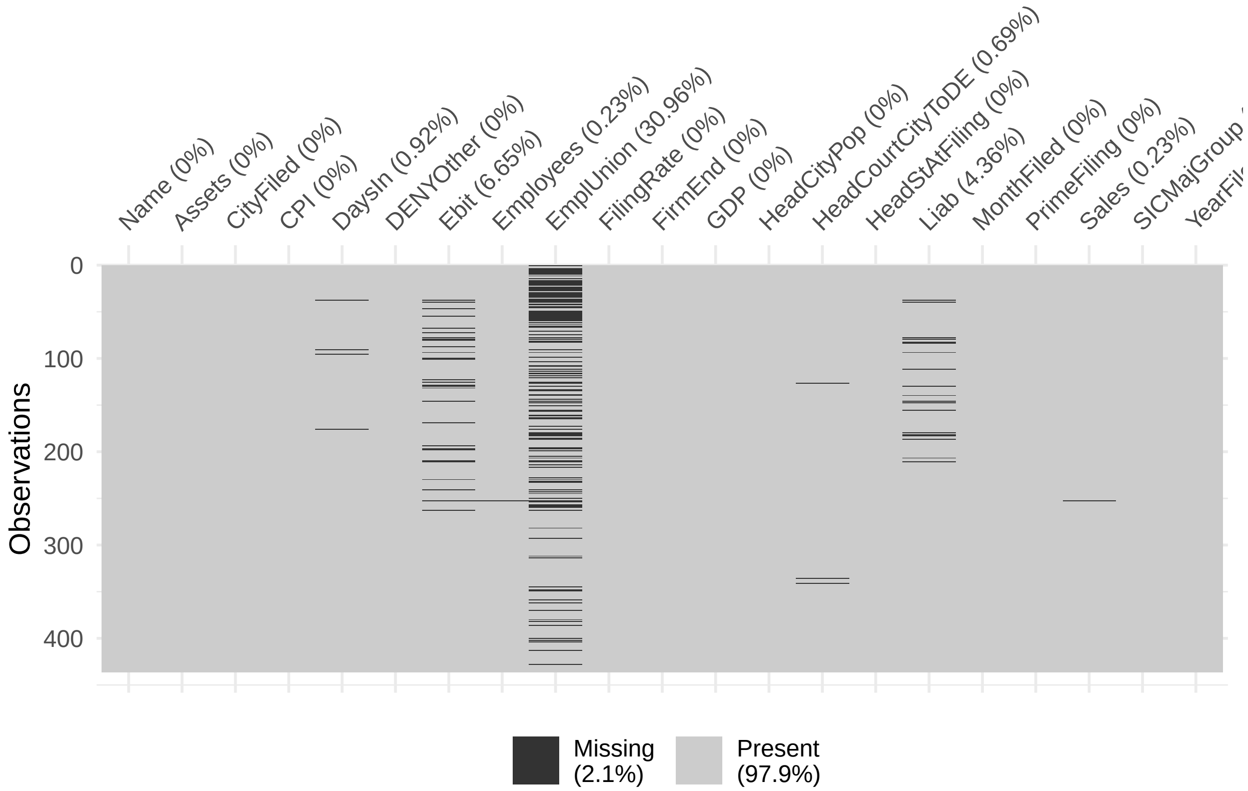


Figure 1: Missing values in the data set

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 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

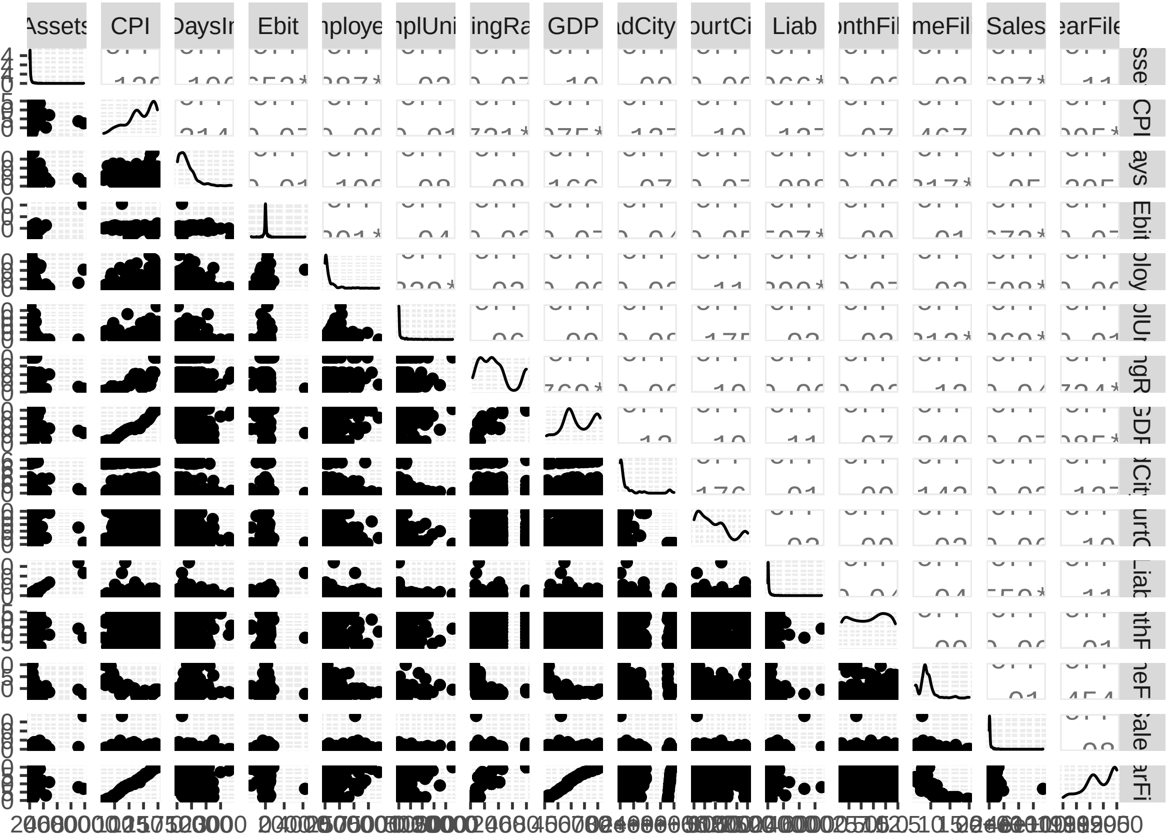


Figure 2: Overview of all numeric variables in the data set

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.

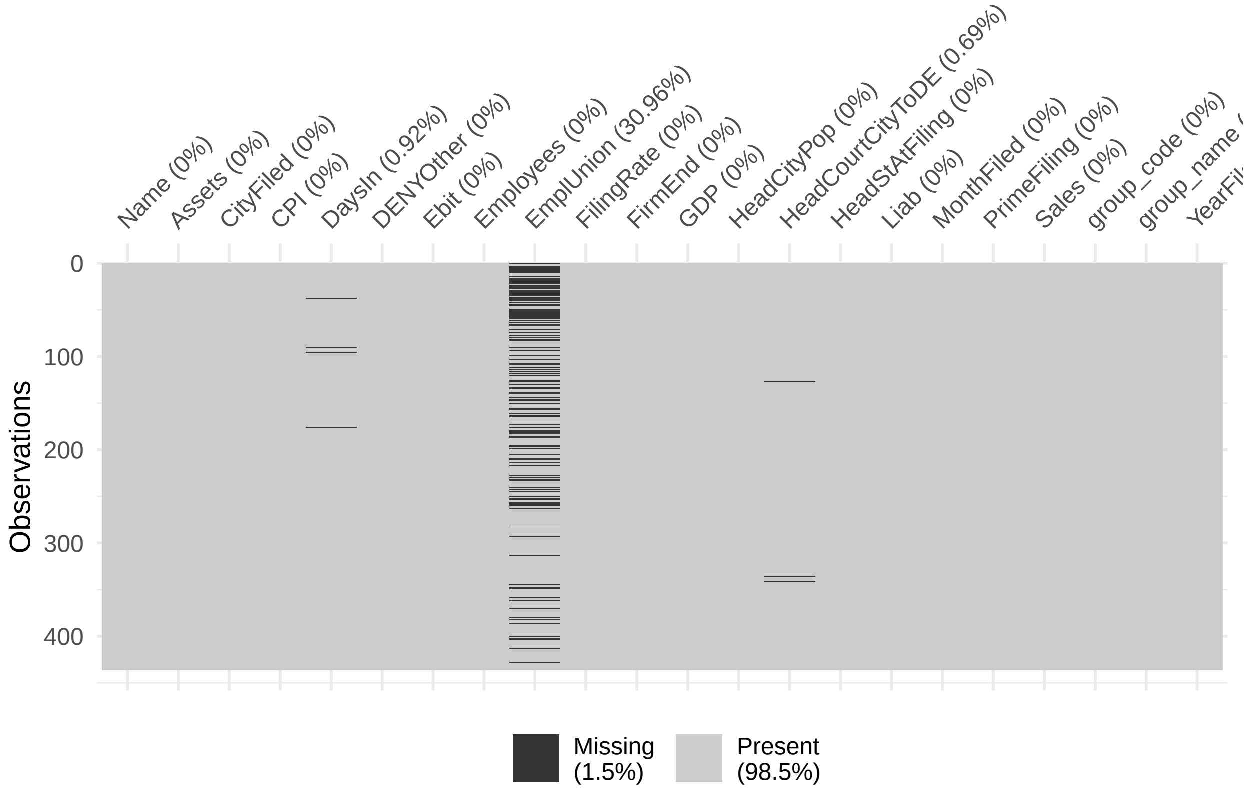


Figure 3: Check missing values after imputation

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

3 Factor Analyasis

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 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

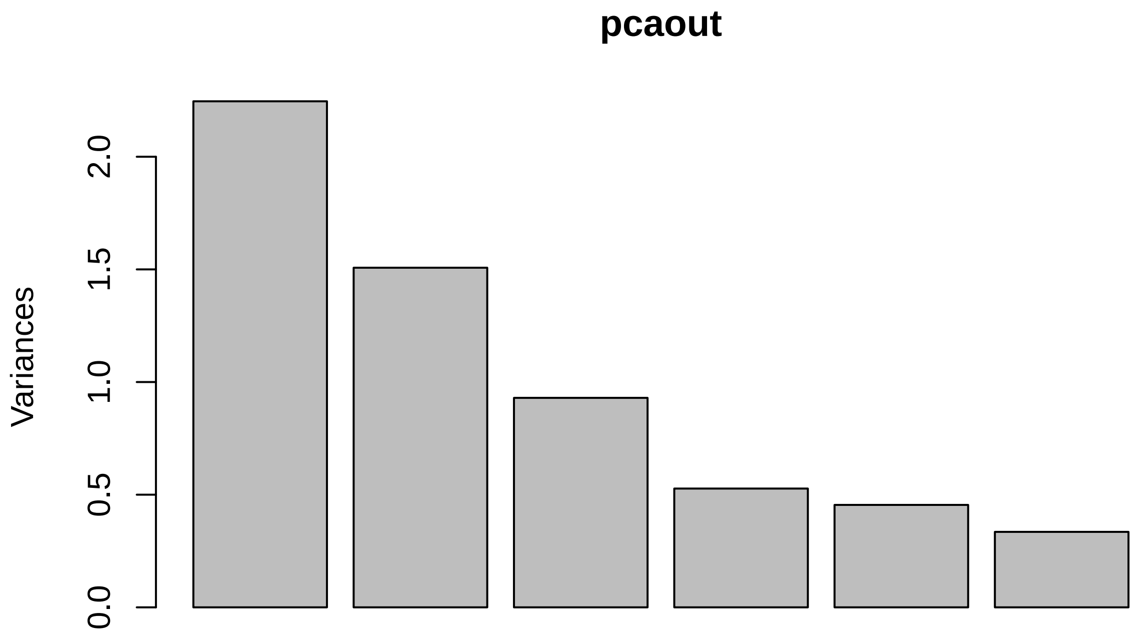
Principal components analysis finds a small number of linear combinations of the orginal variables that explain a large proportion of overall variation in the data. Since the variables in the dataset under investigation are measured in different units we standardise the data by dividing by the standard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

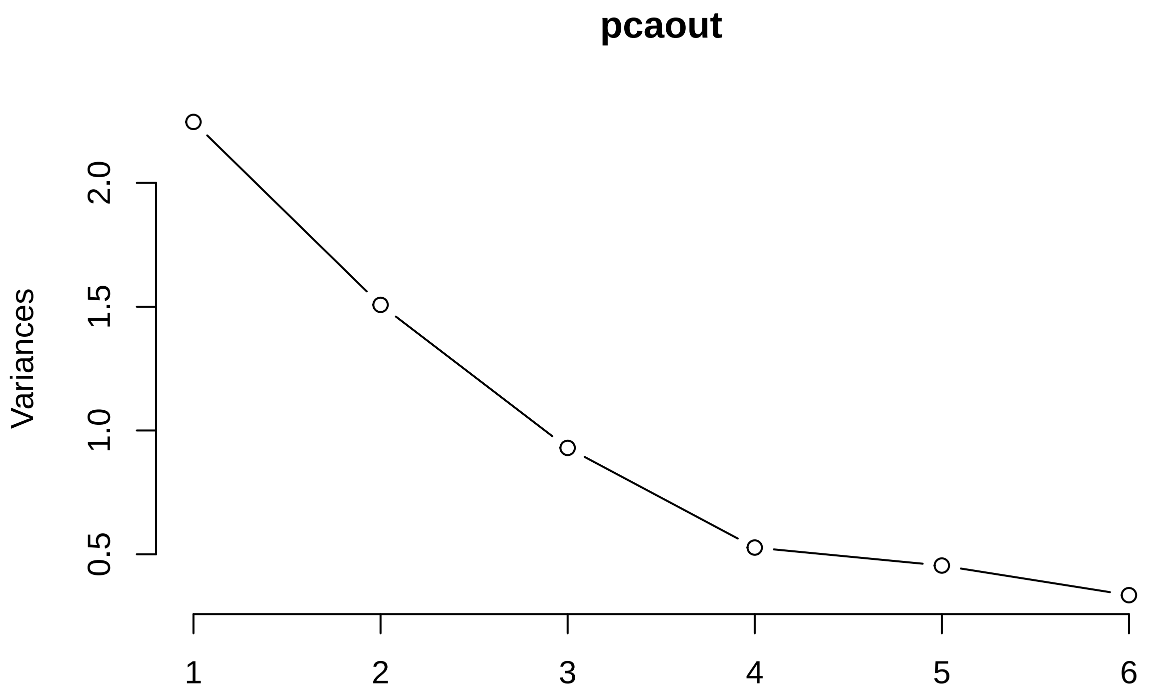
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     1.4987 1.2278 0.9642 0.72607 0.67422 0.57876
## Proportion of Variance 0.3744 0.2512 0.1550 0.08786 0.07576 0.05583
## Cumulative Proportion  0.3744 0.6256 0.7805 0.86841 0.94417 1.00000
```

4.0.1 Proportion of variance explained by the first five PCs together is 52.16%

4.0.2 Proportion of variance explained by the second PC alone is 19.68%

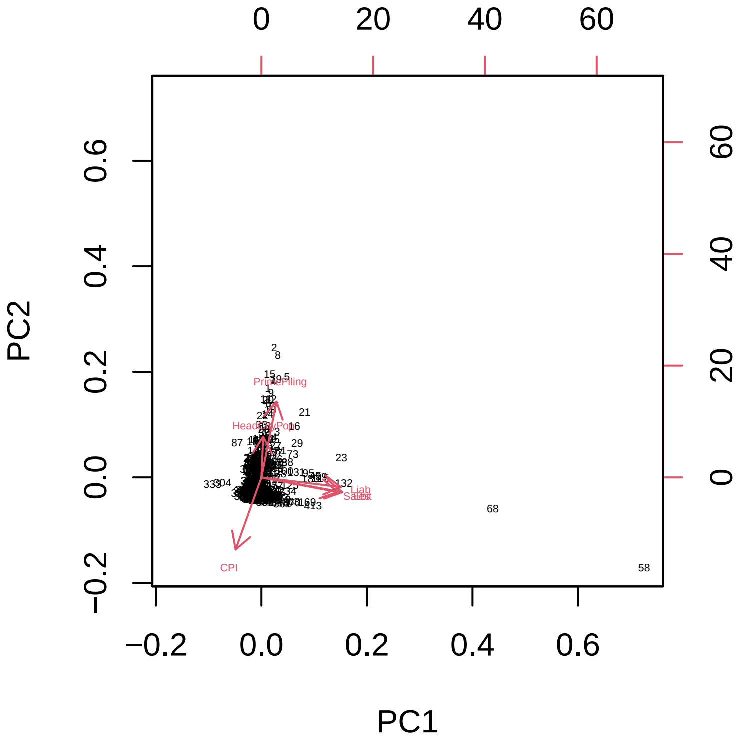
4.0.3 By Kaisers rule select 2 PCs (variance and standard deviation greater than 1.)





```
## integer(0)
```

The elbow appears at the second PC, therefore 2or 3 PCs should be used.



```
## NULL
```

The biplot 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. Those two companies are NGO, so it makes sense that they may be outliers, because any economic index will not effect the NGO.

5 Limitations of the Analysis

Any dimension reduction technique such as principal components analysis represents a loss of information. In this example 0.5216% 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 B-UI 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 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 caused bankruptcy in 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.

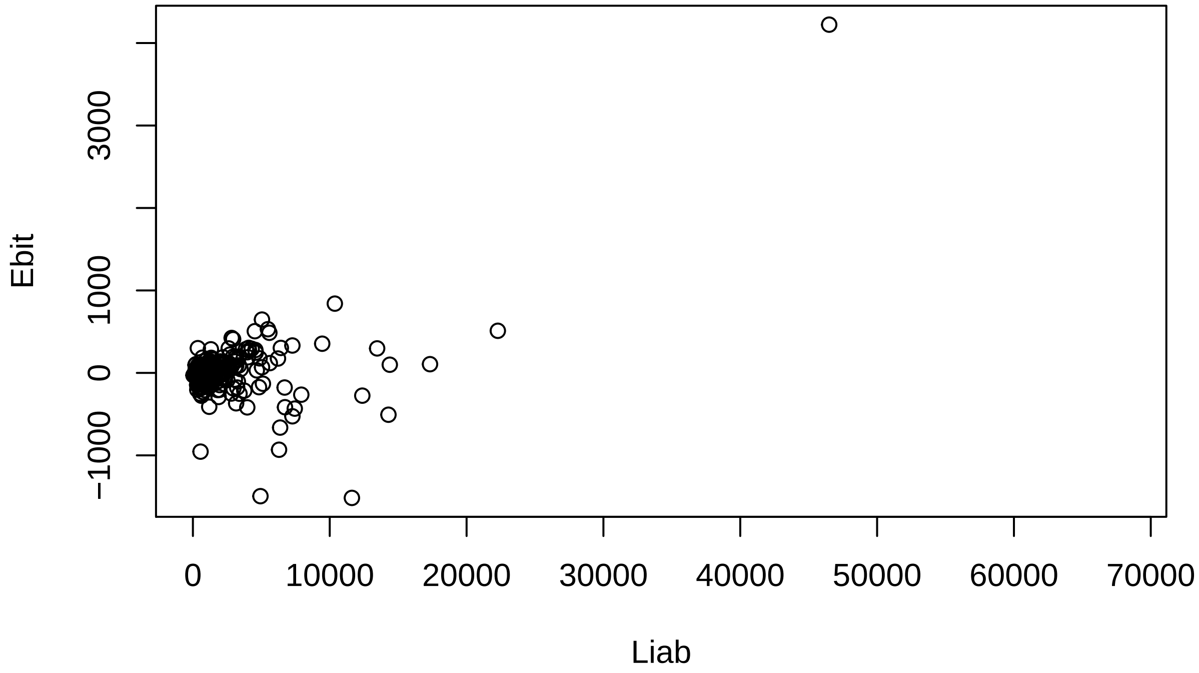


Figure 4: Overview companies financial positions

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

Comparing the BIC, we can find that clusters of 5 are 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
```