What happens to the American Corporation Bankruptcy between 1980 and 2000?

Faculty of Econometrics and Business Statistics
Monash University, Clayton Campus

Team 6 members:

Jason Ching Yuen Siu (31084222)

Liguo Bao (30850894)

Wanxin Liu (30853028)

Zhixiang Yang (30306396)

Date:

19/09/2021

Abstract

This analysis is to examine what factors drive companies in the US to go bankrupt. Regardless of the scale of companies, if businesses are insolvent, one might choose to file a bankruptcy. *Principal Component Analysis* (*PCA*) and *Multidimensional scaling* (*MDS*) are methods that inspect the (dis)similarity data as distances in a low-dimensional space. Therefore, we use these methods to discover the (dis)similarity of the US bankruptcy files.

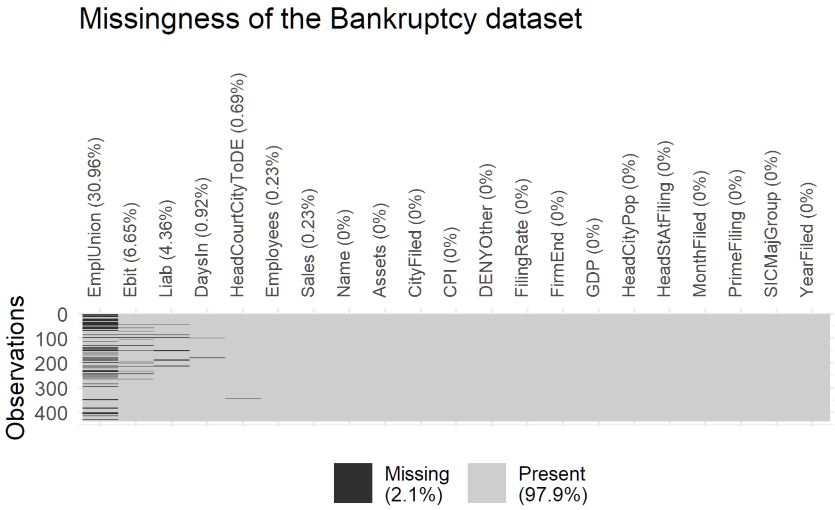
Data description

Sourced from *UCLA-LoPucki Brankruptcy Research Database*, this data is collected on 21 variables each representing different measures of the U.S firms filed for bankruptcy between 1980 and 2000. Details of the features are listed in the appendix.

1. Preliminary Analysis

Before carrying out the analysis, it is worth exploring the features of the original variables.

1. 1 Data cleanness



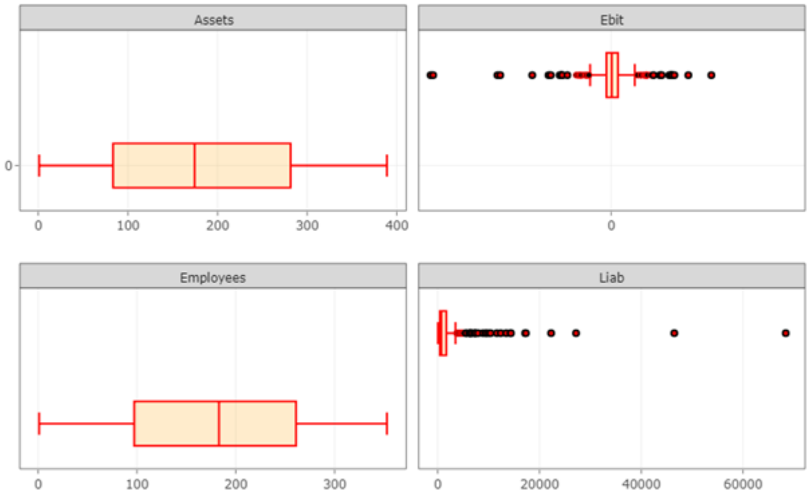
This plot illustrates that the dataset is somewhat clean, with only 2.1 % values missing. However, “EmplUnion” has 30.96 % of NA values which is noticeably high. Furthermore, “FirmEnd” brings little value in this analysis. Therefore, they are eliminated.

Since *MDS* and *PCA* cannot accept values with NA and Character-typed, categorical values are set as ID and missing values are omitted. Both methods only take numerical data into account. In addition, the original variables are measured in different units, so they are standardised.

1.2 Feature transformation

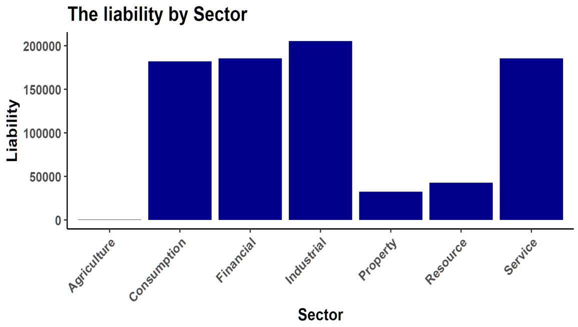
For simplicity, there are 2 features aggregated. Since economy is analysed by quarter, “Monthfiled” is sorted into quarters. Another feature is “SICMajGroup”. Containing as high as 55 unique values, it is difficult to conduct an analysis, hence, they are categorised into 7 sectors.

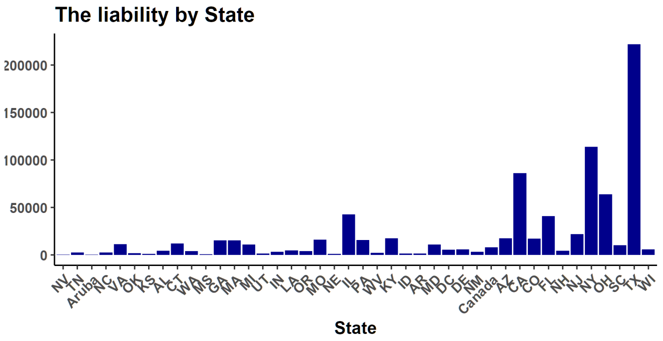
2.1 Financial situation across companies



The boxplot reveals two interesting characteristics. “Assets” and “Employees” follow a Gaussian distribution. Meanwhile, financial values like “Ebit” and “Liab” are right-skewed, suggesting that **the majority of data points are dissimilar**. Also, many outliers occurred in “Ebit” and “Liab”, which means that different companies have different financial situations, many of which have negative “Ebit” though.

3. Liability by states and sectors





The above graphs provide us an interesting characteristic that most values are concentrated into few states and sectors. Each state has its major sector, concluding that a majority of liability is caused by minorities of distinctive pairs. For example, the industrial sector — crucial for **Texas** and **Ohio** — the highest liability. Therefore, the possible conclusion is that many of the liability in the industrial sector comes from these two states. Likewise, this interpretation applies to the services and consumption sector for California and New York.

With the above description, financial instability gives rise to bankruptcy. We suspect that the recession came from two major factors:

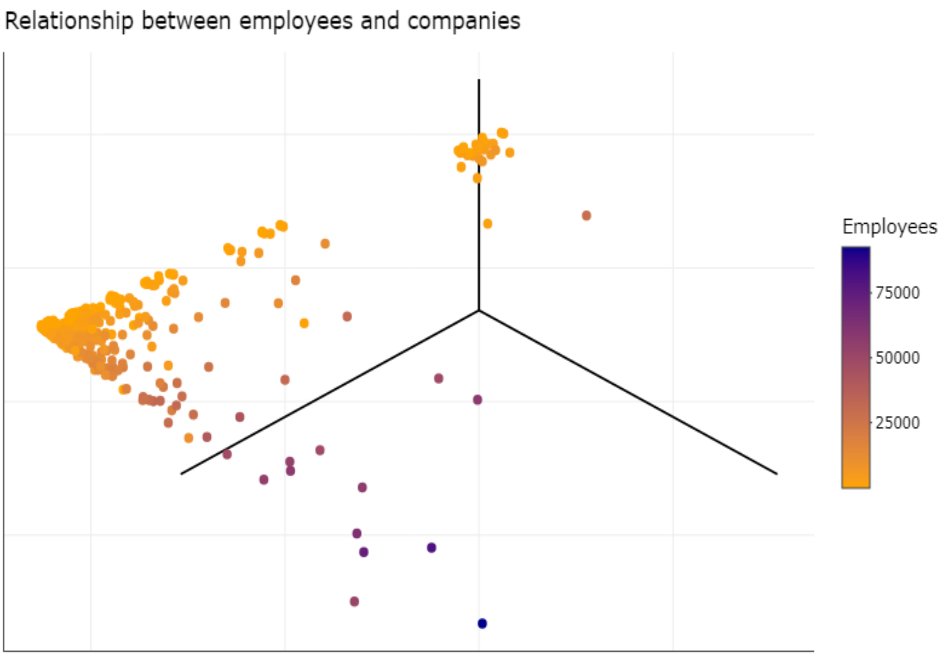
1. *The economics transformation from 1980 to 2000.*
2. *The economic status of the businesses' geolocation*

From below, *PCA* and *MDS* are used to confirm the above hypotheses.

4. Multidimensional scaling (MDS)

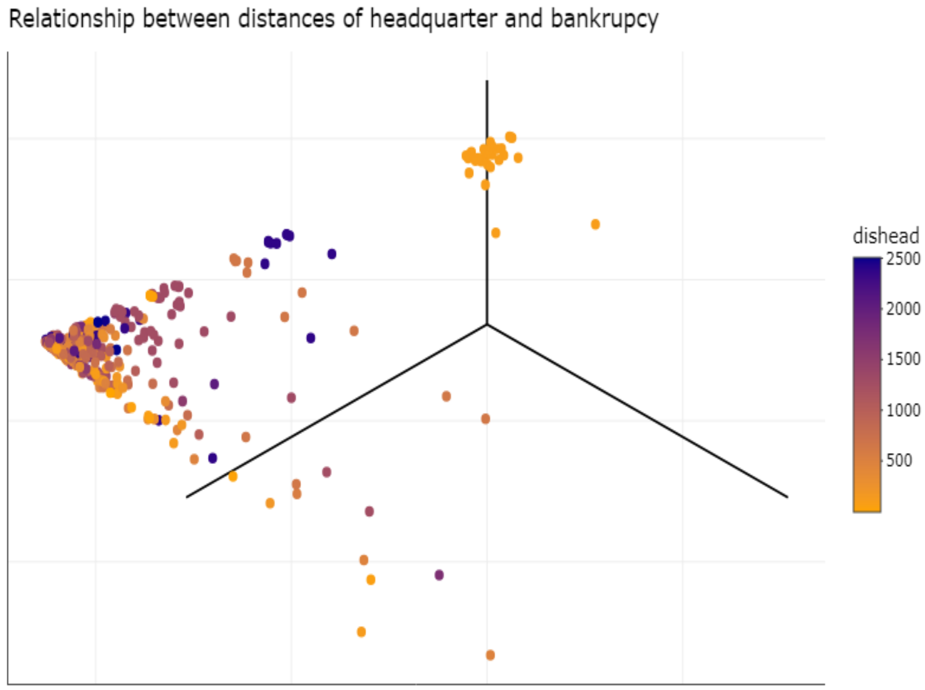
Multidimensional scaling of information about pairwise "distances" among a set of objects or individuals into the configuration of points mapped into an abstract Cartesian coordinate. Here, we first evaluated the Standard Classification into 7 major sectors, listed as :

Sector Name	Financial	Property	Service	Industrial	Consumption	Agriculture	Resource
ID	1	2	3	4	5	6	7



From the classical MDS, we examine the relationship between bankruptcy and number of employees. We can conclude that the majority of the companies have relatively small numbers of employees before bankruptcy.

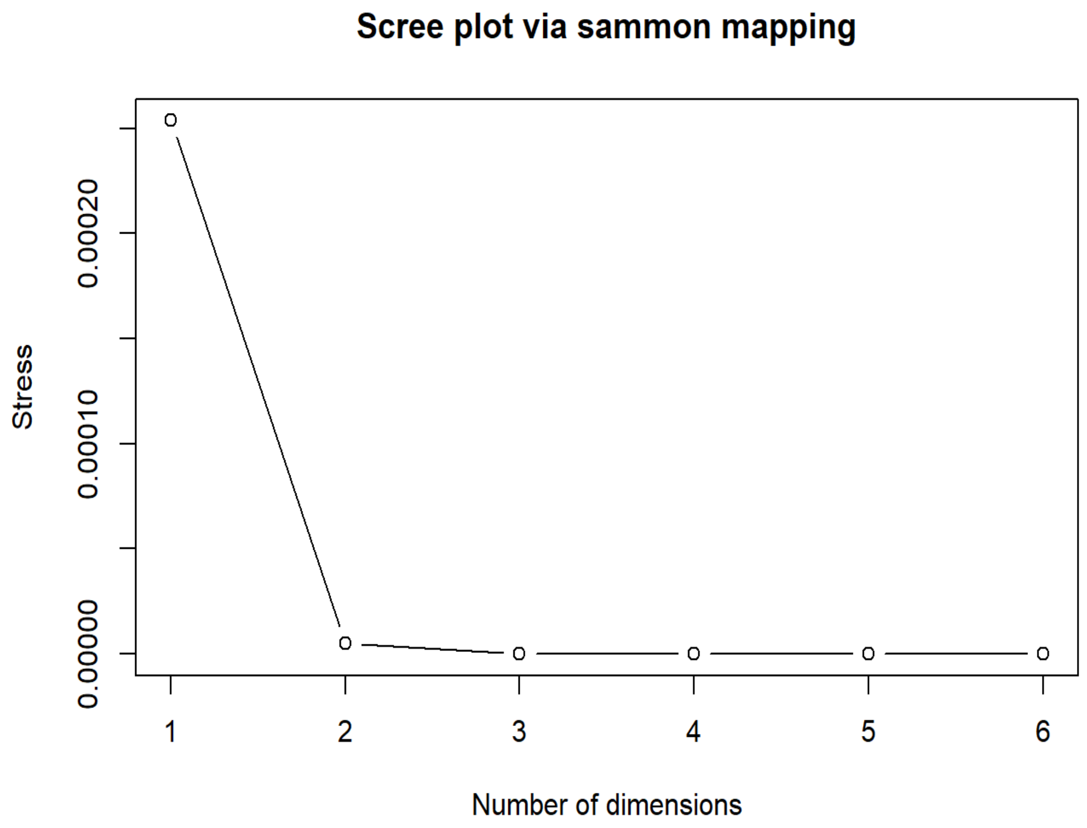
This confirms it because small businesses easily have different issues such as financial hurdles and management issues, which would terminate the businesses in a short time.



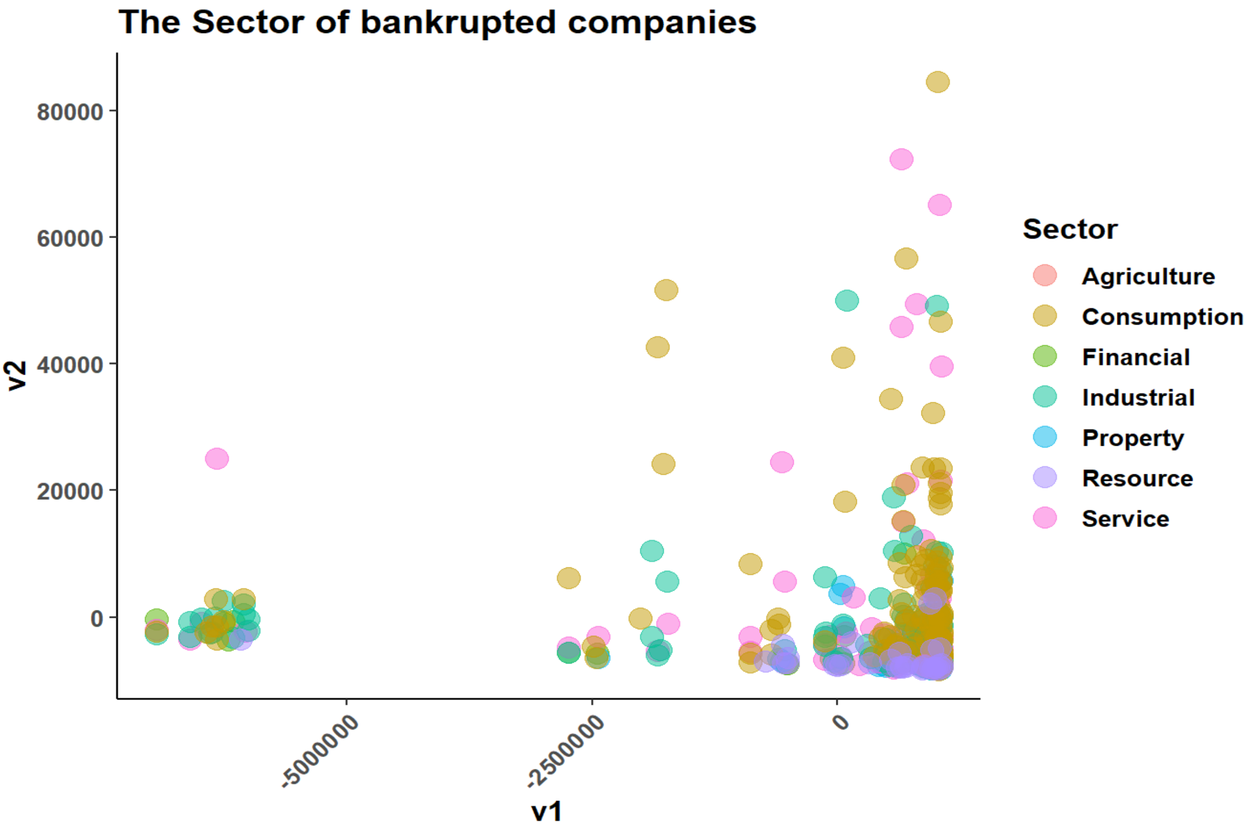
Meanwhile, the above graph displays that majority companies are centralised with different distances between their headquarters, so this proves that the distances between their headquarters have less influence on their bankruptcy.

Distance method	*Goodness of fit 1*	*Goodness of fit 2*
Euclidean distance	0.459	459
Manhattan distance	0.279	0.377

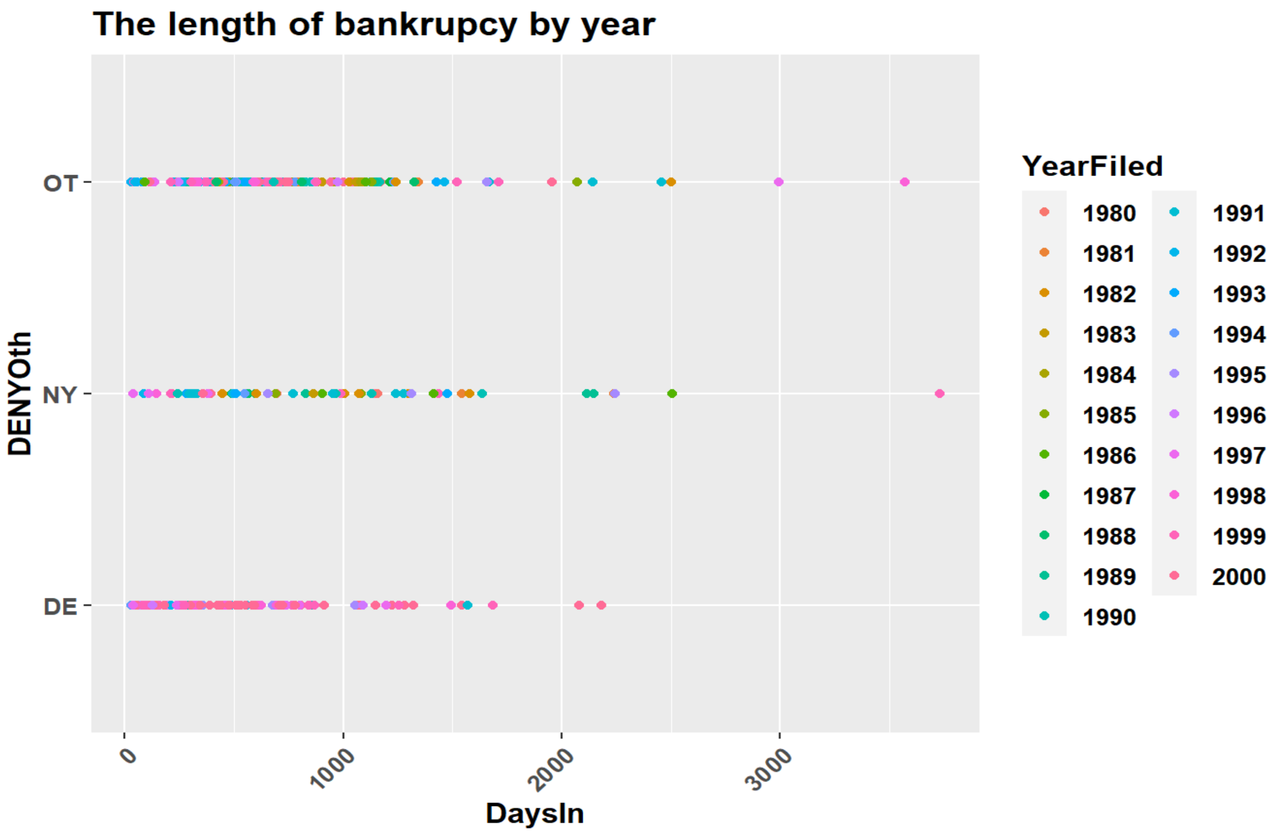
Considering both Euclidean fitnesses are insufficient as they only explain 45.9% of the variation, we conclude that classical MDS may only reflect a part of the data. That said, we try to implement beyond the classical methods to optimise the models.



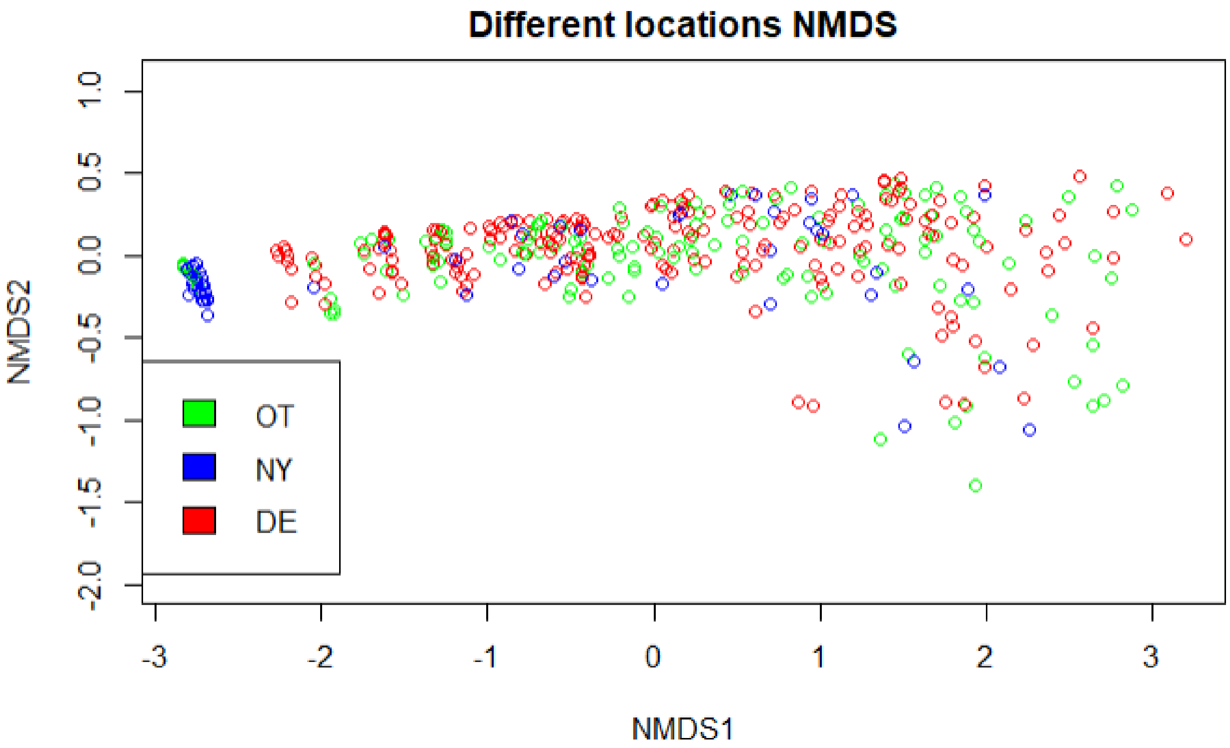
The sammon mapping is a method to project data from high-dimensional space to a lower one while preserving its own structure. Based on the above graph, two dimensions are a good representation for this data.



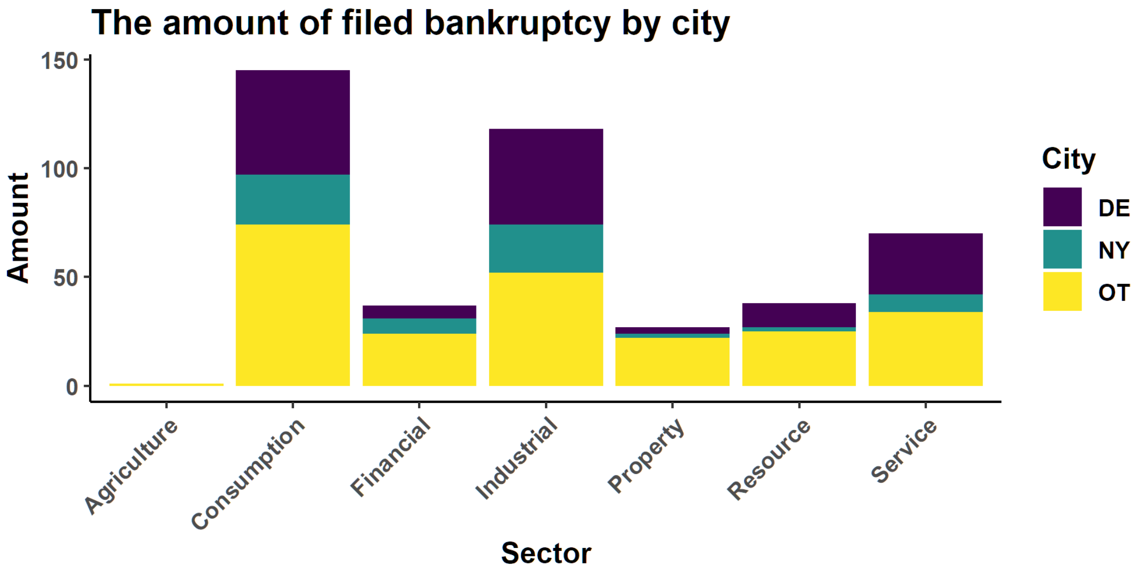
The plot above shows that the sector had a great influence on the bankruptcy of these firms. Among all the 7 sectors, consumption and service are widely spread, which means even if companies are disparate (*displayed as a large distance between them in the MDS graph*), they still go bankrupt. Due to large dissimilarities, it is unlikely that the geographical reasons caused these firms to go bankrupt.



From this plot we can see that companies after 1994, the length of the bankruptcy process has significantly decreased, with the majority of companies being *less than 1000 days*. This is mainly due to the *Bankruptcy Reform Act of 1994*, increasing the efficiency of assessing the bankruptcy and thus shortening the length of the bankruptcy process.



Meanwhile, the non-metric multidimensional scaling (NMDS) shows companies tend to file for bankruptcy in New York and Delaware rather than other states in the US, with the majority of points being blue and red.



And from the second graph, the most favorable sectors are consumption, service and industrial sectors.

The above observation prompts us two questions :

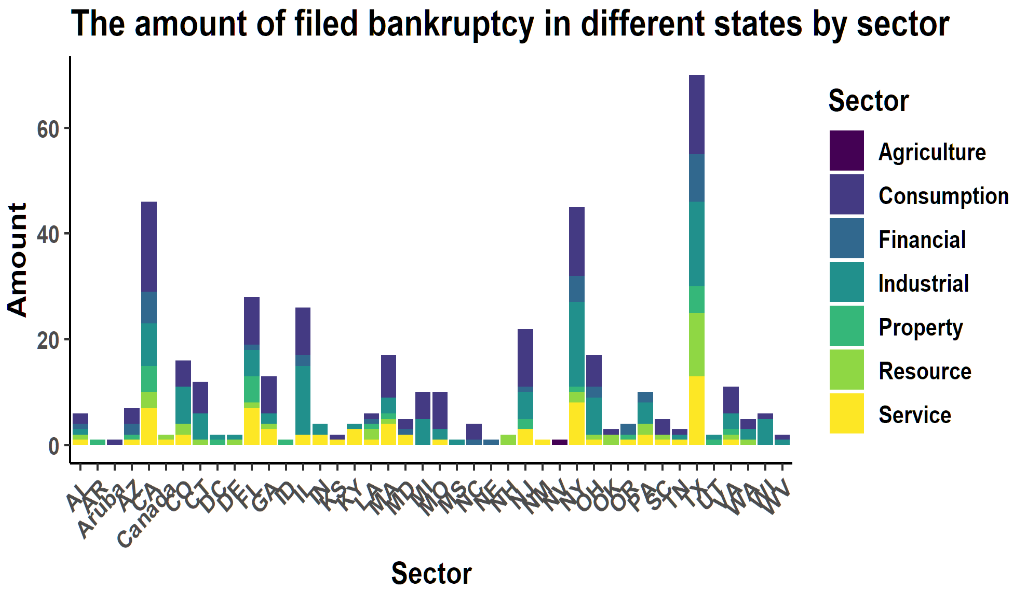
Why were they filed in NY and DE?

According to a paper by LoPucki, L and Doherty J, filing in Delaware and its close competitor New York Division is more likely to survive (Milfold, 2014). Both Delaware and New York have more experienced judges, attracting numerous U.S. companies to file for bankruptcy there.

Another question :

Why did some sectors have such a high number of bankruptcies in NY and DE?

Interestingly, the conclusion we draw is due to the *economic transformation*.



This plot shows the bankrupt companies between 1980 and 2000 are mainly from 3 major traditional sectors, namely *Consumption, Industrial and Service*. This problem happened not only for a few states but ubiquitously for the US. Hence, we draw a conclusion that the shock of the sector triggered the bankruptcy of these companies.

Furthermore, the majority of companies among these sectors are retail or B2C businesses, which easily face financial hardship, management problems and eventually go bankrupt. They lack experience in hiring experienced judges to help them handle bankruptcy so they tend to seek seasoned judges from NY and DE courts to help them.

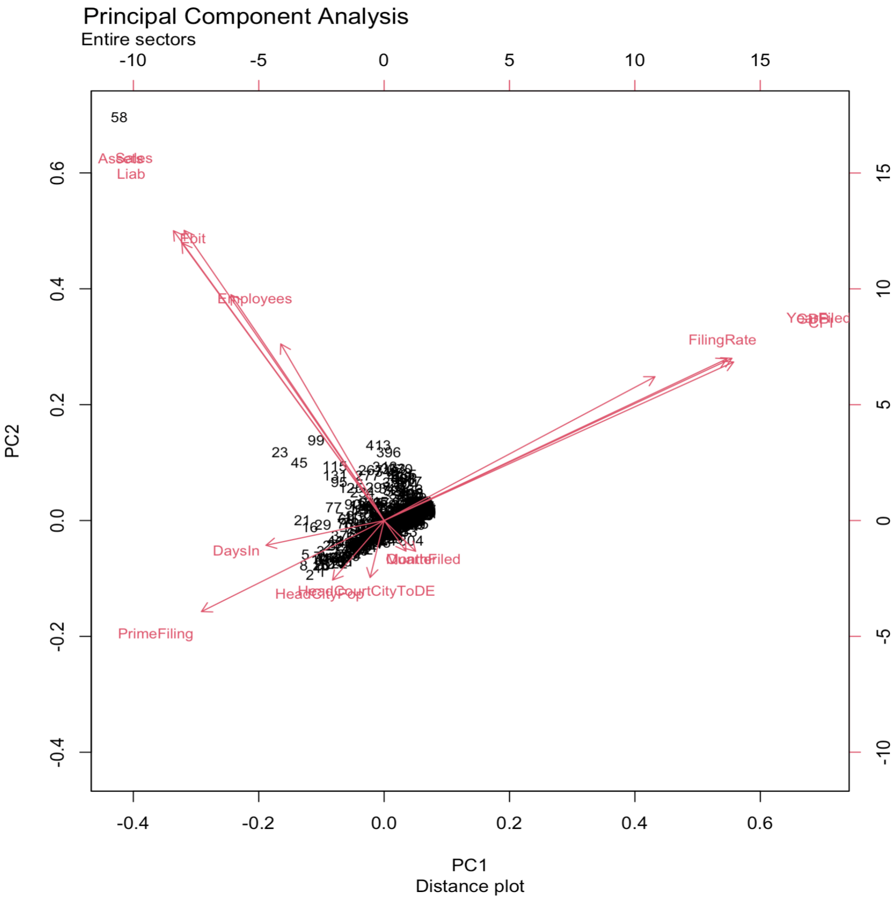
On the other hand, because the size of such companies are small, New York and Delaware could shorten the length of the bankruptcy process within 2000 days.

5. Principal Component Analysis

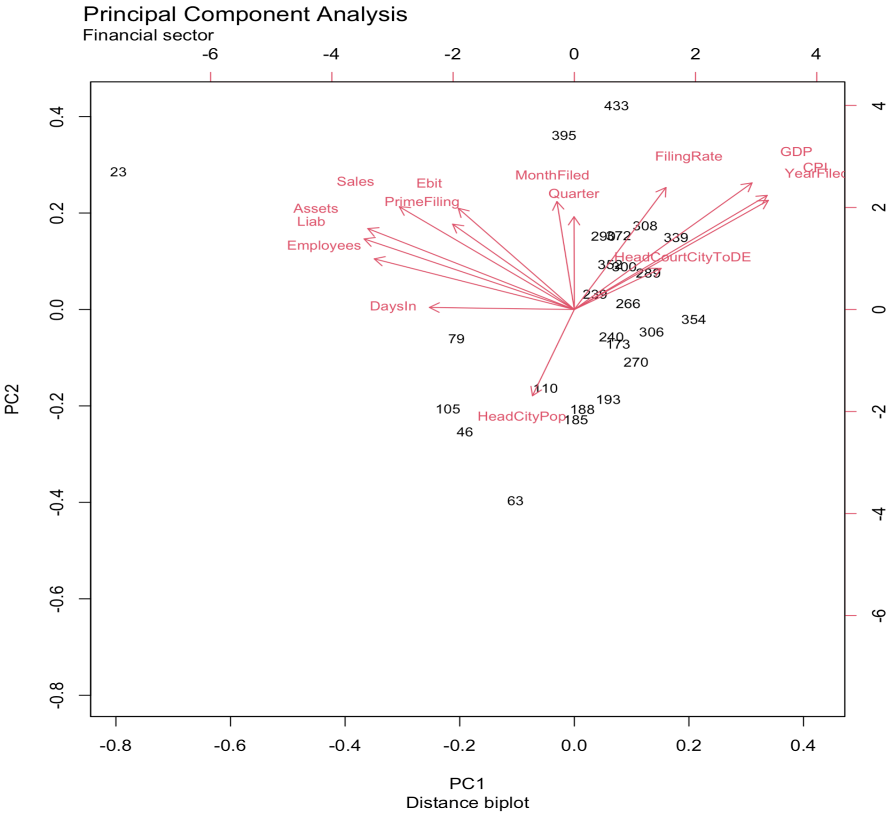
The above MDS analysis represents the relationship between observations. To further explore the relationship between observations and variables, PCA is a sound method since it merges variables through a linear combination.

For the entire PCA analysis, the first principal component is a measure of macroeconomic-related attributes of bankrupt companies, since it is positively correlated with common macro features of improving bankruptcy rate, namely “CPI”, “Yearfield” and “GDP”. Whereas it is negatively correlated with macro features of declining bankruptcy rate, such as PrimeFiling. Likewise, the second principal component is a measure of microeconomic-related attributes of bankrupt companies, such as “Employees”, “Ebit”, “Liab”, “Sales” and “Assets”.

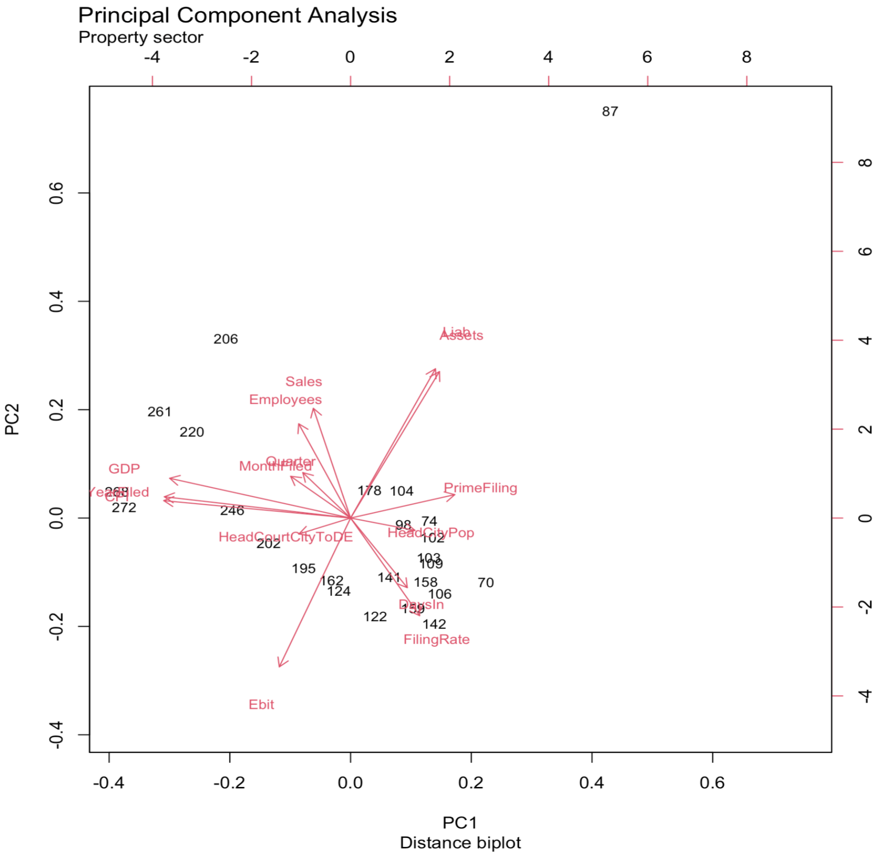
For the entire sector, 58 is a significant outlier called Texaco Inc. — *the world's third-largest oil company, and the biggest firm* ever to seek bankruptcy protection. The jury awarded Pennzoil damages of $10.5 billion plus interest. The PCA plot also confirms that Texaco Inc. has the largest number of assets, liability, earnings and employees.



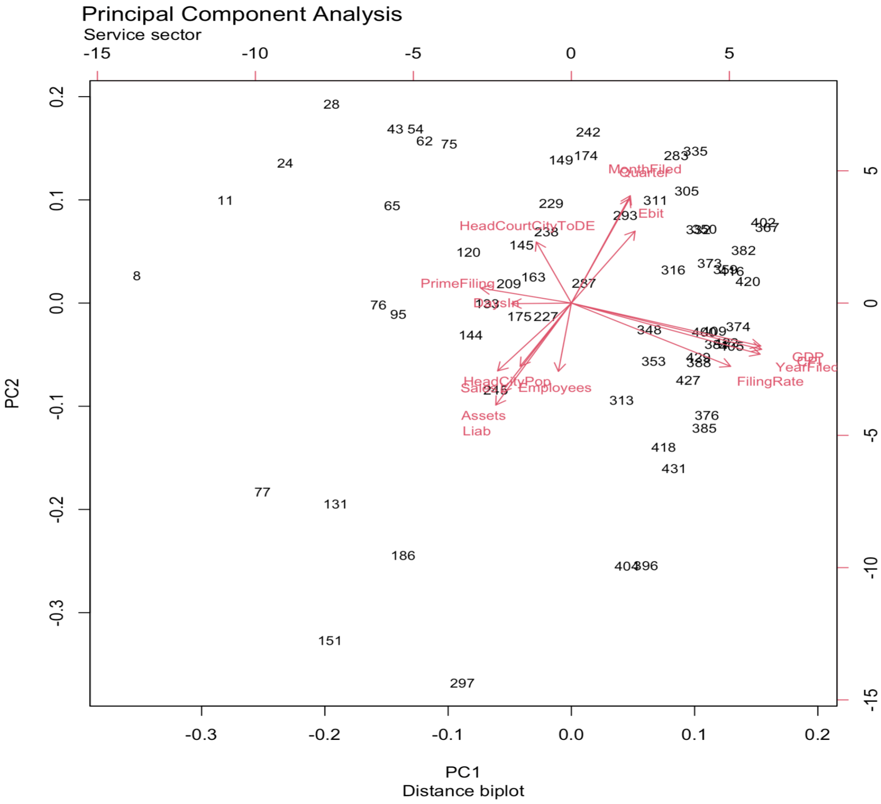
Most observations are concentrated on the same position. It is difficult to analyze and draw conclusions, so transferring to carry out PCA in different sectors.



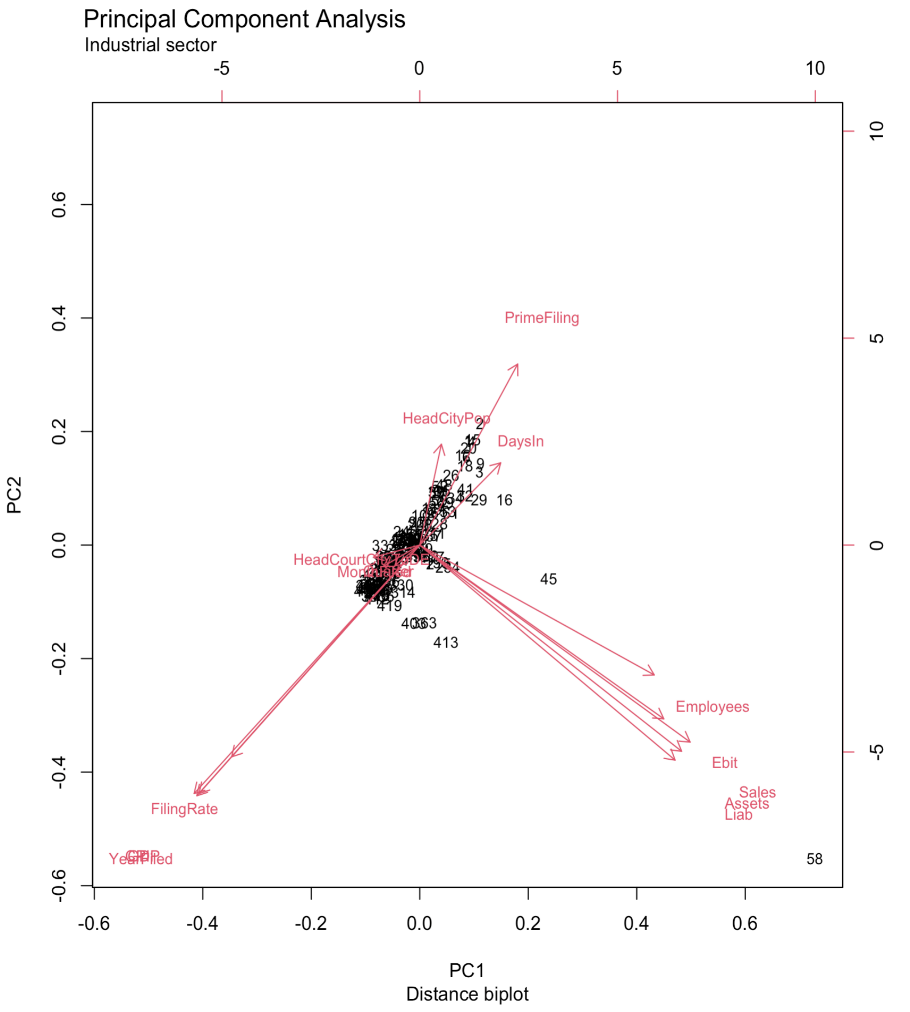
In the financial sector, companies are separated into three parts. The first part is centered on relatively higher “GDP”, “CPI”, “YearField”, “FillingRate”, like 339. The second part is focused on lower “employees”, “liab”, “assets”, “primefiling”, “sales”, like 270. The third part has higher “headCityPop” such as 63. Therefore, we infer that cases in financial sectors are sensitive to changes in macroeconomic factors.



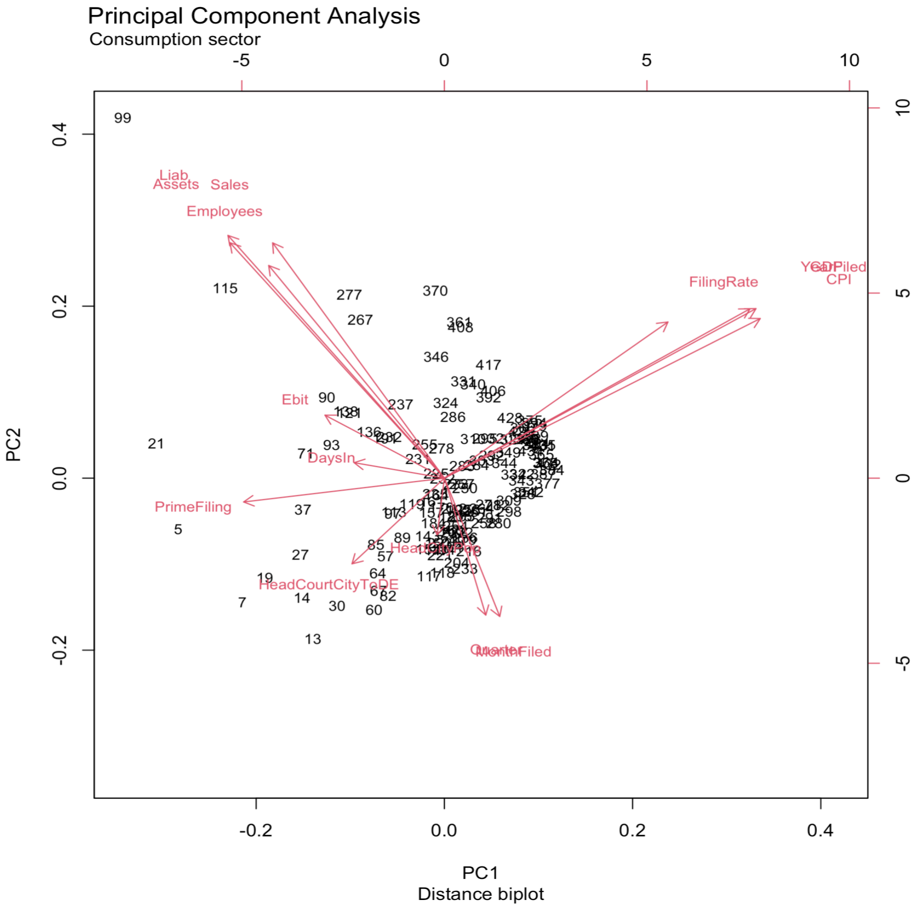
In the property sector, companies are relatively separated. Every variable tends to have more or less effect on companies. No. 87 is the outlier, which is Southmark Corp. , having the highest "Liab", "Assets", lowest "Ebit".



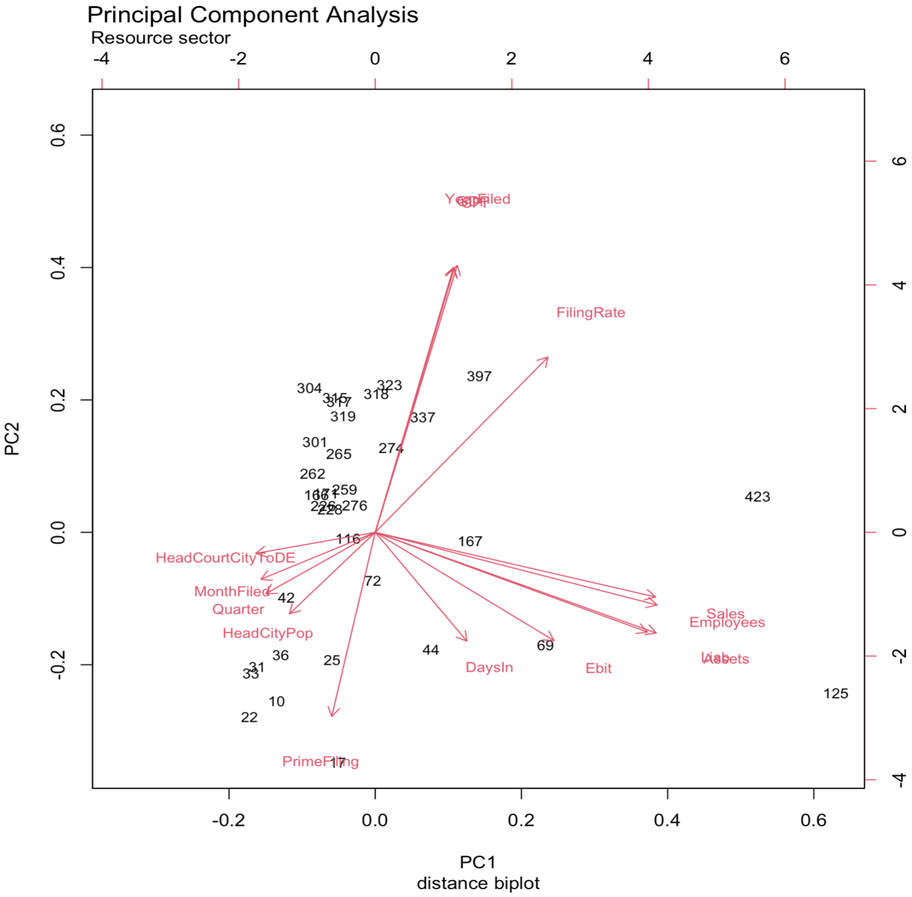
In the service sector, most companies are scattered, corresponding to the above MDS analysis.



In the industrial sector, almost all companies are influenced by the same variables, especially concentrated on the "HeadCityPop", "HeadCourtCityToDE" arrows. Therefore, we infer that most industrial companies are likely to be affected by geographical factors.

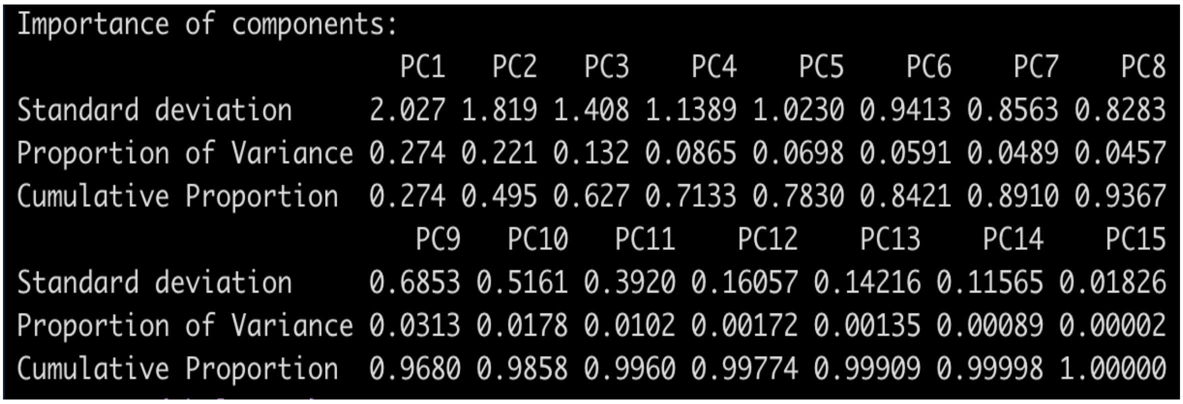


In the consumption sector, large parts concentrate on the arrow of “HeadCityPop”, indicating that almost half headquarters of consumption companies in the cities have relatively large populations. Population might be an indicator for consumption companies to establish headquarters.

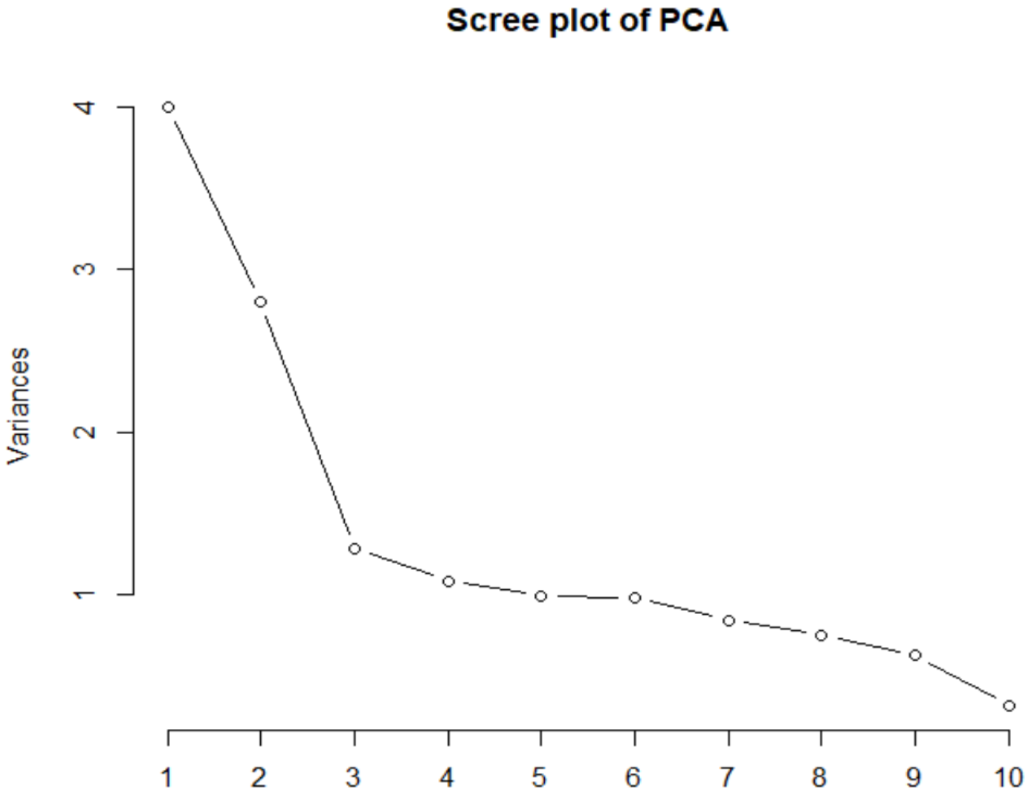


In the resource sector, companies are dispersed in the biplot, most of which are on the positive side of PC2 while the variables of "YearFiled", "GDP" and "FilingRate" contribute to positive PC2.

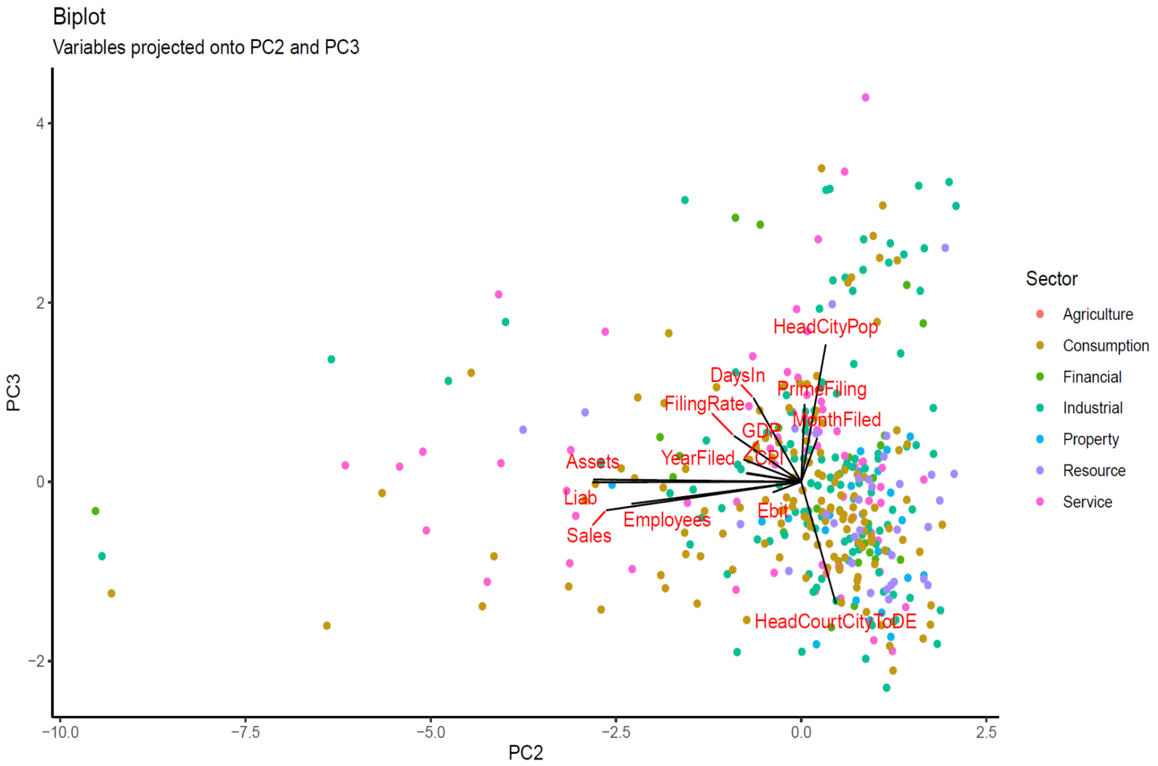
From the scree plot, the elbow point is at the PC3 with about 61% variation of the data so we may use PC3 to analyze the data further.



```
Importance of components:
                          PC1   PC2   PC3    PC4    PC5    PC6    PC7    PC8
Standard deviation     2.027 1.819 1.408 1.1389 1.0230 0.9413 0.8563 0.8283
Proportion of Variance 0.274 0.221 0.132 0.0865 0.0698 0.0591 0.0489 0.0457
Cumulative Proportion  0.274 0.495 0.627 0.7133 0.7830 0.8421 0.8910 0.9367
                          PC9   PC10   PC11    PC12    PC13    PC14    PC15
Standard deviation     0.6853 0.5161 0.3920 0.16057 0.14216 0.11565 0.01826
Proportion of Variance 0.0313 0.0178 0.0102 0.00172 0.00135 0.00089 0.00002
Cumulative Proportion  0.9680 0.9858 0.9960 0.99774 0.99909 0.99998 1.00000
```



Subsequently, we used PC3 to analyse further and removed the outlier of Texaco since its value would largely influence the value of loadings. The difficult interpretation of a 3D graph motives us to project PC3 and PC2 in two dimensions.



In this biplot, the colors of points distinguish different sectors. Most companies from the sector of consumption, property and resource mainly spread in the negative side of PC3 so those companies show the characteristics of relatively low population in headquarter city, low prime rate in loan, shorter process of bankruptcy and far from Wilmington.

6. Limitation

1. It is relatively subjective to divide each observation into different sectors. Maybe it will lead to inaccuracy.

2. Through subjective division, there is just one observation in the agriculture sector. The sample size of the agriculture sector is too small to analyse.

3. For MDS and PCA, we drop out all missing values. It directly loses more than 13% observations, which is also an issue.

4. PCA analysis demonstrates a loss of information. Although we use the first three PCs to reduce dimension, it just explains 62.7% overall variations.

7. Conclusion

Granted the limitation, it is concluded that the hypothesis that **the bankruptcy around these 20 years is caused by businesses' geolocation** was rejected. But rather, it is **the economic transformation** that triggered the rise of bankruptcy.

Reference and Citation

Milfold, M., 2014. Companies turn to Delaware to survive bankruptcy. [online] Delawareonline.com. Available at: <https://www.delawareonline.com/story/money/business/2014/09/19/companies-turn-delaware-survive-bankruptcy/15891887/> [Accessed 18 September 2021].

Citation of packages

1. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

2. Daniel Acker (2021). gg3D: 3D perspective plots for ggplot2. R package version 0.0.0.9000.

3. Duncan Murdoch and Daniel Adler (2021). rgl: 3D Visualization Using OpenGL. R package version 0.107.14. https://CRAN.R-project.org/package=rgl

4. Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

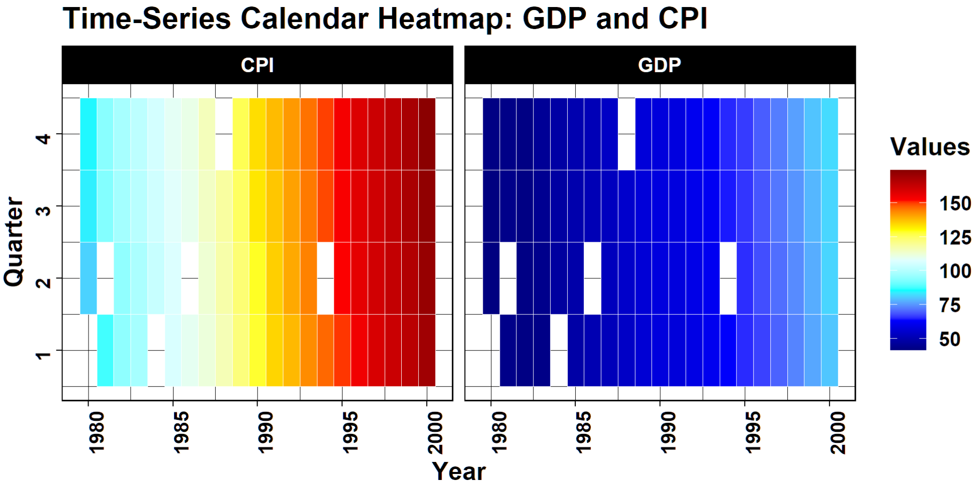
5. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

6. C. Sievert. Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC Florida, 2020.

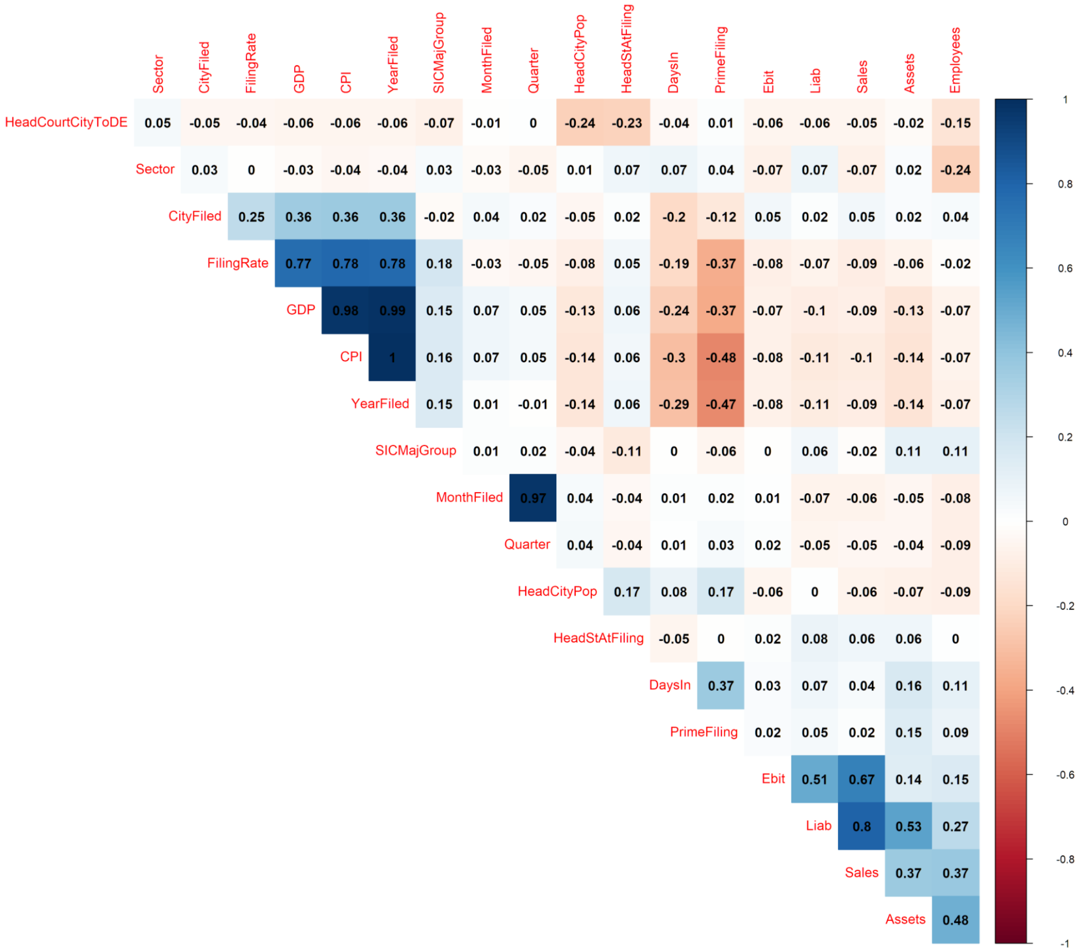
7. Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret

8. Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O''Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs and Helene Wagner (2020). Vegan:
   Community Ecology Package. R package version 2.5-7.
   https://CRAN.R-project.org/package=vegan

Appendix



The heatmap of CPI and GDP



	Sector	CityFiled	FilingRate	GDP	CPI	YearFiled	SICMajGroup	MonthFiled	Quarter	HeadCityPop	HeadStAtFiling	DaysIn	PrimeFiling	Ebit	Liab	Sales	Assets	Employees
HeadCourtCityToDE	0.05	-0.05	-0.04	-0.06	-0.06	-0.06	-0.07	-0.01	0	-0.24	-0.23	-0.04	0.01	-0.06	-0.06	-0.05	-0.02	-0.15
Sector		0.03	0	-0.03	-0.04	-0.04	0.03	-0.03	-0.05	0.01	0.07	0.07	0.04	-0.07	0.07	-0.07	0.02	-0.24
CityFiled			0.25	0.36	0.36	0.36	-0.02	0.04	0.02	-0.05	0.02	-0.2	-0.12	0.05	0.02	0.05	0.02	0.04
FilingRate				0.77	0.78	0.78	0.18	-0.03	-0.05	-0.08	0.05	-0.19	-0.37	-0.08	-0.07	-0.09	-0.06	-0.02
GDP					0.98	0.99	0.15	0.07	0.05	-0.13	0.06	-0.24	-0.37	-0.07	-0.1	-0.09	-0.13	-0.07
CPI						1	0.16	0.07	0.05	-0.14	0.06	-0.3	-0.48	-0.08	-0.11	-0.1	-0.14	-0.07
YearFiled							0.15	0.01	-0.01	-0.14	0.06	-0.29	-0.47	-0.08	-0.11	-0.09	-0.14	-0.07
SICMajGroup								0.01	0.02	-0.04	-0.11	0	-0.06	0	0.06	-0.02	0.11	0.11
MonthFiled									0.97	0.04	-0.04	0.01	0.02	0.01	-0.07	-0.06	-0.05	-0.08
Quarter										0.04	-0.04	0.01	0.03	0.02	-0.05	-0.05	-0.04	-0.09
HeadCityPop											0.17	0.08	0.17	-0.06	0	-0.06	-0.07	-0.09
HeadStAtFiling												-0.05	0	0.02	0.08	0.06	0.06	0
DaysIn													0.37	0.03	0.07	0.04	0.16	0.11
PrimeFiling														0.02	0.05	0.02	0.15	0.09
Ebit															0.51	0.67	0.14	0.15
Liab																0.8	0.53	0.27
Sales																	0.37	0.37
Assets																		0.48

The correlation matrix of all variables

The following pages are the code of this analysis.

In case it cannot be executed, the file can be found :

https://github.com/JasonXsiu/PCA_MDS_US_bankruptcy

~~~
---
title: "PCA and MDS of American Bankruptcy between 1980 and 2000 "
output: html_document
authors : Jason Ching Yuen Siu, Liguo Bao, Wanxin Liu, Zhixiang Yang
---




```{r include=FALSE}

knitr::opts_chunk$set(echo = TRUE, fig.width = 10, fig.height = 5, warning = F)
library(tidyverse)
library(naniar)
library(Hmisc)
library(broom)
library(fastDummies)
library(plotly)
library(viridis)

# MDS
remotes::install_github("AckerDWM/gg3D")
library(gg3D)  # 这是一个 3d 数据包
library(rgl)
library(gg3D)
library(vegan)
library(plotly)
library(caret)

#PCA 2D
library(broom)
library(dummies)
library(tidyr)

#PCA3D
library(plot3D)
library(plot3Drgl)
Bankruptcy <- readRDS('Bankruptcy.RDS')
```

===============================Preliminary Analysis ===============================


View NA
~~~

~~~~
```{r}
png(file="PA_Missingness.png", res=300, width=1600, height=1000)
vis_miss(Bankruptcy, sort=T) +
  ggtitle("Missingness of the Bankruptcy dataset")+
  theme(axis.text.x=element_text( angle = 90,size=8))
```

- Only 2.1% of the data has NA values, which is somewhat clean.
- However, since EmplUnion has 30.96 % of NA values, this feature is eliminated.
- FirmEnd and DENYOther brings little values in this analysis, therefore they are
 also eliminated.

Exclude non-needed attributes
```{r}
Bankruptcy <- Bankruptcy %>% select(-EmplUnion,-FirmEnd)
```

- Since the economy is analysed by quarter usually, therefore, Quarter is sorted.
```{r}
# add an index to the dataset
Bankruptcy <- Bankruptcy %>% mutate(ID = row_number())
# make a quarter
Bankruptcy <- Bankruptcy %>% mutate(Quarter = ceiling(MonthFiled / 3))
```
- There are 55 industries. It will be difficult to conduct an anlysis with so man
y categorical values;
hence they are categorised into different sectors.
```{r}
# Categorise sector
agri <- tibble(Industry = unique(Bankruptcy$SICMajGroup) )%>%
  filter(row_number()==53)
financial <- tibble(Industry = unique(Bankruptcy$SICMajGroup) )%>%
  filter(row_number() %in% c(50,17,30))
property <- tibble(Industry = unique(Bankruptcy$SICMajGroup) )%>%
  filter(row_number() %in% c(27,33,34,37,29,42))
service <- tibble(Industry = unique(Bankruptcy$SICMajGroup) )%>%
  filter(row_number() %in% c(22,28,47,51,8,24,54,55,11,38,20,52))

industrial <- tibble(Industry = unique(Bankruptcy$SICMajGroup) )%>%
  filter(row_number() %in% c(4,6,9,1,21,39,41,13,2,35,19,3,45,14))

consumption <- tibble(Industry = unique(Bankruptcy$SICMajGroup) )%>%
  filter(row_number() %in% c(43,44,40,32,23,31,15,46,25,18,16,5,7,36,12))

resource <- tibble(Industry = unique(Bankruptcy$SICMajGroup) )%>%
~~~~

~~~~
  filter(row_number() %in% c(10,26,48,49))

b <- data.frame()
assign_sector <- function(industry,sector){
  for(i in 1:length(industry)){
  a <- Bankruptcy %>% filter(SICMajGroup %in% industry$Industry)  %>% mutate(Sect
or = sector)
  b <-   rbind(a,b)
  }
  return(b)
}

df1 <- assign_sector(financial,"Financial")
df2 <- assign_sector(property, "Property")
df3 <- assign_sector(service, "Service")
df4 <- assign_sector(industrial, "Industrial")
df5 <- assign_sector(consumption, "Consumption")
df6 <- assign_sector(agri,"Agriculture")
df7 <- assign_sector(resource,"Resource")

Bankruptcy <-  rbind(df1,df2,df3,df4,df5,df6, df7)
#Denyother enumerate
Bankruptcy$DENYOther_id <- Bankruptcy$DENYOther %>% as.factor() %>% as.numeric()
#filter df for PA
df_for_desc_stat<-  rbind(df1,df2,df3,df4,df5,df6, df7)
#filter df for MDS
new_var<-
Bankruptcy[,c("Name","CPI","DaysIn","Ebit","Employees","FilingRate","GDP","HeadCi
tyPop","HeadCourtCityToDE","Liab","Sales","HeadStAtFiling","Sector","YearFiled","
Quarter","CityFiled","DENYOther","MonthFiled","ID","DENYOther_id")]%>%drop_na()
#filter df for PCA3D
Bankruptcy%>%
  filter(ID != 58) %>%
  column_to_rownames('ID') %>%
  select_if(is.numeric) %>%
  select(-Quarter,-DENYOther_id) %>%
  na.omit() %>%
  filter()-> pca_data

```
Categorise variables
```{r factor}

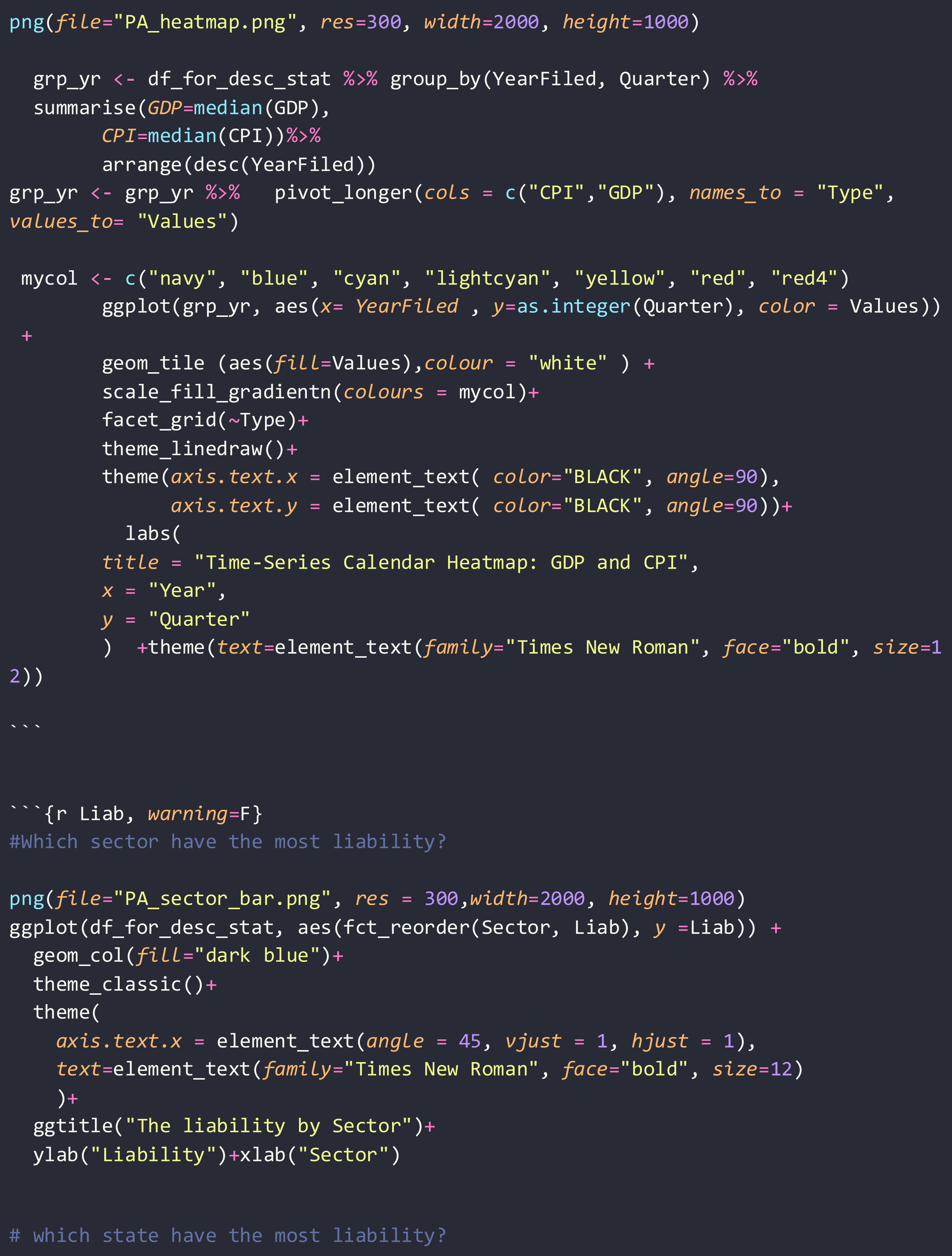
Bankruptcy$CityFiled <- as.factor(Bankruptcy$CityFiled) %>% as.numeric()
~~~~

```
Bankruptcy$HeadStAtFiling <- as.factor(Bankruptcy$HeadStAtFiling) %>% as.numeric()
Bankruptcy$SICMajGroup <- as.factor(Bankruptcy$SICMajGroup) %>% as.numeric()
Bankruptcy$Sector <- as.factor(Bankruptcy$Sector)%>% as.numeric()
Bankruptcy$Assets <- as.factor(Bankruptcy$Assets) %>% as.numeric()
Bankruptcy$DaysIn <- as.factor(Bankruptcy$DaysIn) %>% as.numeric()
Bankruptcy$Employees <- as.factor(Bankruptcy$Employees) %>% as.numeric()
Bankruptcy$FilingRate <- as.factor(Bankruptcy$FilingRate) %>% as.numeric()
Bankruptcy$HeadCourtCityToDE <- as.factor(Bankruptcy$HeadCourtCityToDE) %>% as.numeric()
Bankruptcy$MonthFiled <- as.factor(Bankruptcy$MonthFiled) %>% as.numeric()
Bankruptcy$YearFiled <- as.factor(Bankruptcy$YearFiled) %>% as.numeric()


```

```{r select-numeric}
bc_numeric <- Bankruptcy %>% select_if(is.numeric)
```

```{r boxplot}
company <- c("Assets","Ebit","Employees","Liab")
general <- c("FilingRate","Quarter","CityFiled")

#Company
boxp<-ggplot(
  gather(
    #remove unneeded columns
    bc_numeric %>% select(company))
  , aes(value)) +
    geom_boxplot(bins = 10) +
    facet_wrap(~key, scales = 'free_x')+
  theme_bw()+
    theme(axis.text.x = element_text(angle = 360, vjust = 1, hjust = 1))+
    geom_boxplot(color="red", fill="orange", alpha=0.2)+
  xlab("")+ylim(10,500)
ggplotly(boxp)



```
```{r heatmap-macro, warning=F}
```

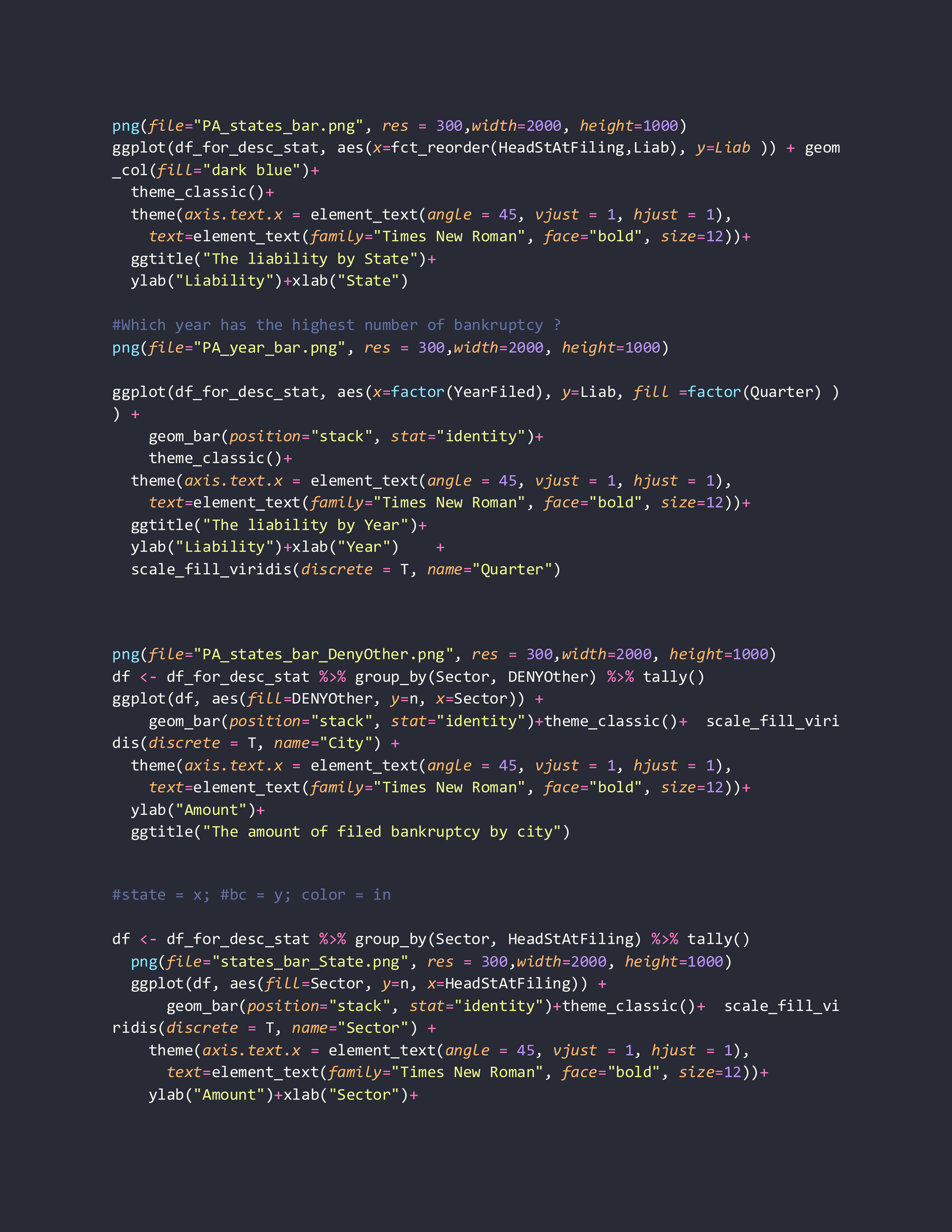


```
png(file="PA_heatmap.png", res=300, width=2000, height=1000)

  grp_yr <- df_for_desc_stat %>% group_by(YearFiled, Quarter) %>%
  summarise(GDP=median(GDP),
        CPI=median(CPI))%>%
        arrange(desc(YearFiled))
grp_yr <- grp_yr %>%   pivot_longer(cols = c("CPI","GDP"), names_to = "Type",
values_to= "Values")

 mycol <- c("navy", "blue", "cyan", "lightcyan", "yellow", "red", "red4")
        ggplot(grp_yr, aes(x= YearFiled , y=as.integer(Quarter), color = Values))
 +
        geom_tile (aes(fill=Values),colour = "white" ) +
        scale_fill_gradientn(colours = mycol)+
        facet_grid(~Type)+
        theme_linedraw()+
        theme(axis.text.x = element_text( color="BLACK", angle=90),
              axis.text.y = element_text( color="BLACK", angle=90))+
          labs(
        title = "Time-Series Calendar Heatmap: GDP and CPI",
        x = "Year",
        y = "Quarter"
        )  +theme(text=element_text(family="Times New Roman", face="bold", size=1
2))

```

```{r Liab, warning=F}
#Which sector have the most liability?

png(file="PA_sector_bar.png", res = 300,width=2000, height=1000)
ggplot(df_for_desc_stat, aes(fct_reorder(Sector, Liab), y =Liab)) +
  geom_col(fill="dark blue")+
  theme_classic()+
  theme(
    axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
    text=element_text(family="Times New Roman", face="bold", size=12)
    )+
  ggtitle("The liability by Sector")+
  ylab("Liability")+xlab("Sector")


# which state have the most liability?
```



```
png(file="PA_states_bar.png", res = 300,width=2000, height=1000)
ggplot(df_for_desc_stat, aes(x=fct_reorder(HeadStAtFiling,Liab), y=Liab )) + geom
_col(fill="dark blue")+
  theme_classic()+
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
    text=element_text(family="Times New Roman", face="bold", size=12))+
  ggtitle("The liability by State")+
  ylab("Liability")+xlab("State")

#Which year has the highest number of bankruptcy ?
png(file="PA_year_bar.png", res = 300,width=2000, height=1000)

ggplot(df_for_desc_stat, aes(x=factor(YearFiled), y=Liab, fill =factor(Quarter) )
) +
    geom_bar(position="stack", stat="identity")+
    theme_classic()+
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
    text=element_text(family="Times New Roman", face="bold", size=12))+
  ggtitle("The liability by Year")+
  ylab("Liability")+xlab("Year")    +
  scale_fill_viridis(discrete = T, name="Quarter")



png(file="PA_states_bar_DenyOther.png", res = 300,width=2000, height=1000)
df <- df_for_desc_stat %>% group_by(Sector, DENYOther) %>% tally()
ggplot(df, aes(fill=DENYOther, y=n, x=Sector)) +
    geom_bar(position="stack", stat="identity")+theme_classic()+  scale_fill_viri
dis(discrete = T, name="City") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
    text=element_text(family="Times New Roman", face="bold", size=12))+
  ylab("Amount")+
  ggtitle("The amount of filed bankruptcy by city")


#state = x; #bc = y; color = in

df <- df_for_desc_stat %>% group_by(Sector, HeadStAtFiling) %>% tally()
  png(file="states_bar_State.png", res = 300,width=2000, height=1000)
  ggplot(df, aes(fill=Sector, y=n, x=HeadStAtFiling)) +
      geom_bar(position="stack", stat="identity")+theme_classic()+  scale_fill_vi
ridis(discrete = T, name="Sector") +
    theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
      text=element_text(family="Times New Roman", face="bold", size=12))+
    ylab("Amount")+xlab("Sector")+
```

~~~
    ggtitle("The amount of filed bankruptcy in different states by sector")

```


```{r #of File}
Bankruptcy %>%
 mutate(Quarter = factor(tools::toTitleCase(as.character(Quarter)))) %>%
   ggplot() +
    geom_histogram(aes(Quarter, fill = Quarter), stat = "count") +
   theme_minimal(base_size = 10) +
   ggtitle("Number of companies filed per quarter") +
   scale_fill_brewer(palette = "Set3") +
   xlab("Quarter ") +
   ylab("Num. Companies")->p11

p11
```

```{r corr-matrix}
library(corrplot)

png(file="corr.png", res=300, width=4500, height=4500)

corbank<-cor(bc_numeric %>% drop_na()) %>% as.matrix()

corrplot(corbank,
         method="color",
         type="upper",
         order="hclust",diag = FALSE,
         addCoef.col = "black",sig.level = 0.05,
         insig="blank",
         number.cex=1,
         mar=c(0,1,1,5),
         main="The correlation graph between numerical variables")

```


===============================MDS ===============================


```{r df_for_MDS}
~~~

~~~
NumSec<-as.matrix(new_var["Sector"])
NumSector<-as.numeric(factor(NumSec))
NumSector<-tibble(NumSector)
new_var<-cbind(new_var,NumSector)



num_var<-
new_var[,c("CPI","DaysIn","Ebit","Employees","FilingRate","GDP","HeadCityPop","He
adCourtCityToDE","Liab","NumSector","Quarter","ID","DENYOther_id")]

dropped<-num_var%>%drop_na()



## here how to adjust the margin again with more clearer display
```




## Goodness of fit

```{r}

delta1<-filter(num_var)%>%
   select_if(is.numeric)%>%
   scale%>%
   dist(method="manhattan")

attributes(delta1)$Labels<-filter(new_var)%>%
   pull(HeadStAtFiling)%>%
   abbreviate(6)
 ## delta1 is manhattan distance but same as Euclidean Classical method

delta<-filter(num_var)%>%
   select_if(is.numeric)%>%
   scale%>%
   dist



 ## Euclidean distance data in delta


attributes(delta)$Labels<-filter(new_var)%>%
~~~

~~~~
   pull(CityFiled)%>%
   abbreviate(6)



mdsout1<-cmdscale(delta1,eig = TRUE) ##delta1 表示的 Manhattan  距离
str(mdsout1$GOF)

mdsout<-cmdscale(delta,eig=TRUE)## delta 表示 Euclidean 距离

str(mdsout$GOF)

eigval<-mdsout$eig

```



## 3D Euclidean MDS



```{r scaling}
normalize <- function(x, na.rm = TRUE) {
    return((x- min(x)) /(max(x)-min(x)))
}
```



```{r}

# 这个就是转换成三维坐标写的，还是 classical 的模型

data.dist<-dist(dropped %>% normalize()) ## euclidean distance , normalised
data.mds<-cmdscale(data.dist,k=3) ## Generate euclidean

#plot3d(data.mds,col=dropped$NumSector)


df11<-
data.frame(var1=data.mds[,1],var2=data.mds[,2],var3=data.mds[,3],Employees=droppe
d$Employees) ## EUCLIDEAN  datast
~~~~

~~~~
d3plot<-
ggplot(df11,aes(x=var1,y=var2,z=var3,color=Employees))+theme_void() +axes_3D()+st
at_3D() + scale_colour_gradient(
  low = "orange",
  high = "dark blue",
  space = "Lab",
  na.value = "grey50",
  guide = "colourbar",
  aesthetics = "colour"
)+ggtitle("Relationship between employees and companies")+theme(axis.text.x = ele
ment_text(angle = 45, vjust = 1, hjust = 1),text=element_text(family="Times New R
oman", face="bold", size=12))

png(file="MDS_relation_bt_emp_comp.png", res=300, width=4500, height=4500)

ggplotly(d3plot)


#plot_ly(data=df11,x=var1,y=var2,z=var3,type="scatter3d", mode="markers")

```



##relationships between headquarters

```{r}

# 这个就是转换成三维坐标写的，还是 classical 的模型

data.dist<-dist(dropped %>% normalize()) ## euclidean distance , normalised
data.mds<-cmdscale(data.dist,k=3) ## Generate euclidean



df11<-
data.frame(var1=data.mds[,1],var2=data.mds[,2],var3=data.mds[,3],dishead=dropped$
HeadCourtCityToDE) ## EUCLIDEAN  datast

d3plot<-
ggplot(df11,aes(x=var1,y=var2,z=var3,color=dishead))+theme_void() +axes_3D()+stat
_3D() + scale_colour_gradient(
  low = "orange",
  high = "dark blue",
  space = "Lab",
~~~~

```
  na.value = "grey50",
  guide = "colourbar",
  aesthetics = "colour"
)+ggtitle("Relationship between distances of headquarter and bankrupcy")

png(file="MDS_relation_bt_head_bank.png", res=300, width=4500, height=4500)
ggplotly(d3plot)


```
```

```
# Sammon mapping
```{r SM, warning=F}
library(MASS)
dropped.sammon<-sammon(dist(dropped),k=2) ## 二维 sammon 的坐标在这里
# stress
dropped.sammon$stress


scree.plot = function(d, k) {
    stresses=sammon(d, k=k)$stress
    for(i in rev(seq(k-1)))
        stresses=append(stresses,sammon(d, k=i)$stress)
    plot(seq(k),rev(stresses), type="b", xaxp=c(1,k, k-1), ylab="Stress", xlab="Number of dimensions",main="Scree plot via sammon mapping")
}


png(file="MDS_screeplot.png", res=300, width=4500, height=4500)
scree.plot(dist(dropped),k=6)## scree plot 对于 sammon 来说


## sammon mapping
df13<-data.frame(v1=dropped.sammon$points[,1],
                 v2=dropped.sammon$points[,2],
                 Sector=factor(NumSec),DENYother=factor(new_var$DENYOther)) ## 这个 sammon 的图建议还是把数字换成 industry 的名称!!

colors <- c("Financial" = "#F51720", "Property" ="Blue" , "Industrial" = "Yellow", "Industrial"="#4C5270" , "Consumption" = "#F8D210", "Agriculture" = "#FA26A0", "Resource"  = "dark blue" )
```

```
png(file="MDS_sector.png", res=300, width=4500, height=4500)

ggplot(df13,aes(x=v1,y=v2,col=Sector))+
   geom_point(size = 4, alpha =.5)+
   scale_shape_manual(values = c(10:16))+
  theme_classic() +ggtitle("The Sector of bankrupted companies")+theme(axis.text.
x = element_text(angle = 45, vjust = 1, hjust = 1),
    text=element_text(family="Times New Roman", face="bold", size=12))

```
```

```
## relationship between year filed and the legnth of Bankrupcy
```{r rug , warning=F}

## with the year increased , the length of bankrupcy becomes faster than previous
 years
## previous year ggplot(new_var,aes(x=DaysIn,col=YearFiled))+geom_rug()+ggtitle("
The length of bankrupcy by year")

df15<-
data.frame(DaysIn=new_var$DaysIn,DENYOth=new_var$DENYOther,YearFiled=factor(new_v
ar$YearFiled))
png(file="MDS_length_BC_by_yr.png", res=300, width=4500, height=4500)
ggplot(df15,aes(x=DaysIn,y=DENYOth,col=YearFiled))+
  geom_point()+
  ggtitle("The length of bankrupcy by year")+
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
    text=element_text(family="Times New Roman", face="bold", size=12))+
  scale_fill_viridis(discrete = T, name="YearFiled") +theme_bw()
```

## NMDS

```{r NMDS, warning=F}
set.seed(123456)
```

~~~~
dropped.nmds<-metaMDS(dropped,distance="bray",autotransform = FALSE)

dropped.envfit<-envfit(dropped.nmds,dropped,permutations=999)



dropped.nmds
```


#DENYother

```{r ordiplot}
png(file="MDS_ordiplot.png", res=300, width=4500, height=4500)


ordiplot(dropped.nmds,type="n", main = "Different locations NMDS")

#ordiplot(dropped.nmds,type="n", main = "different quarters")

orditorp(dropped.nmds, display = "sites", labels = F, col = c("green", "blue","red") [as.numeric(dropped$DENYOther_id)], cex = 1)

# other green, NY blue, DE red
legend("bottomleft", legend=c("OT", "NY", "DE"),
       col = c("green", "blue","red"),lty=1, cex=1,box.lty=2, box.lwd=1)

```


===============================PCA2D ===============================



# Principal Components Analysis and Biplot

Principal components analysis is a convenient way to merge variables through a linear combination. Since this report wants to investigate the reason of bankruptcy in different observations and more convenient views, we set ID as row names instead of certain names. After removing missing values, numerical variables are selected because categorical variables cannot be used by PCA. In addition, original variables are measured in different units, so we standardise the data by dividing by the standard deviation.

In the whole PCA analysis, the first two components just explain 49.46% variance, which is lower. However, we can firstly investigate some insights from the first
~~~~

two components. The first principal component is a measure of macro attributes of bankrupt companies, since it is positively correlated with common macro features of improving bankrupt rate, such as filling rate, CPI, yearfield and GDP whereas it is negatively correlated with common macro features of declining bankrupt rate, such as PrimeFiling. Similarly, the second principal components is a measure of micro attributes of bankrupt companies because it is positively correlated with common micro features of improving bankrupt rate, such as Employees, Ebit, Liab, Sales and Assets. Moreover, some loadings of variables are small and no significant effect on PC1 and PC2, so we decide to explain those variables in 3d PCA plots later.

Most observations are focused on the same position, which means that those bankruptcy companies have similar features. No.58 company is an significant outlier in the plot, and this company is called Texaco Inc. It is the world's third-largest oil company, and the biggest firm ever to seek bankruptcy protection. Texaco was forced into bankruptcy as a result of a 1985 decision by a Houston jury that Texaco had illegally interfered with Pennzoil's plan to acquire Los Angeles-based Getty Oil Co. The jury awarded Pennzoil damages of $10.5 billion plus interest. The PCA plot also confirms it. Texaco Inc. has the largest number of assests, liab, ebit and employees.

~~~~
```{r function-for-adding-title}
add_title <-  function(sector,title_at,subtitle_at){
mytitle = "Principal Component Analysis"
mysubtitle = paste("",sector)
mtext(side=3, line=3, at=title_at, adj=0, cex=1.3, mytitle)
mtext(side=3, line=2, at=subtitle_at, adj=0, cex=1, mysubtitle)
}
```
~~~~

~~~~
```{r}
Bankruptcy%>%
 column_to_rownames('ID') %>%
 select_if(is.numeric) %>%
 na.omit() %>% prcomp(scale.=TRUE)->whole_pca

#screeplot
summary(whole_pca)
png(file="PCA2D_Screeplot.png", res=300, width=2000, height=1700)
screeplot(whole_pca,type = 'l')

#biplot
png(file="PCA2D_biplot.png", res=300, width=2000, height=1700)
biplot(whole_pca,cex=0.8,sub="Distance plot")
add_title("Entire sectors", -175.1, -175.4)
~~~~

~~~
```
~~~

In order to analysis the features of bankruptcy companies, we make PCA analysis in different industry.

- In financial industry, companies are separated into three parts. The first part is center on relative higher GDP, CPI, YearField, FillingRate, such as 339 and 308. The second part is focused on relative lower employees, liab, assests, primefiling, sales, such as 270 and 173. The third part has higher headCityPop such as 63 and 46.

~~~
```{r fin-PCA}
#df1 financial
financial_df1_pca <- df1 %>%
 column_to_rownames('ID') %>%
 select_if(is.numeric) %>%
 na.omit() %>% prcomp(scale.=TRUE)

summary(financial_df1_pca)

#Distance biplot
png(file="PCA2D_Fin_biplot.png", res=300, width=2000, height=1700)
biplot(financial_df1_pca,sub="Distance biplot",cex=0.8)
add_title("Financial sector",-8.1, -8.2)
```
~~~

- In property industry, companies are relative separated. No. 206 is the outlier, which is JWP Inc. It has highest sales, employees and lowest filling rate and daysin.

~~~
```{r prop-PCA}
#df2 property
property_df2_pca <- df2 %>%
 column_to_rownames('ID') %>%
 select_if(is.numeric) %>%
 na.omit() %>% prcomp(scale.=TRUE)

summary(property_df2_pca)
#Distance biplot
png(file="PCA2D_Prop_biplot.png", res=300, width=2000, height=1700)
biplot(property_df2_pca,sub="Distance biplot",cex=0.8)
add_title("Property sector", -5.5,-5.6)
~~~

~~~
```

- In service industry, most companies are widely distributed. Probably those serv
ice companies do not have more similar features than other industry.


```{r serv-PCA}
service_df3_pca <- df3 %>%
 column_to_rownames('ID') %>%
 select_if(is.numeric) %>%
 na.omit() %>% prcomp(scale.=TRUE)

summary(service_df3_pca)
#Distance biplot
png(file="PCA2D_Serv_biplot.png", res=300, width=2000, height=1700)
biplot(service_df3_pca,sub="Distance biplot",cex=0.8)
add_title("Service sector", -15.5,-15.6)
```

- In industry, almost all companies are concentrated on the same position, which
means that they all have similar features. It is worth noting No.45 and No.413. N
o.45 is LTV Corp, which have the highest micro features, such as highest employee
s, ebit, sales and liab.
No.413 is Owens Corning, which has the lowest population of the firms headquarter
s city. Because its headquarters city is Toledo ohio, which just has 270871 popul
ation in 2020.

```{r ind-PCA}
industry_df4_pca <- df4 %>%
 column_to_rownames('ID') %>%
 select_if(is.numeric) %>%
 na.omit() %>% prcomp(scale.=TRUE)

#Distance biplot
png(file="PCA2D_Ind_biplot.png", res=300, width=2000, height=1700)
biplot(industry_df4_pca,sub="Distance biplot",cex=0.8)
add_title("Industrial sector", -8.5,-8.6)
```

- In consumption industry, large parts concentrate on the arrow of 'HeadCityPop'
which indicates that almost half headquarters of consumption companies in the cit
ies with relative large populations and population might be a indicator for consu
mption companies to establish headquarters. For other major parts of consumption,
 they did not show a specific distribution with loading directions, further PCs m
ay be needed.
~~~

~~~~
```{r cons-PCA}
consumption_df5_pca <- df5 %>%
 column_to_rownames('ID') %>%
 select_if(is.numeric) %>%
 na.omit() %>% prcomp(scale.=TRUE)
#Distance biplot
png(file="PCA2D_Cons_biplot.png", res=300, width=2000, height=1700)
biplot(consumption_df5_pca,sub="Distance biplot",cex=0.8)
add_title("Consumption sector", -8.8,-8.9)
```

- In resource industry, companies spread dispersed in the biplot, most of them ar
e on the positive side of PC2 while the variables of 'YearFiled', 'GDP' and 'Fili
ngRate' cotributes to positive PC2. However, those companies only go over a bit o
f PC2 and none of variables could specifically explain their distribution so thos
e 3 variables might no be able to explain those major observation well.



```{r res-PCA}
resource_df7_pca <- df7 %>%
 column_to_rownames('ID') %>%
 select_if(is.numeric) %>%
 na.omit() %>% prcomp(scale.=TRUE)

png(file="PCA2D_Res_biplot.png", res=300, width=2000, height=1700)
biplot(resource_df7_pca,sub="distance biplot",cex=0.8)
add_title("Resource sector",-4.22,-4.24)
```

===============================PCA3D ===============================




```{r screeplot}
bkc <- Bankruptcy %>% semi_join(pca_data, by = 'Ebit')
#PCA
bkc_pca<- prcomp(pca_data[1:14], center = TRUE, scale = TRUE)
join <- bkc %>% mutate(PC1 = bkc_pca$x[,1], PC2 = bkc_pca$x[,2], PC3 = bkc_pca$x[
,3])

bkc_pca_var <- tibble(PC=1:length(bkc_pca$sdev), prop.var=bkc_pca$sdev^2)
bkc_pcs <- as_tibble(bkc_pca$x[,1:3]) %>%
~~~~

```
  mutate(company=bkc$Name)
bkc_pcs_evc <- as_tibble(bkc_pca$rotation[,1:3]) %>%
  mutate(origin = rep(0,14), variables = colnames(pca_data)) %>%
  mutate(PC1s = PC1*(bkc_pca_var$prop.var[1]*2),
         PC2s = PC2*(bkc_pca_var$prop.var[2]*2),
         PC3s = PC3*(bkc_pca_var$prop.var[3]*2)
         )
```
added in Appendix
```{r }

par(mar = c(1.2,1.2,1.2,1.2))
scatter3D(bkc_pcs$PC1, bkc_pcs$PC2, bkc_pcs$PC3,
          pch = 20, theta = 150, phi = 10,
          ticktype = "detailed",
     colvar= as.integer(factor(bkc$DENYOther )),
     col = gg2.col(3),
     colkey = list(
       at = c(
         unique(as.integer(factor(bkc$DENYOther)))),
       side = 4, length = 1, width = 0.5, labels = unique(factor(bkc$DENYOther)))
,
      xlab = "PC1", ylab = "PC2", zlab = "PC3", main = "PCs agaisnt DENYOther", c
lab = "DENYOther"
     )
points3D(bkc_pcs_evc$origin, bkc_pcs_evc$origin, bkc_pcs_evc$origin, add = TRUE,
col="blue",
          colkey = FALSE, pch = 19, cex = 1)
arrows3D(bkc_pcs_evc$origin, bkc_pcs_evc$origin, bkc_pcs_evc$origin,
         bkc_pcs_evc$PC1s, bkc_pcs_evc$PC2s, bkc_pcs_evc$PC3s,
         phi = 0, theta = 90,
         lwd = 2, d = 3, clab = c("Quality", "score"),
         main = "Loading", bty ="g", ticktype = "detailed", overlay = T,add = T)
text3D( bkc_pcs_evc$PC1s, bkc_pcs_evc$PC2s, bkc_pcs_evc$PC3s,
       c(bkc_pcs_evc$variables),
       add=TRUE, colkey = FALSE, col = "red")
```

```{r biplot-sector, warning=F}
biplot_sector <- ggplot() +
  geom_point(data=join, aes(x=PC2, y=PC3, col = Sector)) +

  xlab("PC2") + ylab("PC3") +
  theme(aspect.ratio=1)  +
  theme_classic() +
  geom_segment(data=bkc_pcs_evc, aes(x=origin, xend=PC2s, y=origin, yend=PC3s), c
olour="black") +
  ggrepel:: geom_text_repel(data=bkc_pcs_evc, aes(x=PC2s, y=PC3s, label=variables
), colour="red", nudge_y=sign(bkc_pcs_evc$PC3)*0.2)+
  labs(x = "PC2", y = "PC3", title = "Biplot", subtitle = "Variables projected on
to PC2 and PC3") +
   theme( text=element_text(family="Times New Roman", face="bold", size=12))

png(file="PCA3D_biplot_Sector.png", res=300, width=2000, height=1000)
biplot_sector
```
```{r biplot-DNother, warning=F}
biplot_deny <- ggplot() +
  geom_point(data=join, aes(x=PC2, y=PC3, col = DENYOther)) +
  xlab("PC2") + ylab("PC3") +
  theme(aspect.ratio=1) +
  theme_classic() +
  geom_segment(data=bkc_pcs_evc, aes(x=origin, xend=PC2s, y=origin, yend=PC3s), c
olour="black") +
  ggrepel:: geom_text_repel(data=bkc_pcs_evc, aes(x=PC2s, y=PC3s, label=variables
), colour="red", nudge_x=0.1, nudge_y=sign(bkc_pcs_evc$PC3)*0.2)+
  labs(x = "PC2", y = "PC3", title = "Biplot", subtitle = "Variables projected on
to PC2 and PC3") +
   theme( text=element_text(family="Times New Roman", face="bold", size=12))

png(file="PCA3D_biplot_DENYother.png", res=300, width=2000, height=1000)
biplot_deny

```
```
```