A18 - Machine Learning (principal components)

Bryana Benson
Conner Bryan

March 27, 2019

Using FT Global 500 data, apply forcats::fct_lump() to create five groups of sectors (the top four by number of companies and the rest). Do a principal components analysis.

Library

```
library(readxl)
library(dplyr)
library(corrplot)
library(ggplot2)
```

Data

Remove omitted data in the dataset to prevent errors.

```
#Select only numeric variables
data.n <- select_if(data, is.numeric)
```

Since PCA works best with numerical data, you'll exclude the categorical variables, and view the correlations.

```
#Correlations of numeric data
corr <- round(cor(data.n),2)
```

The Global Rank in 2014 and 2015 are highly correlated. The Market Value in Millions has a strong negative correlation to Global Rank in 2014 and 2015. Market value and Net income have a moderately - strong positive corrleation.

PCA prep and Plot

Data Types

```
#check data types
str(data)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    377 obs. of  14 variables:
##  $ Global rank 2015  : num  1 2 5 6 7 8 9 10 11 12 ...
##  $ Global rank 2014  : num  1 2 3 16 7 6 21 14 25 10 ...
##  $ Company           : chr  "Apple" "Exxon Mobil" "Microsoft" "PetroChina" ...
##  $ Country           : chr  "US" "US" "US" "China" ...
##  $ Market value $m   : num  724773 356549 333525 329715 279920 ...
##  $ Sector            : chr  "Technology hardware & equipment" "Oil & gas producers" "Software & compu
##  $ Turnover $m       : chr  "182795" "364763" "86833" "367853.66999999998" ...
##  $ Net income $m     : num  39510 32520 22074 17269 23057 ...
##  $ Total assets $m   : num  231839 349493 172384 385178 1687155 ...
##  $ Employees         : num  92600 75300 128000 534652 264500 ...
##  $ Price $           : num  124.43 85 40.66 1.11 54.4 ...
##  $ P/e ratio         : num  19.3 11.2 15.5 12.3 13.3 ...
##  $ Dividend yield (%): num  1.45 3.18 2.75 3.61 2.48 ...
##  $ Year End          : POSIXct, format: "2014-09-27" "2014-12-31" ...
##  - attr(*, "na.action")= 'omit' Named int  3 4 29 33 43 54 77 86 113 126 ...
```

```
##   ..- attr(*, "names")= chr  "3" "4" "29" "33" ...
```

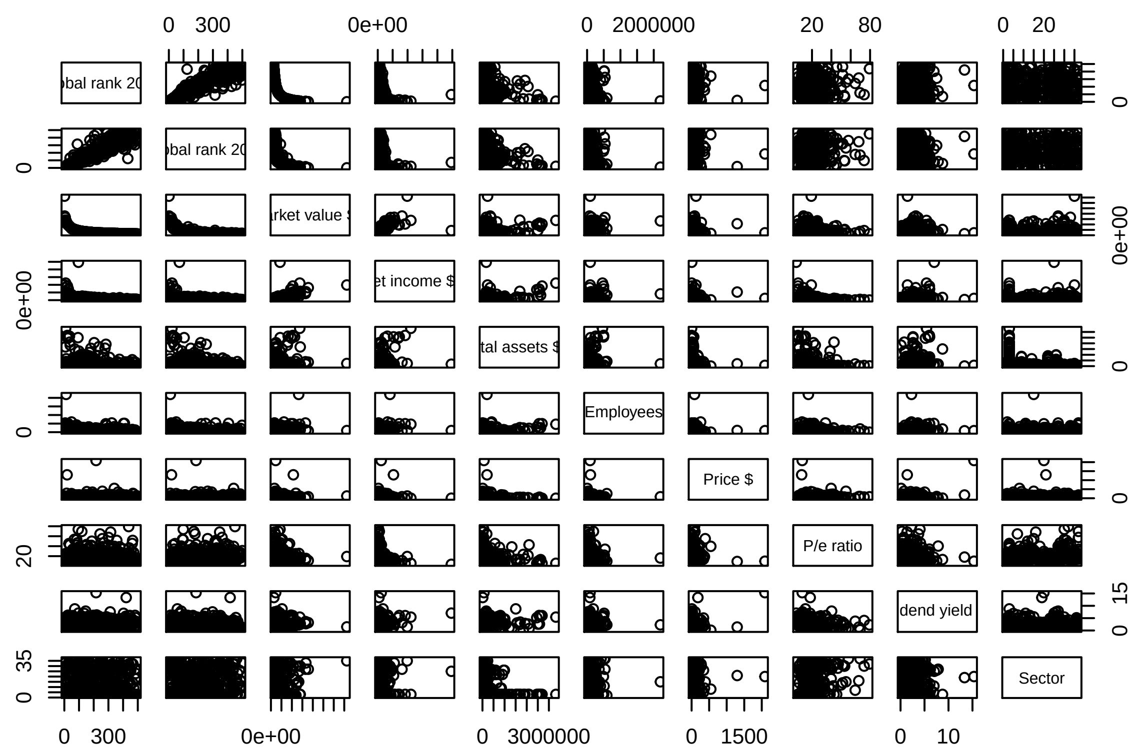
```r
#change Sector data type to factor
data$Sector <- as.factor(data$Sector)
```

```r
#Add back Sector column to numeric dataset
data.n$Sector <- data$Sector

#Rename data.n back to data
data <- data.n
```

Plots

```r
plot(data)
```



Grouping

```r
#Use negative value to collapse the 4 most common groups of Sectors
data$Sector<- forcats::fct_lump(data$Sector, n= -4)
```

fact_lump lumps together least/most common factor levels into "other". In this case, we want to lump together the 4 most common Sectors and leave the rest in Other.

Principal Component Analysis

The purpose of principal components analysis is to reduce the complexity of a multivariate dataset into a principal components space. It includes the mathematical transformation of number of possibly correlated variables into a smaller number of uncorrelated variables.

```r
#run PCA
pca <- prcomp(data[,c(1:9)],center = T, scale=T)
pca
```

```
## Standard deviations (1, .., p=9):
## [1] 1.8293613 1.1757078 1.0553088 0.9134785 0.8619354 0.8272603 0.7421603
## [8] 0.5070885 0.2963034
##
## Rotation (n x k) = (9 x 9):
##                              PC1         PC2         PC3         PC4
## Global rank 2015    0.465259313 -0.29169795  0.08995880 -0.20127746
## Global rank 2014    0.471004785 -0.19106904  0.11007891 -0.23138882
## Market value $m    -0.458394801  0.20326155 -0.03274427 -0.03305945
## Net income $m      -0.407616480 -0.19996862  0.03936084 -0.04448147
## Total assets $m    -0.280379195 -0.31560221  0.30786911  0.17254042
## Employees          -0.263308129  0.04461274  0.17922605 -0.85664444
## Price $            -0.005047851 -0.02314386 -0.84865337 -0.27193859
## P/e ratio           0.132692888  0.62905936 -0.09399864  0.15010967
## Dividend yield (%) -0.140835821 -0.54704283 -0.34815934  0.20683729
##                            PC5         PC6         PC7         PC8
## Global rank 2015    0.01315155 -0.15045400  0.28420790 -0.20445500
## Global rank 2014    0.07776897 -0.25877187  0.29275341 -0.18829663
## Market value $m     0.20037396 -0.06159416  0.28941893 -0.78632200
## Net income $m       0.34689428 -0.18777246  0.60833937  0.51381737
## Total assets $m    -0.30003446 -0.71832360 -0.29422669 -0.06629101
## Employees          -0.37485179  0.12251814 -0.04497224  0.07183488
## Price $             0.11862673 -0.38967752 -0.19363081  0.03388369
## P/e ratio          -0.52824143 -0.29273331  0.41592071  0.12534634
## Dividend yield (%) -0.55646007  0.32063612  0.28391273 -0.11839510
##                              PC9
## Global rank 2015   -0.710250330
## Global rank 2014    0.694131032
## Market value $m    -0.019612824
## Net income $m      -0.025578745
## Total assets $m    -0.015995427
## Employees           0.003089727
## Price $            -0.026718331
## P/e ratio          -0.043298045
## Dividend yield (%)  0.099164783
```

You obtain 9 principal components, which shows the correlation between each variable and each prinicpal components.

```r
summary(pca)
```

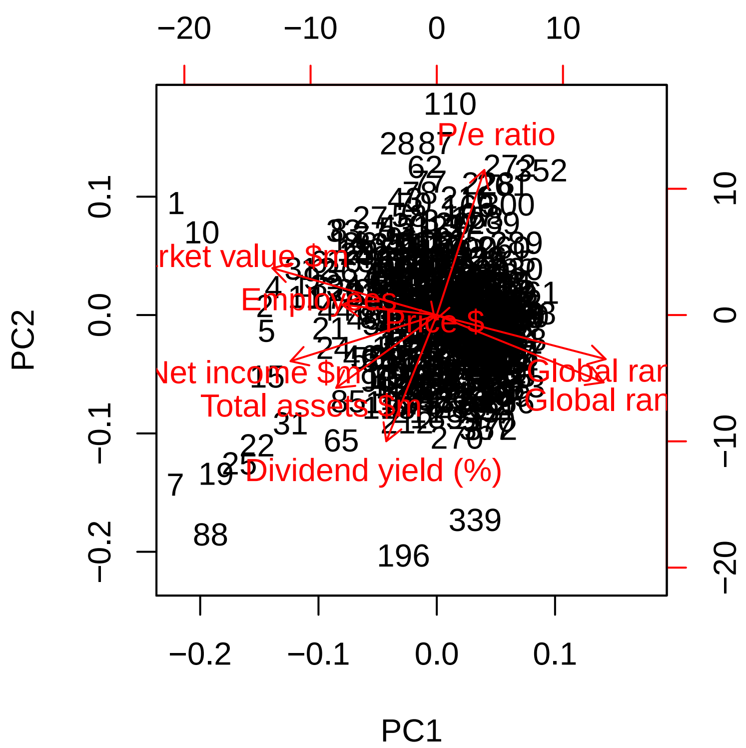
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6    PC7
## Standard deviation     1.8294 1.1757 1.0553 0.91348 0.86194 0.82726 0.7422
## Proportion of Variance 0.3718 0.1536 0.1237 0.09272 0.08255 0.07604 0.0612
## Cumulative Proportion  0.3718 0.5254 0.6492 0.74189 0.82443 0.90047 0.9617
##                            PC8     PC9
## Standard deviation     0.50709 0.29630
## Proportion of Variance 0.02857 0.00976
## Cumulative Proportion  0.99024 1.00000
```

For example, PC1 explains 37% of the total variance, which means that 37% of the information in the dataset can be encapsulated by just that one Principal Component.

PC2 explains 15% of the variance.

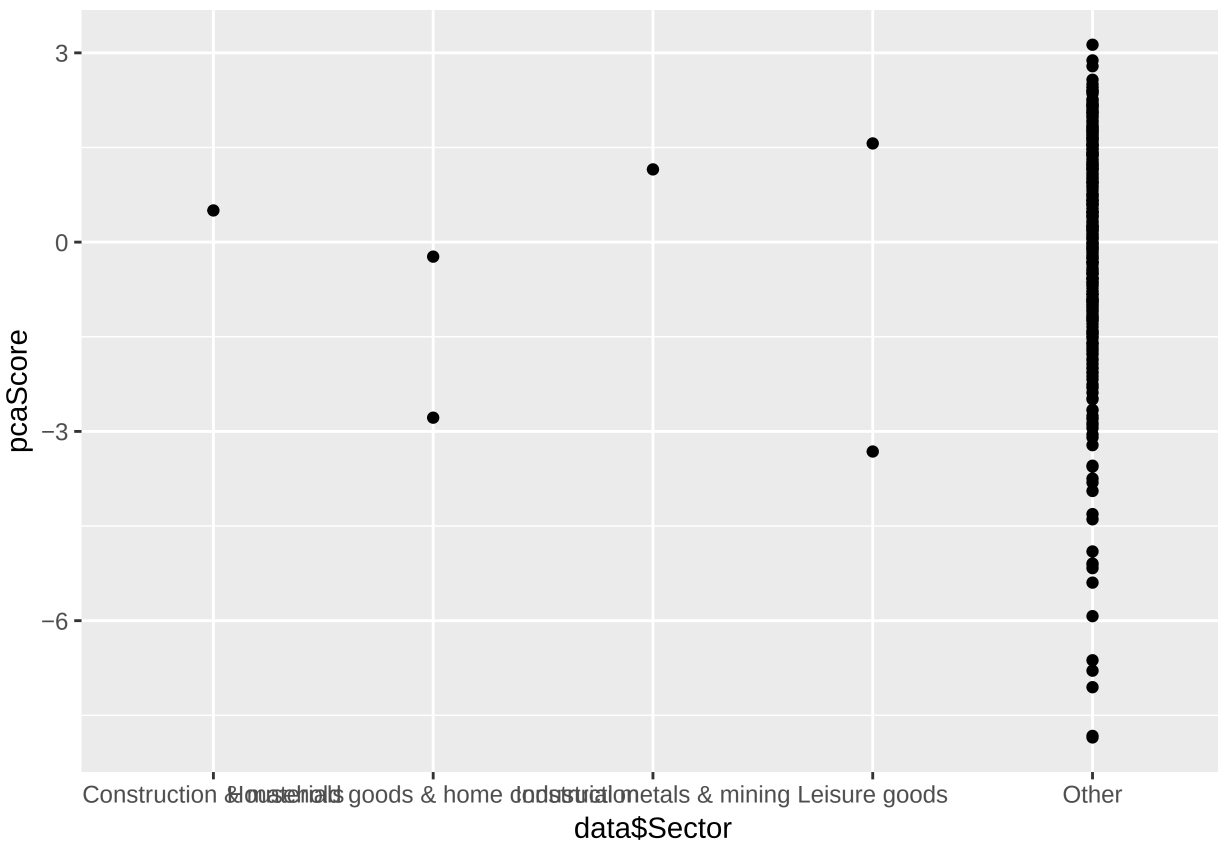
PC1 and PC2 can explain 52% of the variance.

```r
biplot(pca)
```




```r
#principal component score for each company
pcaScore <- pca$x[,1]

#Plot Sector and Score
ggplot(data=NULL, aes(x=data$Sector,y=pcaScore)) +
 geom_point()
```


3
0
−3
−6
pcaScore
Construction & materials
Household goods & home construction
Industrial metals & mining
Leisure goods
Other
data$Sector