College Scoreboard

Matt Richey

05/07/2020

Libraries

The usual libraries. You will probabilty need to install the factoextra package.

```
suppressMessages(library(tidyverse))
suppressMessages(library(factoextra))
```

Application of Clustering: College Scorecard Data

Here is a data set of private colleges. Data is taken from College Scorecard https://collegescorecard.ed.gov and https://collegescorecard.ed.gov/data/ I've created a subset of the data consisting of private colleges and some selected variables.

```
college.df <- read.csv("../Colleges2015.csv")
names(college.df)
```

```
##  [1] "INSTN"           "ADM_RATE"        "SAT_AVG"         "DEG_MS"
##  [5] "PERC_WHITE"      "PERC_PT"         "NET_PRICE"       "NET_PRICE_30k"
##  [9] "FAC_SAL"         "PCT_PELL"        "SIX_YR_CP"       "LOAN_DEF"
## [13] "PELL_DEBT_MDN"   "NOPELL_DEBT_MDN" "FAM_INC"
```

- ADM_RATE: admit rate
- SAT_AVG: Average SAT (ACT converted)
- DEG_MS: Degrees in Math/Stat
- PERC_WHITE: Percent White
- PERC_PT: Percent Part-time
- NET_PRICE: Net Price
- NET_PRICE_30k: Net Price 0-30k Income bracket
- FAC_SAL: Average Fac Salary
- PCT_PELL: Percent Pell Eligible
- SIX_YR_CP: 6 year completion rate
- LOAN_DEF: 3 year load default rate
- PELL_DEBT_MED: Median debt Pell Eligible
- NOPELL_DEBT_MED: Median debt Not Pell Eligible
- FAM_INC: Family Income

Pull of the instiution names.

```
inst <- college.df[,1]
```

Scale and check...

```
college.df <- data.frame(scale(college.df[,-1]))
colMeans(college.df)
```

```
##        ADM_RATE         SAT_AVG          DEG_MS      PERC_WHITE         PERC_PT
##   -1.712044e-16    2.381840e-16   -4.919547e-17   -2.527235e-17    6.386352e-18
##       NET_PRICE   NET_PRICE_30k         FAC_SAL        PCT_PELL       SIX_YR_CP
##    5.702422e-17    8.734654e-17   -1.644775e-17    1.310230e-16   -2.217234e-17
##        LOAN_DEF   PELL_DEBT_MDN NOPELL_DEBT_MDN         FAM_INC
##    6.133854e-17   -1.114335e-16   -2.128955e-16   -1.471816e-16
```

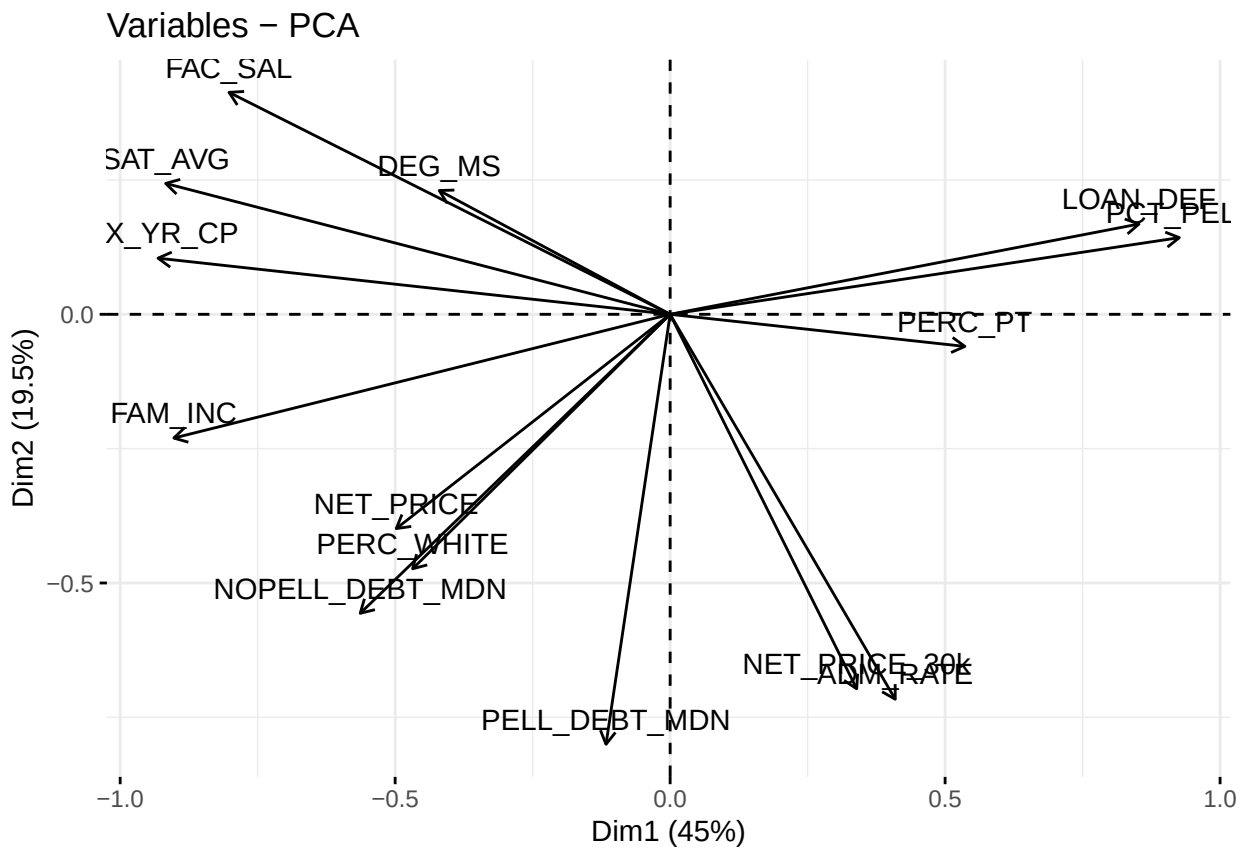
Intial Clustering with three clusters

```
K <- 3
mod.km <- kmeans(college.df,K,nstart=25)
mod.km$cluster
```

```
##   [1] 2 3 3 2 2 2 3 1 1 1 3 1 1 3 2 2 2 2 2 3 3 3 2 2 2 2 2 2 2 3 2 2 2 2 2 1 2
##  [38] 2 2 2 2 3 2 3 3 1 2 1 3 2 3 1 1 2 2 2 2 2 1 1 2 2 1 2 3 3 2 2 2 2 3 1 1 2
##  [75] 2 1 1 2 1 3 3 3 2 2 1 1 2 2 1 2 3 1 1 1 1 2 2 2 1 2 3 2 2 2 3 3 2 2 2 3 2
## [112] 1 2 2 2 2 2 2 2 2 1 2 3 2 2 1 1 3 2 2 2 2 2 1 2 2
```

Here's how the fviz_cluster works, with a little more detail Compute the principal components (more on this next time)

```
mod.pc <- prcomp(college.df)
fviz_pca_var(mod.pc)
```



Principal Components

The Principal Components method identifies a change af basis that better represents the variability of the data. Basically, the new basis vectors, in order, represent directions of greatest variability.

Pull off the rotation and check the dimension.

```
rot.mat <- mod.pc$rotation
dim(rot.mat)
```

```
## [1] 14 14
```

The dimensions should be determined by the number of predictors.

Change of basis time!!!

```
college.mat <- as.matrix(college.df)
college.rot <-  college.mat%*% rot.mat
```

Fix it up as a data frame.

```
collegeRot.df <- data.frame(college.rot)
```

We can rerun the kmeans on the rotated data frame.

```
mod.km2 <- kmeans(collegeRot.df,K,nstart=25)
```

As it turns out, this produces the same clustering.

```
table(mod.km$cluster,mod.km2$cluster)
```

```
##
##      1  2  3
##   1 31  0  0
##   2  0 77  0
##   3  0  0 27
```

Now add the clusters and the institional names

```
collegeRot.df$cluster <- factor(mod.km$cluster)
collegeRot.df$inst <- inst
```

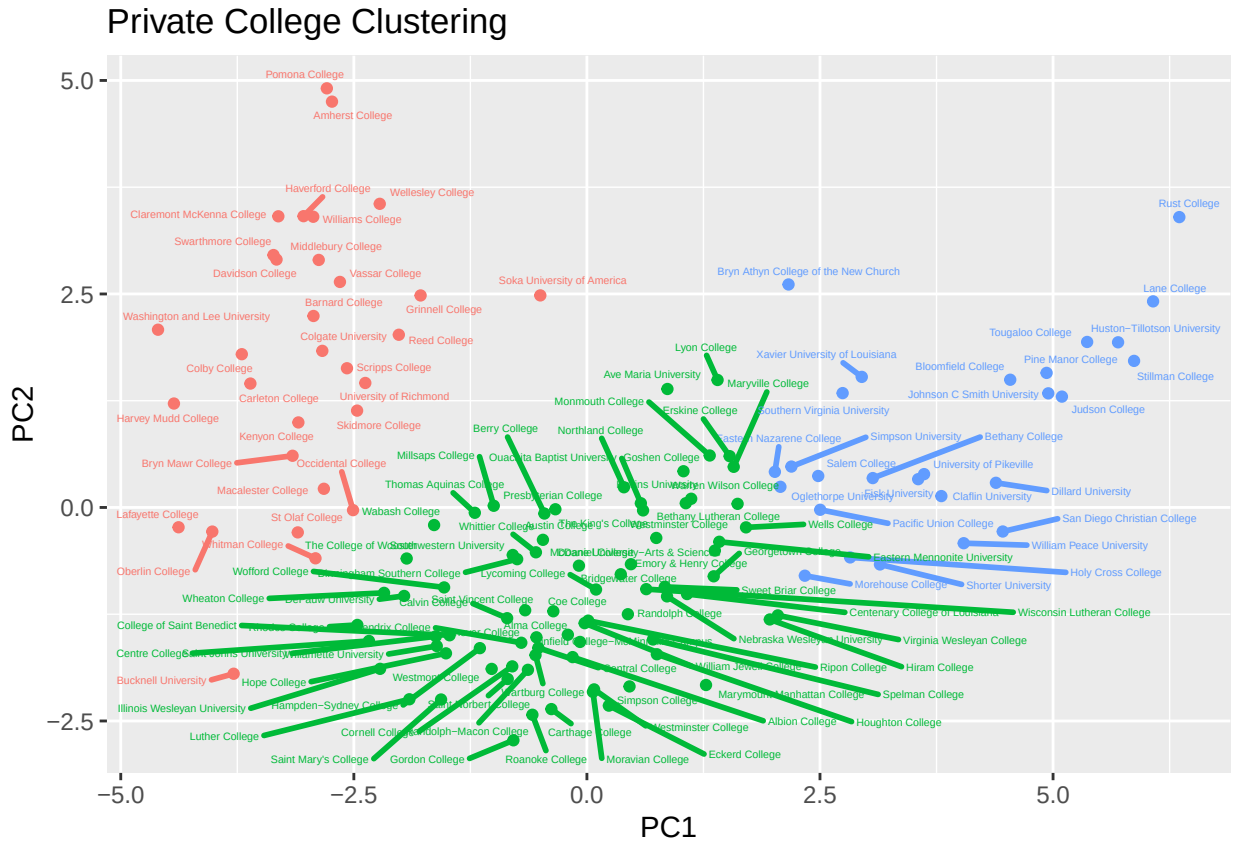
Here is the same vizualization as fviz_nbcluster only with the institution names included. We will add the loading vectors next.

Use the ggrepel library to offset the label names

```
library(ggrepel)
```

Ok...

```
collegeRot.df%>%
    ggplot()+
    geom_point(aes(PC1,PC2,color=cluster))+
    geom_text_repel(aes(PC1,PC2,color=cluster,label=inst),
                    size=1.5,segment.size = 1)+
    guides(color=F)+
    ggtitle("Private College Clustering")
```

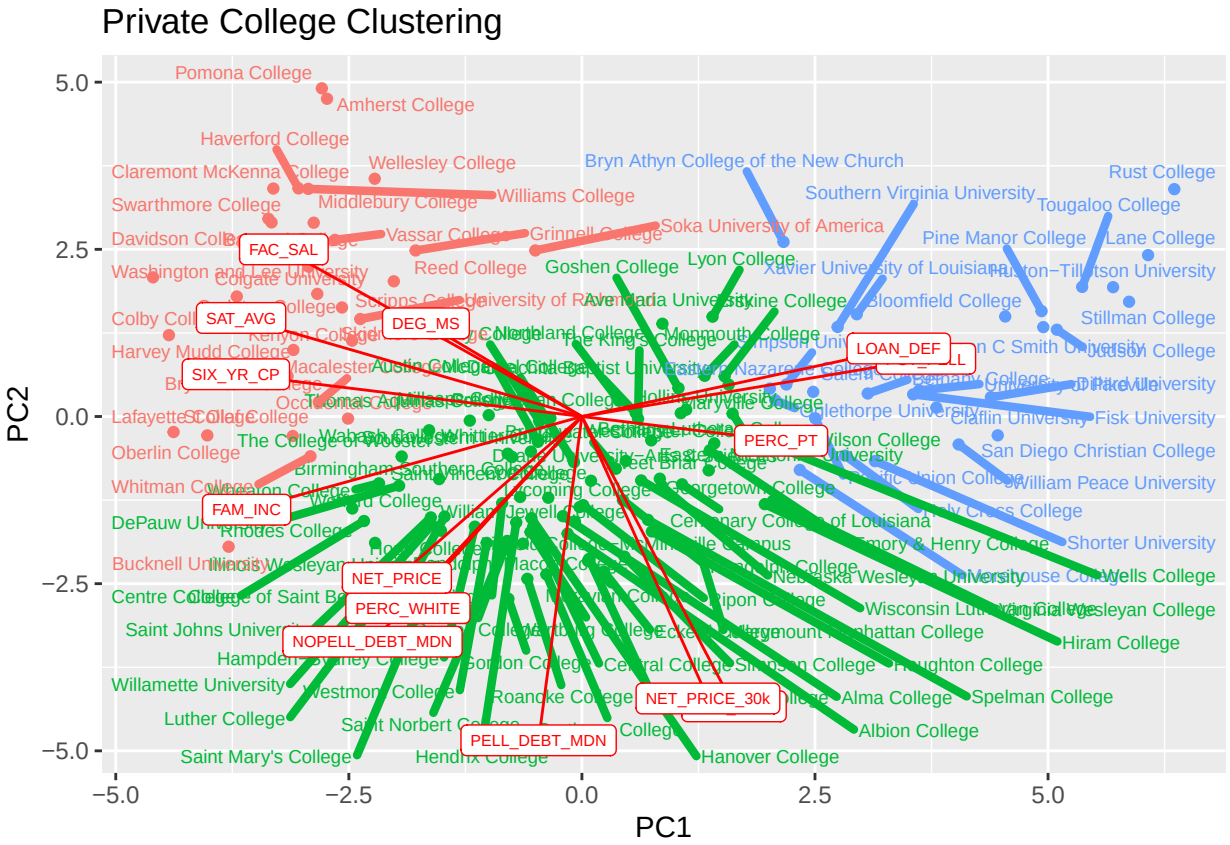


No add the loading vectors.

```
loading.df <- data.frame(rot.mat)
loading.df$pred <- rownames(rot.mat)
```

Redraw the picture. We need to scale the loading vectors to get everything in the picture properly.

```
load.scale <- 10
collegeRot.df%>%
  ggplot()+
  geom_point(aes(PC1,PC2,color=cluster))+
  geom_text_repel(aes(PC1,PC2,color=cluster,label=inst),
                  size=2.5,segment.size = 1.5)+
  geom_segment(data=loading.df,
               aes(x=0,xend=load.scale*PC1,
                   y=0,yend=load.scale*PC2),
               color="red")+
   geom_label(data=loading.df,
              aes(load.scale*PC1,
                  load.scale*PC2,
                  label=pred),
              color="red", size=2)+
  guides(color=F)+
    ggtitle("Private College Clustering")
```



We can infer a great deal about how private colleges cluster. The green cluster of schools (upper left corner) are the so-called "elite" schools with lots of resources. The loading show that they tend to have high faculty salaries, SAT scores, and 6-year completion rates. Interestingly, they also have a relatively high percentage of Math/Stat degrees. The blue cluster comprises pretty good schools with decent resources but with more of a focus on inclusion and diversity. We can see that the are aligned along the loading axes with predictors such as family income and PELL grants intertwined. The red cluster of schools appear to be "resource challenged" with high load defaults and perpendicular to qualities such as completion rates and SAT scores.