College ScoreBoard: PCA and K-means Clustering

Kshitij Gurung

5/6/2020

Introduction:

In this project, I attempt to classify, compare, and study 135 different private colleges in the U.S. bsaed on 14 different variables like admission rate, averaet SAT score, percentage on white and non-white, faculty salary, etc. I implement Principal Component Analysis (PCA) and K-means clustering to achieve my goal.

Libraries

Loading the usual libraries

```
## Load the libraries
suppressMessages(library(tidyverse))
suppressMessages(library(ggrepel))
suppressMessages(library(factoextra))
```

Reading the dataset

```
college.df <- read.csv("~/Desktop/Classes/ADM/Data/Colleges2015.csv")
head(as.tibble(college.df))
```

```
## # A tibble: 6 x 15
##   INSTN ADM_RATE SAT_AVG DEG_MS PERC_WHITE PERC_PT NET_PRICE NET_PRICE_30k
##   <fct>    <dbl>   <int>  <dbl>      <dbl>   <dbl>     <int>         <int>
## 1 Birm~    0.531    1167 0.0405     0.791   0.0067     21562         18551
## 2 Juds~    0.629    1031 0          0.780   0.332      13152         12695
## 3 Stil~    0.566     831 0          0.0581  0.0878     16988         15781
## 4 Lyon~    0.591    1120 0.065      0.748   0.003      17003         13631
## 5 Hend~    0.824    1270 0.0138     0.778   0.0031     25503         21841
## 6 Ouac~    0.684    1120 0.0207     0.850   0.0177     19251         14948
## # ... with 7 more variables: FAC_SAL <int>, PCT_PELL <dbl>, SIX_YR_CP <dbl>,
## #   LOAN_DEF <dbl>, PELL_DEBT_MDN <dbl>, NOPELL_DEBT_MDN <dbl>, FAM_INC <dbl>
```

Pulling off the variables and the schools name.

```
collegevars.df <- college.df[,-1]
collegeNames <- college.df[,1]
varNames <- colnames(collegevars.df)
varNames # variable names
```

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

The variables are as follows:

- 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

Principal Component Analysis (PCA)

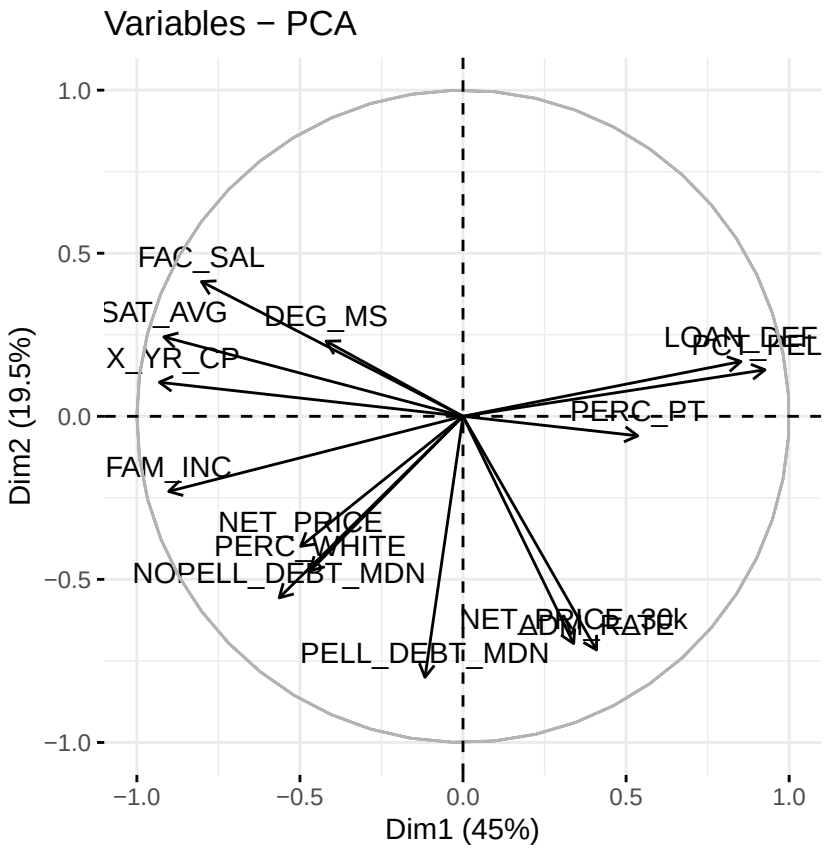
Now we can apply Principal Component Analysis to this data set. We will scale the data just in case.

```
college.mat <- data.matrix(collegevars.df)
rownames(college.mat) <- collegeNames # adding collage names/observations
mod.pca <- prcomp(college.mat,scale=T)
```

```
summary(mod.pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.5089 1.6526 1.11430 1.05662 0.86406 0.82450 0.58989
## Proportion of Variance 0.4496 0.1951 0.08869 0.07975 0.05333 0.04856 0.02486
## Cumulative Proportion  0.4496 0.6447 0.73336 0.81310 0.86643 0.91499 0.93984
##                           PC8     PC9    PC10    PC11   PC12    PC13    PC14
## Standard deviation     0.45115 0.40245 0.38392 0.33788 0.2970 0.25949 0.24410
## Proportion of Variance 0.01454 0.01157 0.01053 0.00815 0.0063 0.00481 0.00426
## Cumulative Proportion  0.95438 0.96595 0.97648 0.98463 0.9909 0.99574 1.00000
```

```
fviz_pca_var(mod.pca)
```



- A quick glimse on pca dimesions and the variables loading.

Building a better biplot

Let's pull off the rotated scores. These are the points in the biplot.

```
scoresRotated <- mod.pca$x
rotation.mat <- mod.pca$rotation
```

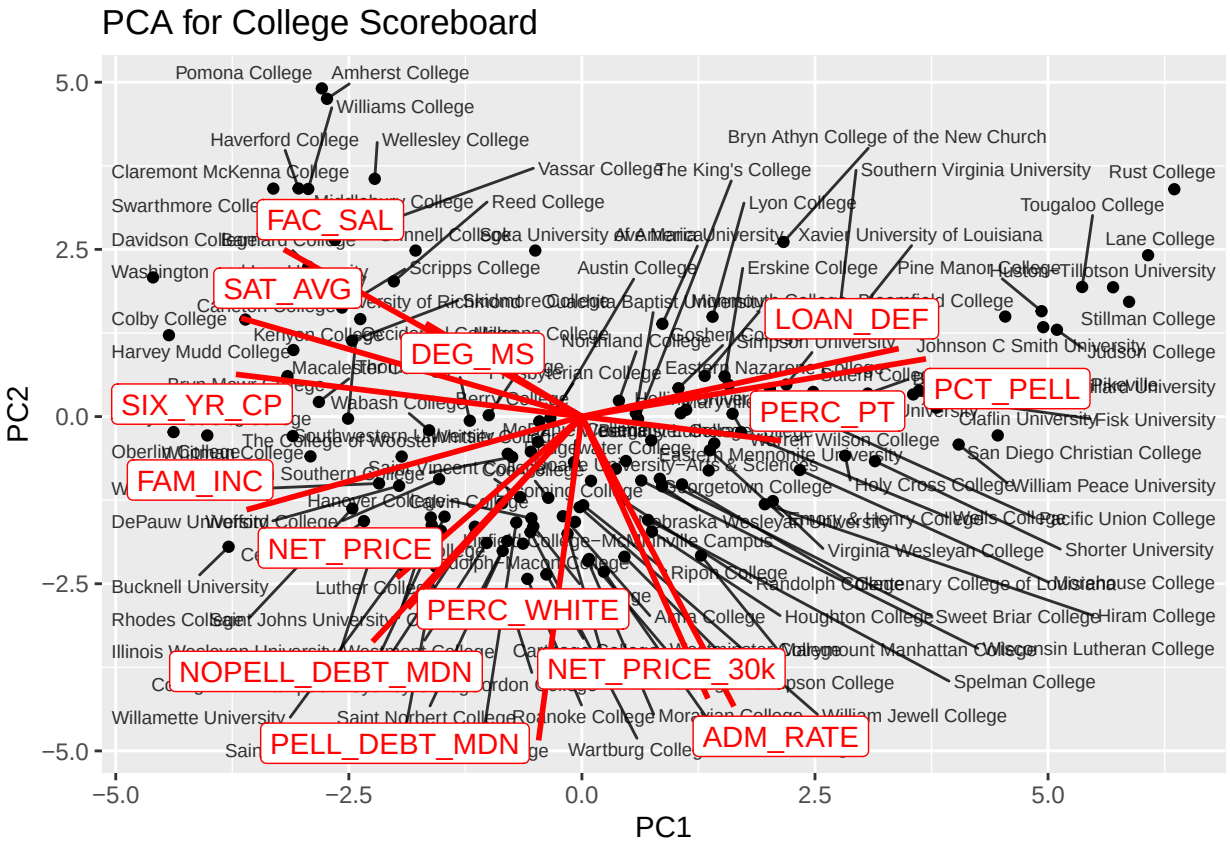
Put the PCA info into data frames.

```
scoresRotated.df <- data.frame(scoresRotated)
scoresRotated.df$names <- collegeNames
rotation.df <- data.frame(rotation.mat)
rotation.df$events <- varNames
```

Now we can build the biplot from scratch.

- The points are the individual college's scores rotated into the basis defined by the Principal Components. We only see the first two components.

```
sc <- 10 ## get everything on the same scale
scoresRotated.df %>%
  ggplot()+
  geom_point(aes(PC1,PC2))+
  geom_text_repel(aes(PC1,PC2,label=names), size=2.5,segment.size = .5, alpha = .8)+
  ## Add the loadings,  these are just the coordinates in the PC1 and PC2 vectors
  geom_segment(data=rotation.df,
               aes(x=0,y=0,xend=sc*PC1,yend=sc*PC2),size=1,color="red")+
  geom_label_repel(data=rotation.df,
                   aes(sc*PC1,sc*PC2,label=events),color="red")+
  labs(title="PCA for College Scoreboard")
```



- The PC1 axes tends to feature a rate measurement like Percent Pell Eligible (PCT_PELL), 6 yr completion rate (SIX_YR_CP), Percent part-time (PERC_PT), and soon. The PC2 axes tends to measure larger numbers like Median debt Pell Eligible (PELL_DEBT_MDN) and not Eligible, Net Price (NET_PRICE_30K), Faculty salary (FAC_SAL) and so forth.

K-means clustering

Now, we will include a k-means clustering on top of the PCA analysis. We will use an elbow plot to determine a candidate for the optimal number of clusters. We weill also include a plot combining PCA and clustering. Does this show anything special about the schools?

Let's do some k-means clustering in search of the optimal number of clusters The plan is simple, for each value k=1,2,...M (large enough), build a k-means clustering. For each, extract the Total Within Sum of

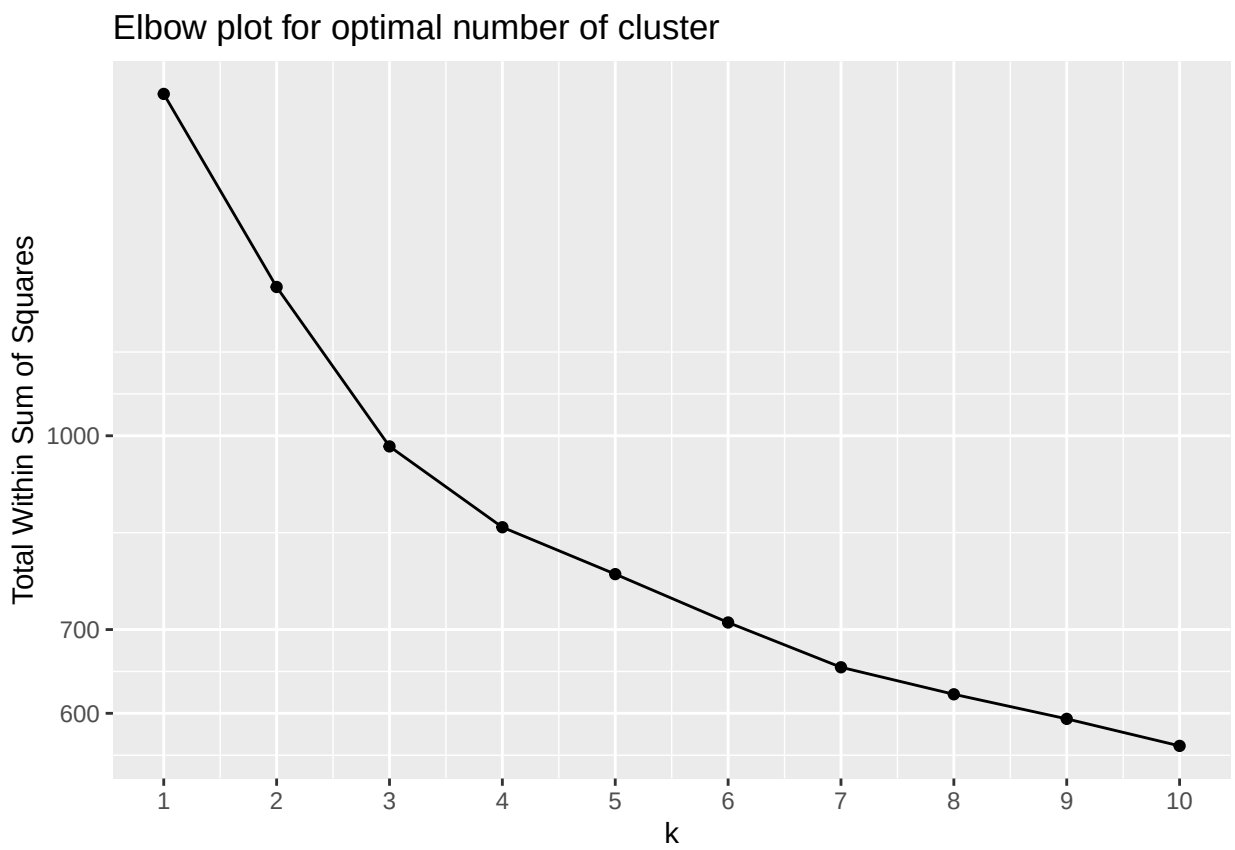
Squares (tot.withinss).

Building k cluster on top of the PCA analysis.

```
## Search up to M
M <- 10
twissVals <- numeric(M)
for(k in 1:M){
    mod.kmeans <- kmeans(scoresRotated.df[,-15],centers=k,nstart=25)
    twissVals[k] <- mod.kmeans$tot.withinss
}
```

What do we have?

```
data.frame(k=1:M,
           twiss=twissVals) %>%
    ggplot()+
    geom_point(aes(k,twiss))+
    geom_line(aes(k,twiss))+
  scale_y_log10()+
    scale_x_continuous(breaks=1:M)+
  labs(title='Elbow plot for optimal number of cluster', y='Total Within Sum of Squares')
```



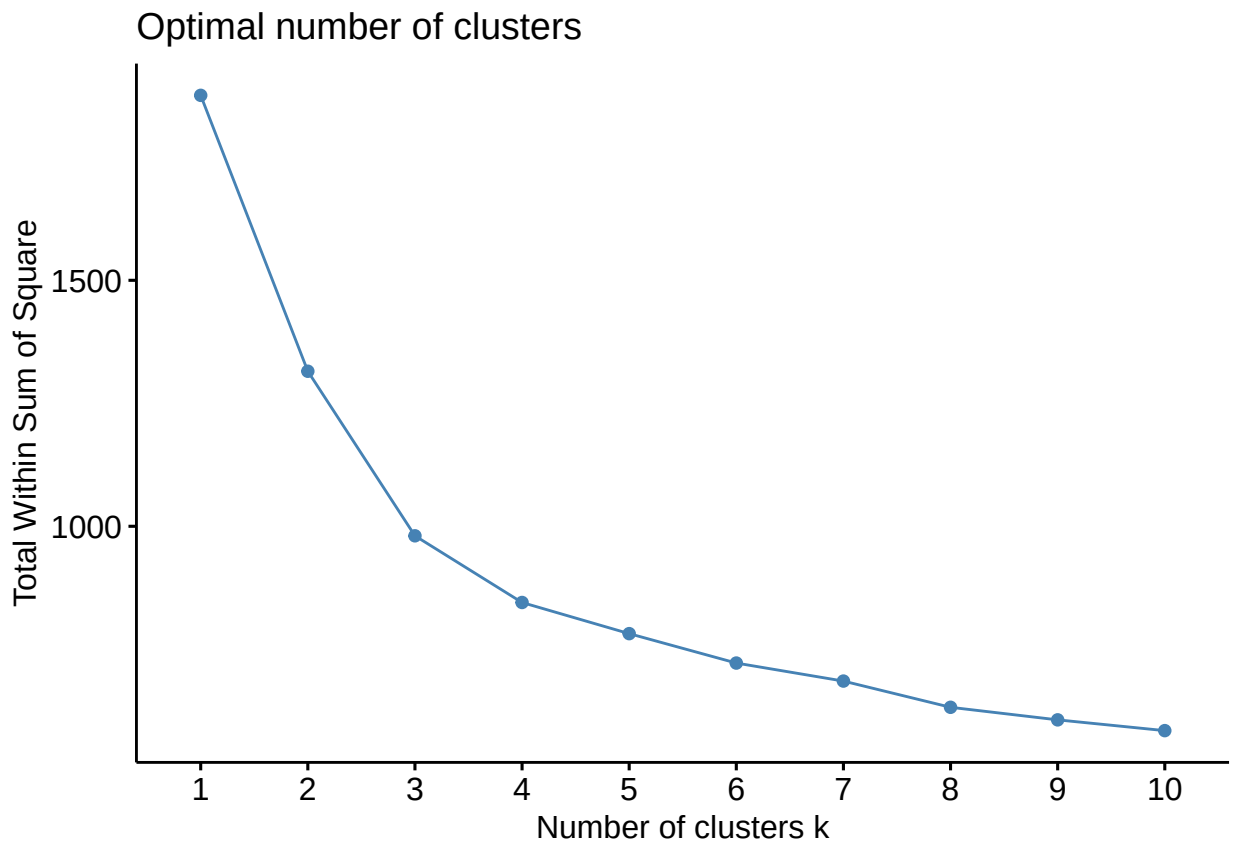
We do not see a sharp "elbow" but around k=6 seems to indicate that this is a good choice for the clustering.

In general, this elbow plot can help identify the optimal k. In practice, the elbow can be hard to precisely identify. Close enough is good enough. Of course, in practice, we would use some sort of train/test to see if the elbow persists.

Let's use factoextra package to run combine both k-means clustering and PCA to generate a nice 2-D reduced cluster plot.

However, we can still cluster. Just to be safe, let's scale the data and repack into data frame.

```
data.df <- scale(college.mat)
data.df <- data.frame(data.df)
fviz_nbclust(data.df,kmeans,method="wss")
```



- Looks like cluster k=4 might be a good choice. Note: This figure and previous are almost identical. It is because the former one was when we ran kmeans on top of the PCA whereas the later one is running factoextra packege onto the original dataset, which uses PCA under the hood.

Let's apply k-means with, say, K=4 means.

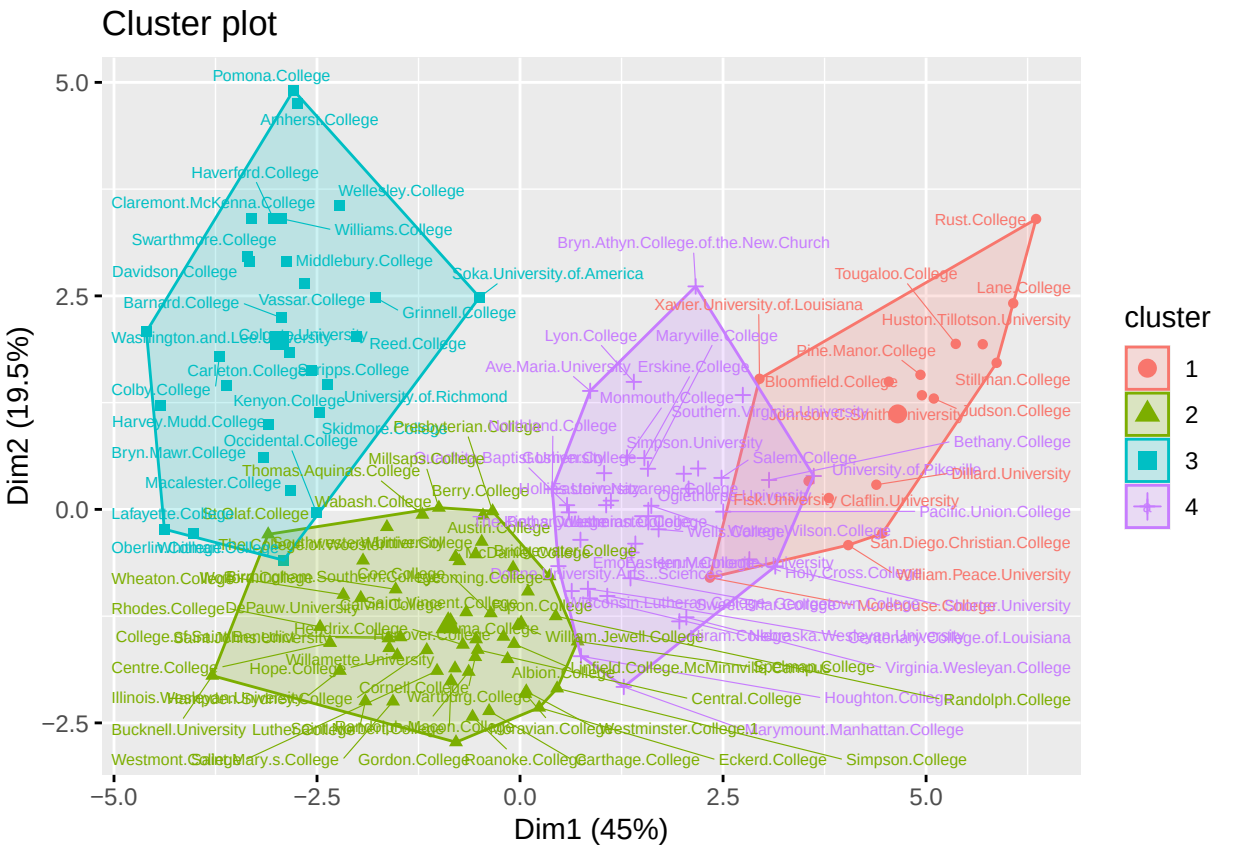
```
K<-4
mod.km <- kmeans(data.df,K,nstart=25)
data.df$cluster <- factor(mod.km$cluster)
```

What do we do now, we have a clustering, but how does it look?

Here's the plan: Perform a Principal Component Analysis and project into 2-dimensional space. Carry the clusters along with the projection and see what we have.

I.e, fviz_cluster will project onto the "best" two dimensions. This is essentially the biplot with clustering information included.

```
## make sure we only use the original data!
fviz_cluster(mod.km,data=data.df[,-15], labelsize = 6, repel = T)
```



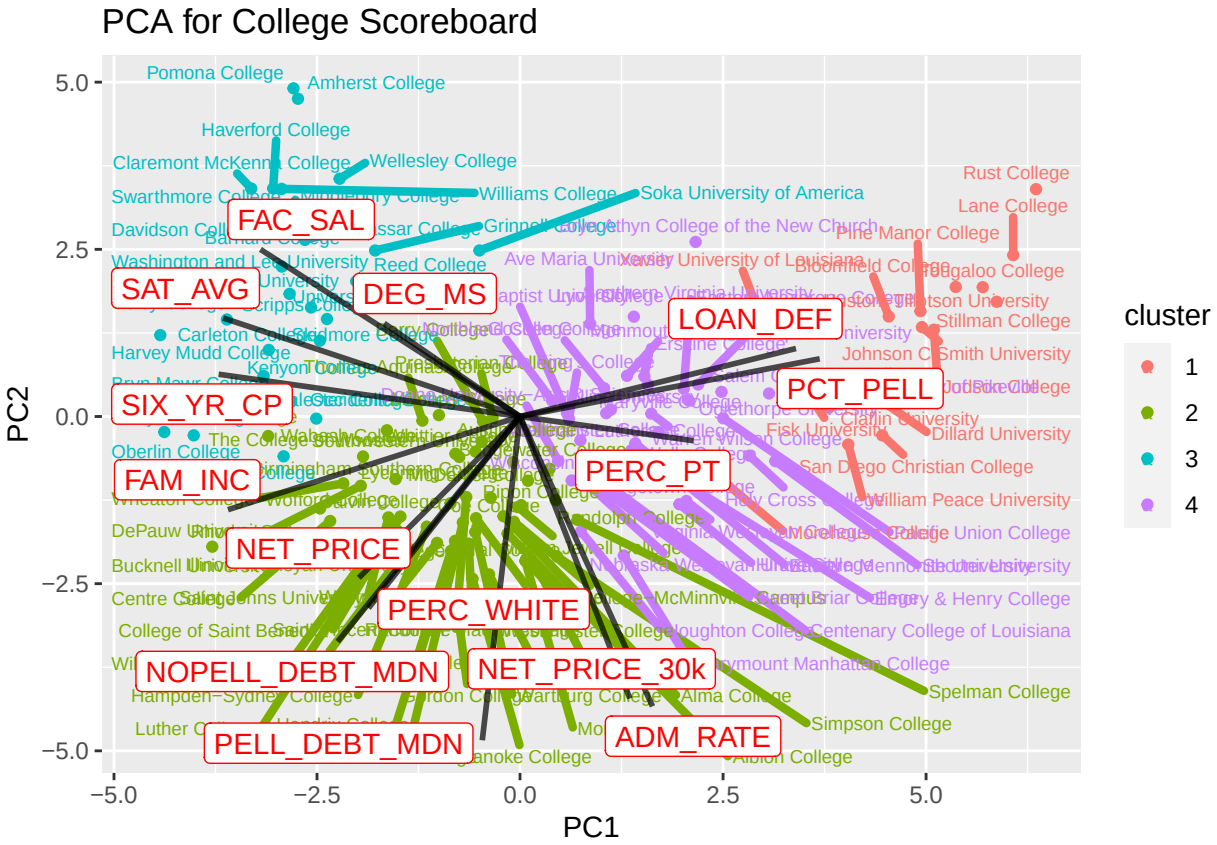
- These clusters seem to be agree with some of the PCA generate vectors (directions) in the PCA Collegescore figure from part 1. However, since our cluster plot diagram does not have varibale names in the cluster area, it is hard to tell what these cluster actually represnt. Nevertheless, it looks pretty cool.

Now, let's add the clusters and the institutions name.

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

Redraw the PCA picture with clusters.

```
sc <- 10 ## get everything on the same scale
scoresRotated.df %>%
  ggplot()+
  geom_point(aes(PC1,PC2, color=cluster))+
  geom_text_repel(aes(PC1,PC2,label=names, color=cluster), size=2.5,segment.size = 1.5)+
  ## Add the loadings,  these are just the coordinates in the PC1 and PC2 vectors
  geom_segment(data=rotation.df,
               aes(x=0,y=0,xend=sc*PC1,yend=sc*PC2),size=1,color="black", alpha = 0.7)+
  geom_label_repel(data=rotation.df,
                aes(sc*PC1,sc*PC2,label=events),color="red")+
  labs(title="PCA for College Scoreboard")
```



- We can infer a great deal about how private colleges cluster. The blue 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 green 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 and purple cluster of schools appear to be "resource challenged" with high load defaults and perpendicular to qualities such as completion rates and SAT scores.