Appollis\_PartB

The data set **Universities.csv** contain information on 1,302 American colleges and universities offering an undergraduate program. For each university, there are 17 measurements that include continuous measurements (such as tuition and graduation rate) and categorical measurements (such as location by state and whether it is a private or public school). You should remove all categorical variables and missing numeric measurements from the data set prior to doing the tasks below.

1. Conduct a principal component analysis (PCA) of the given data set and comment on the results.
2. Should the data be normalized? Discuss what characterizes the components you consider key.

**Answer:**

1. There is a lot of missing data, with variables like, room, board, add fees and the % of students from the top 10% having the highest number of missing data. There are only 471 schools that have populated all the variables. There is one school that has only populated 5 of the 20 variables and a total of 37 schools that has populated 12 or fewer variables.

I have conducted a principal component analysis (PCA) on the University data set and this is my results.

The table below is an extract of the result of PCA. This is a view of the standard deviation of the different principal components, with the variance explained for the different principal components. These principal components is in a hierarchical order with the pc1 explaining the biggest portion of the variance.

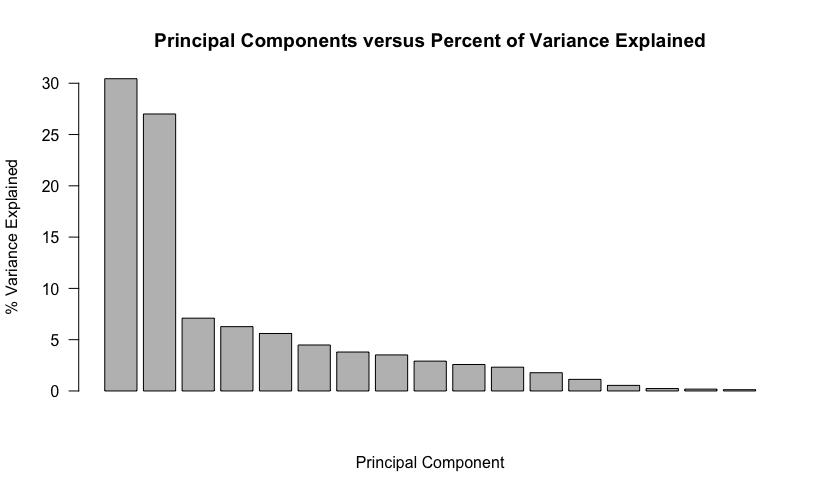
PC1 PC2 PC3 PC4 PC5 PC6 PC7

Standard deviation 2.27 2.14 1.10 1.03 0.98 0.87 0.80

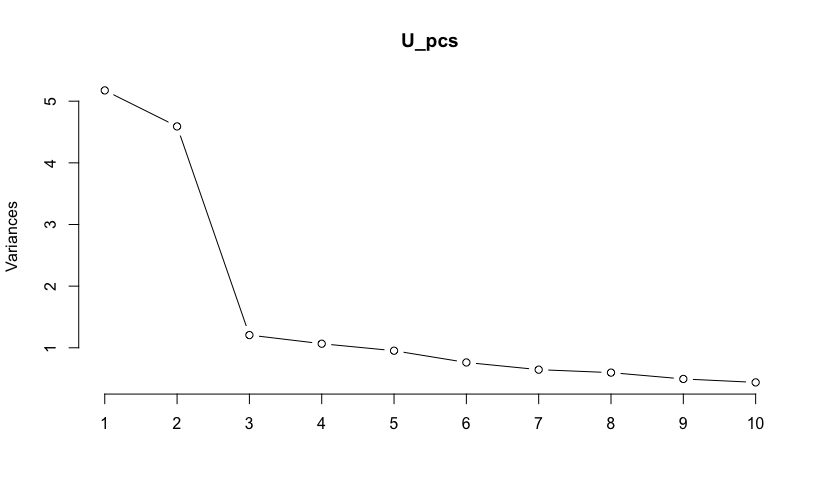
Proportion of Variance 30.44% 27.00% 7.10% 6.27% 5.60% 4.48% 3.80%

Cum Proportion 30.44% 57.44% 64.54% 70.81% 76.41% 80.89% 84.69%

Based on the table and histogram below 84.69% of the variance can be explained by the first 7 pc’s.



Based on the elbow plot below – I choose the number of pc (principal components) of 7.



1. It is important to normalize the data as some of the variables are in absolute numbers and some variables is in percentage. If I run the model on these numbers at face value it will be bias towards large numbers.

The characteristics of the different components are:

* PC1: These schools are characterized by out of state and instate tuition as well as the percentage of students in the top 10% and the percentage of students in the top 25%.
* PC2: These universities are characterized by the number of students enrolled and the number of full time undergrads as well as the number of applications received and accepted.
* PC3: These universities are characterized by the estimated book cost and the estimated personal cost, and to a lesser extend room and boarding fees.
* PC4: These universities are characterized by room fees, additional fees, estimated personal cost and the percentage of students in the top 25% and the top 10%
* PC5: Additional fees, estimated book cost, and the number of part-time characterized these universities.
* PC6: Student faculty ratio, estimated personal cost, book cost, and additional fees characterized these universities.
* PC7: These universities are characterized by the percentage of faculty that has a PHD as well as the number of part-time undergrads.

The scatter plot look like this.

