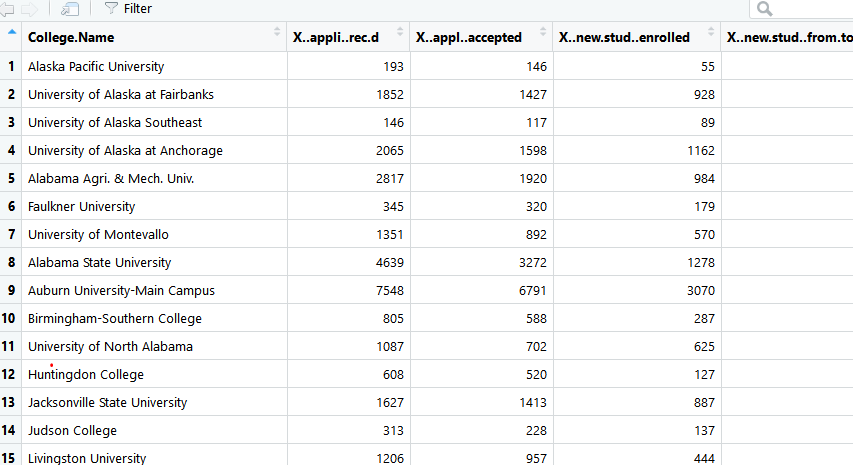
****Problem 4.2.****

****A.****

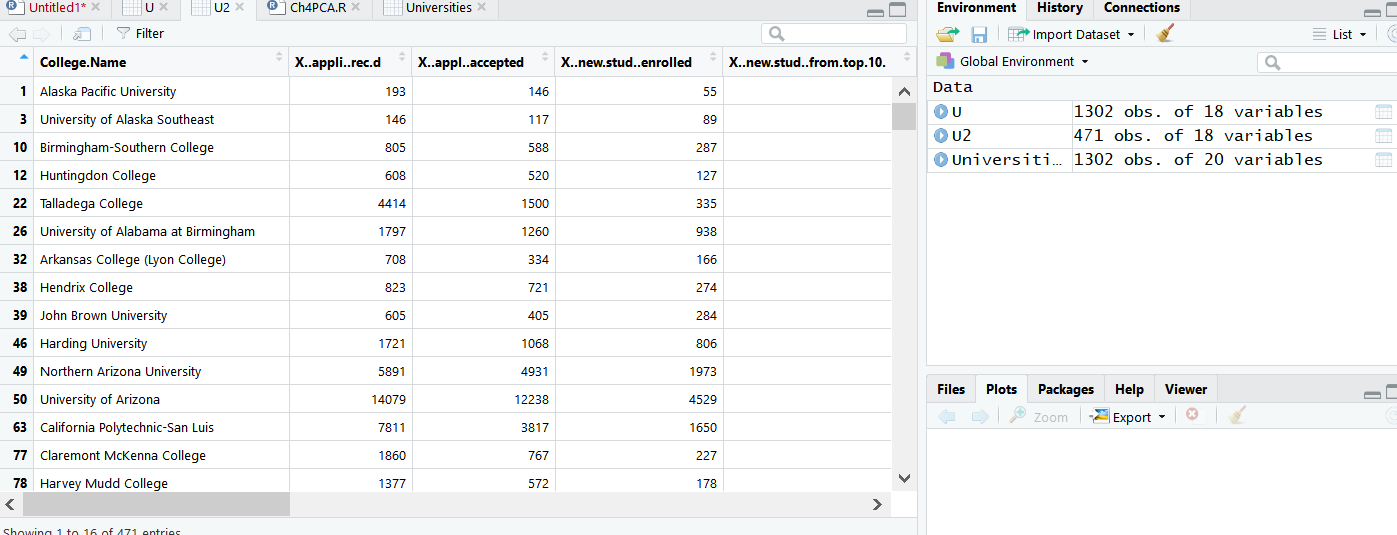
****#Remove all categorical variables****

****U <- Universities[,-c(2,3)]****



****#Remove all records with missing numerical measurements****

****U2 <- na.omit(U)****

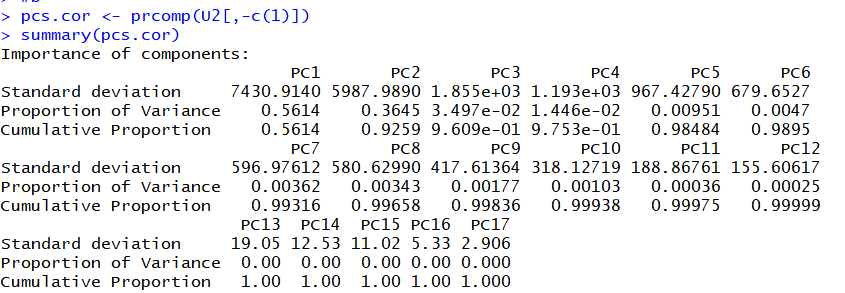


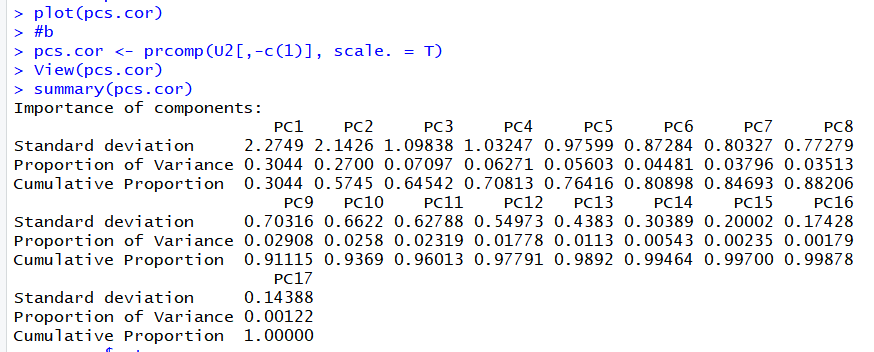
There were 1302 original variables, after removing. There are 471 variables in the dataset.

B.

No,I think this data set needn’t to be normalized. Because the variables which belong to a same attribute (dollars, number of things, percentage) are using the same unit.

But the result of them has a huge difference, so I choose to normalize it.





The first 9 principal components account for more than 90% variability of all.

The first principal component is most affected by the *out of state tuition*

The second principal component is most affected by the *students come form top10%, application accepted and application received*

The third principal component is most affected by the *estimate..book cost*

The fourth principal component is most affected by the *room,add..fees and estim..personal*

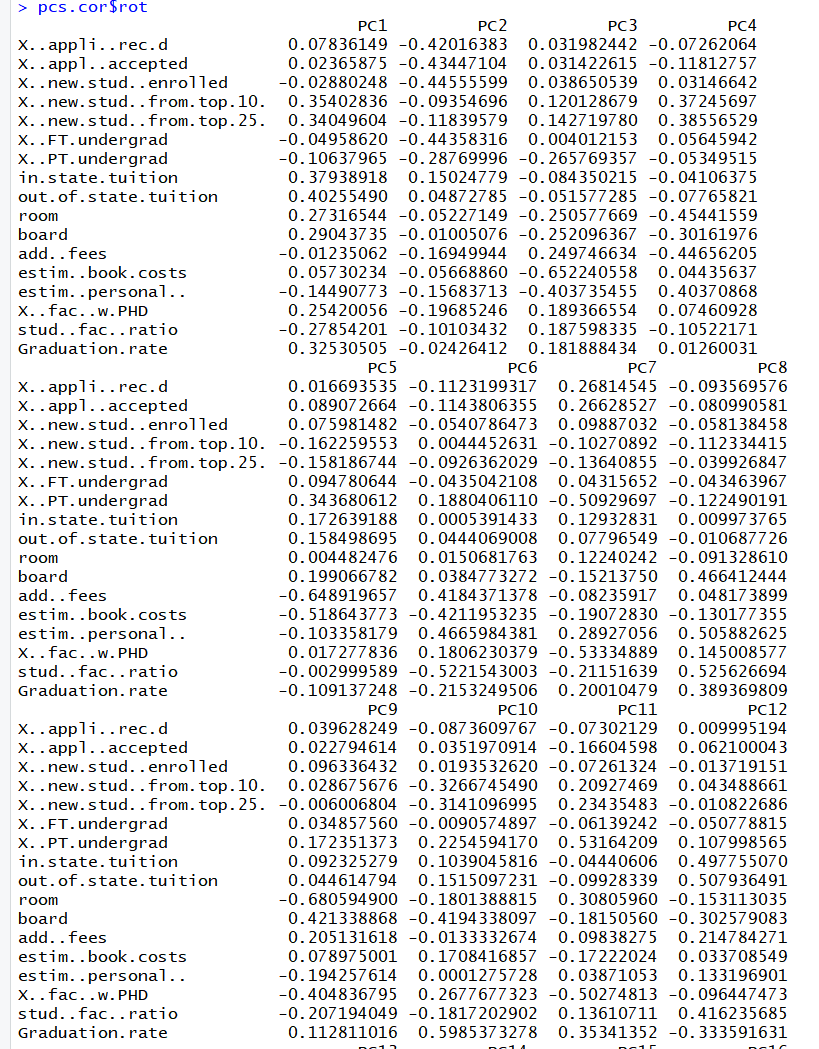
The fifth principal component is most affected by the *add.fees, book cost.*

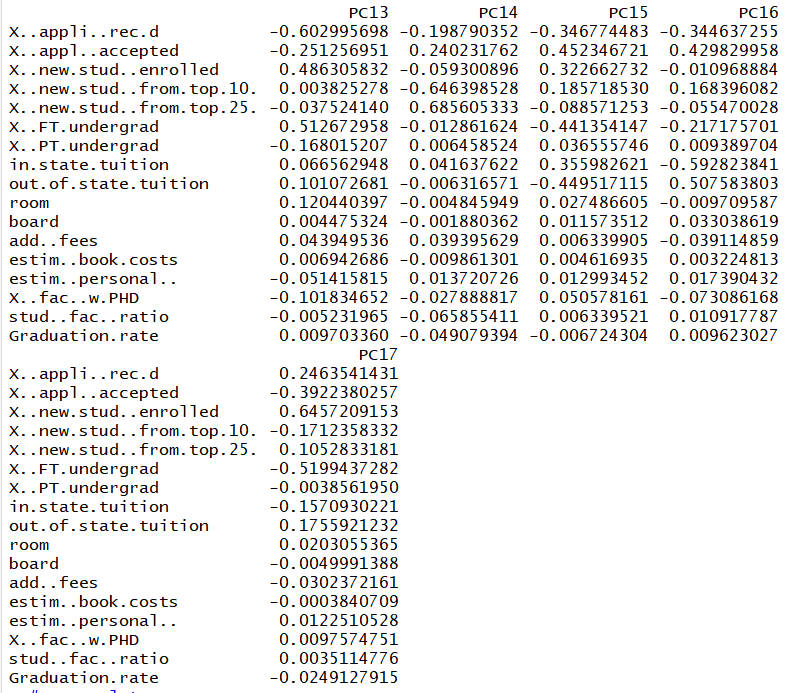
The sixth principal component is most affected by the *stud.fac.ratio, add.fees, book cost and estim..personal*

The seventh principal component is most affected by the *x..PT.undergrad and x.fac..w.phd*

The eighth principal component is most affected by the *board, estim..personal and stud..fac..ratio*

The ninth principal component is most affected by the *room,board and x..fac..w.phd*





In part (b), also create a scree plot.

