UNIVERSIDAD CARLOS III

FINAL PROJECT

PROFESSOR : Pedro Galeano

SUBJECT : Statistical Learning

DEGREE : Master in Big Data Analytics

STUDENTS : Miguel Ángel De Moya Jiménez, NIA: 100467131
Miguel Zabaleta Sarasa, NIA: 100463947
Javier López-Tello Morales, NIA: 100464692

INDEX

0. Description of the dataset

The dataset used to carry out this project contains data about America's Top College Ranking of 2019 from Forbes Magazine. It contains information about 650 colleges from the United States, ranked based on alumni salary (20%), student satisfaction (20%), debt (20%), American leaders (15%), on-time graduation rate (12.5%), and academic success (12.5%), along with various other statistics pertaining to each school.

The exact variables present in the dataset are the following:

- **Rank**: College ranking of 2019 by Forbes Magazine.
- **Name**: Name of the college.
- **City**: City where campus is located.
- **State**: State where campus is located.
- **Public.Private**: Whether school is publicly or privately funded.
- **Undergraduate.Population**: Number of enrolled undergraduate students.
- **Student.Population**: Total number of students enrolled.
- **Net.Price**: Average cost for one year of education, subtracting any financial aid received by the students.
- **Average.Grant.Aid**: Average amount of money students receive each year to help pay for college, from sources such as the government.
- **Total.Annual.Cost**: Total cost of tuition, room and board, and any additional fees that the college charges per year.
- **Alumni.Salary**: Median salary for workers with 10 or more years of experience.
- **Acceptance.Rate**: Percentage of students who apply to a college that are admitted.

- **SAT.Lower**: Average first quartile composite SAT score.
- **SAT.Upper**: Average third quartile composite SAT score.
- **ACT.Lower**: Average first quartile composite ACT score.
- **ACT.Upper**: Average third quartile composite ACT score.
- **Website**: College website url.

There are 6 categorical variables and 11 quantitative variables.

We can see in the next screenshots a sample with the first 5 observations of the dataset:

Rank	Name	City	State	Public.Private	Undergraduate.Population	Student.Population	Net.Price
1	Harvard University	Cambridge	MA	Private	13844	31120	14327
2	Stanford University	Stanford	CA	Private	8402	17534	13261
3	Yale University	New Haven	CT	Private	6483	12974	18627
4	Massachusetts Institute of Technology	Cambridge	MA	Private	4680	11466	20771
5	Princeton University	Princeton	NJ	Private	5659	8273	9327

Average.Grant.Aid	Total.Annual.Cost	Alumni.Salary	Acceptance.Rate	SAT.Lower	SAT.Upper	ACT.Lower	ACT.Upper	Website
49870	69600	146800	5	1460	1590	32	35	www.harvard.edu
50134	69109	145200	5	1390	1540	32	35	www.stanford.edu
50897	71290	138300	7	1460	1580	32	35	www.yale.edu
43248	67430	155200	7	1490	1570	33	35	web.mit.edu
48088	66150	139400	6	1430	1570	31	35	www.princeton.edu

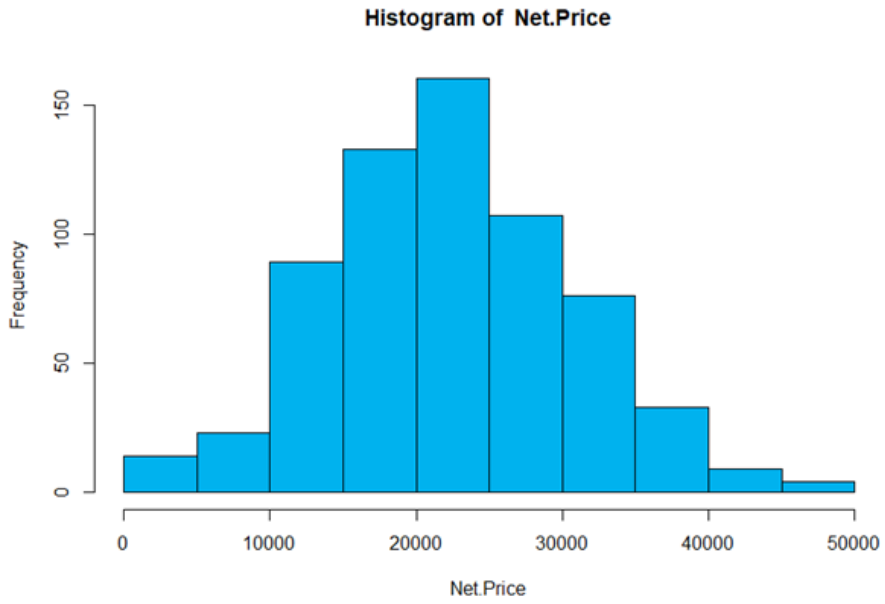
Analyzing this dataset could provide us with valuable insights into the top universities of the United States, and hopefully, there will be interesting findings regarding which colleges are worth attending the most.

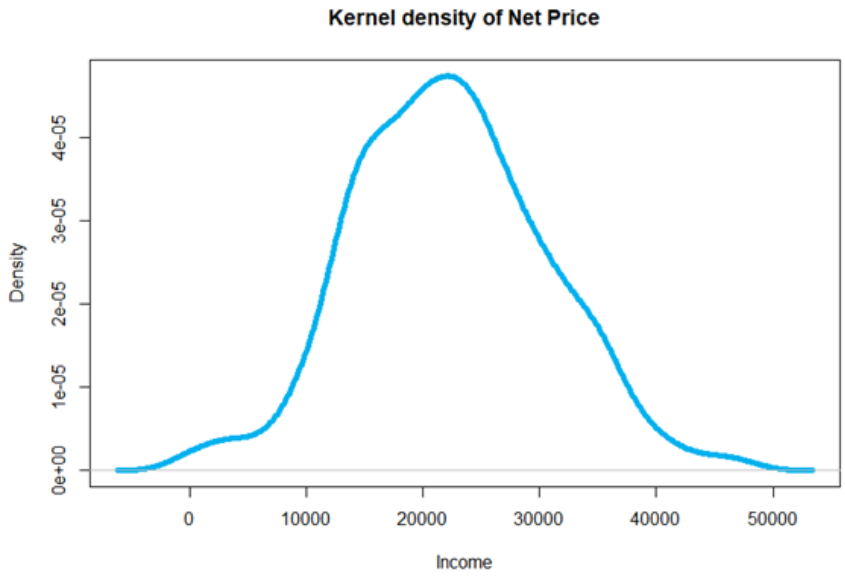
1. Graphical analysis

We start by plotting the histograms of all the continuous variables in order to search for distinguished distributions and differences between groups. We will also plot their respective kernel density plot when considered appropriate.

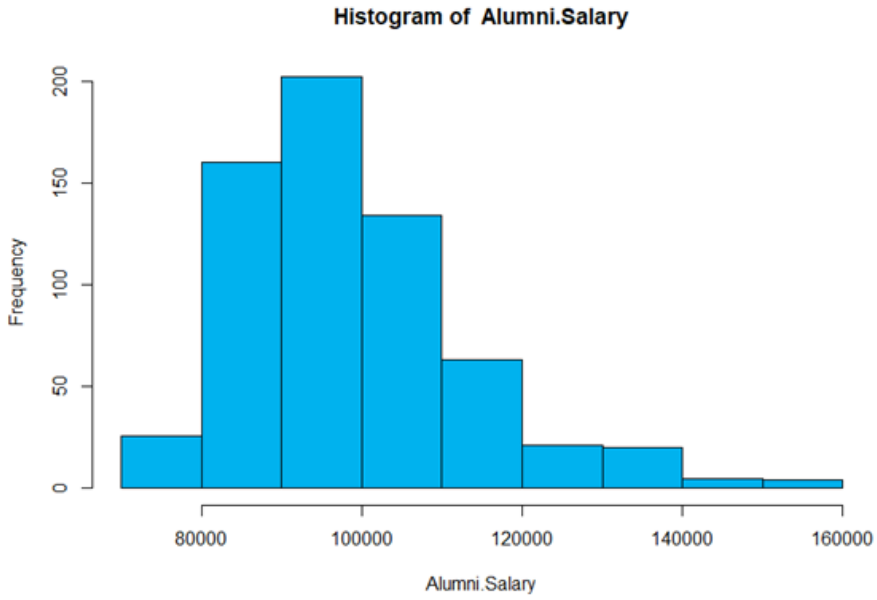
The results show a few remarkable findings, which are the following:

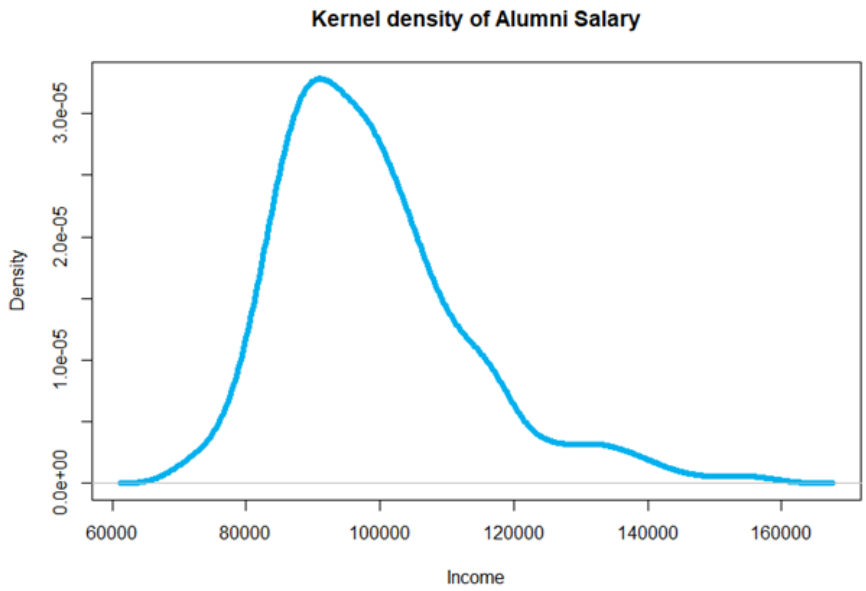
- Variables with a somewhat apparent Gaussian distribution: only *Net Price*

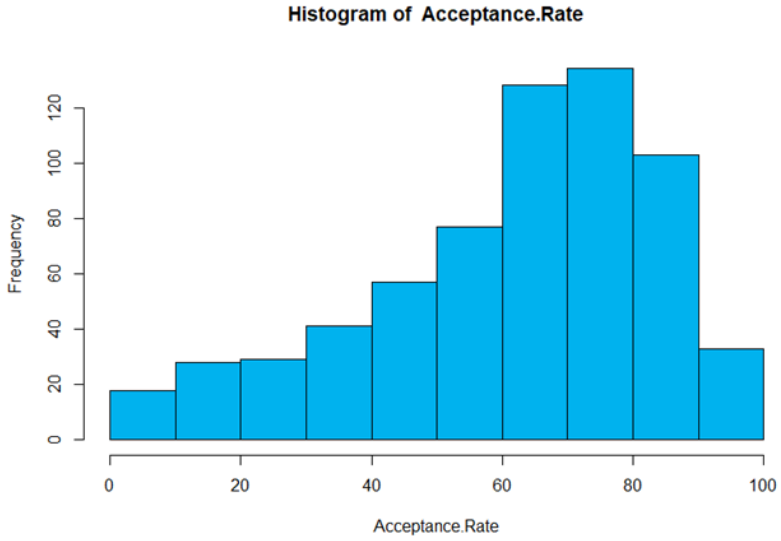


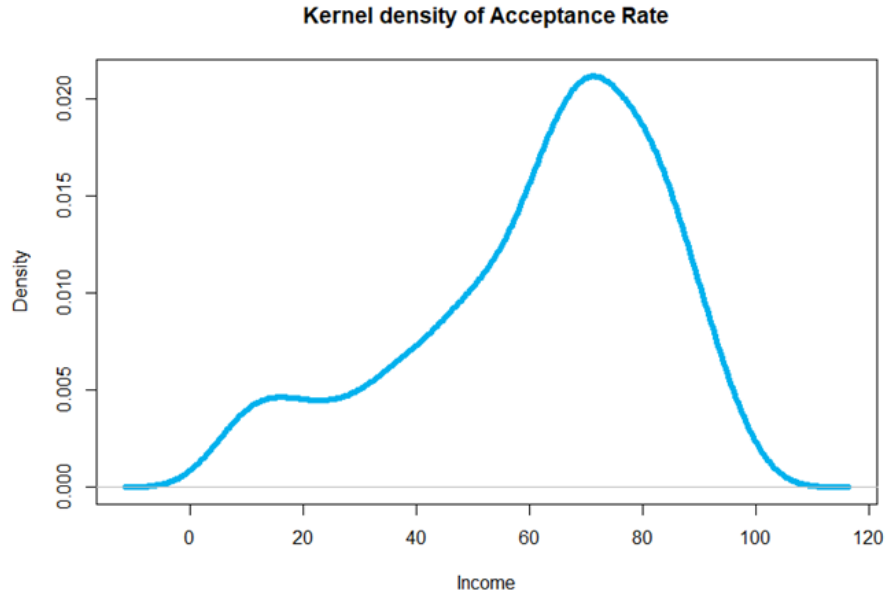


- Gaussian variables with a strong level of skewness: *Alumni Salary* (right), *Acceptance Rate* (left), *SAT Lower* (right); and variables slightly skewed: *SAT Upper* (right), *ACT Lower* (right), *ACT Upper* (right).


Histogram of Alumni.Salary
Frequency
0
50
100
150
200
80000
100000
120000
140000
160000
Alumni.Salary

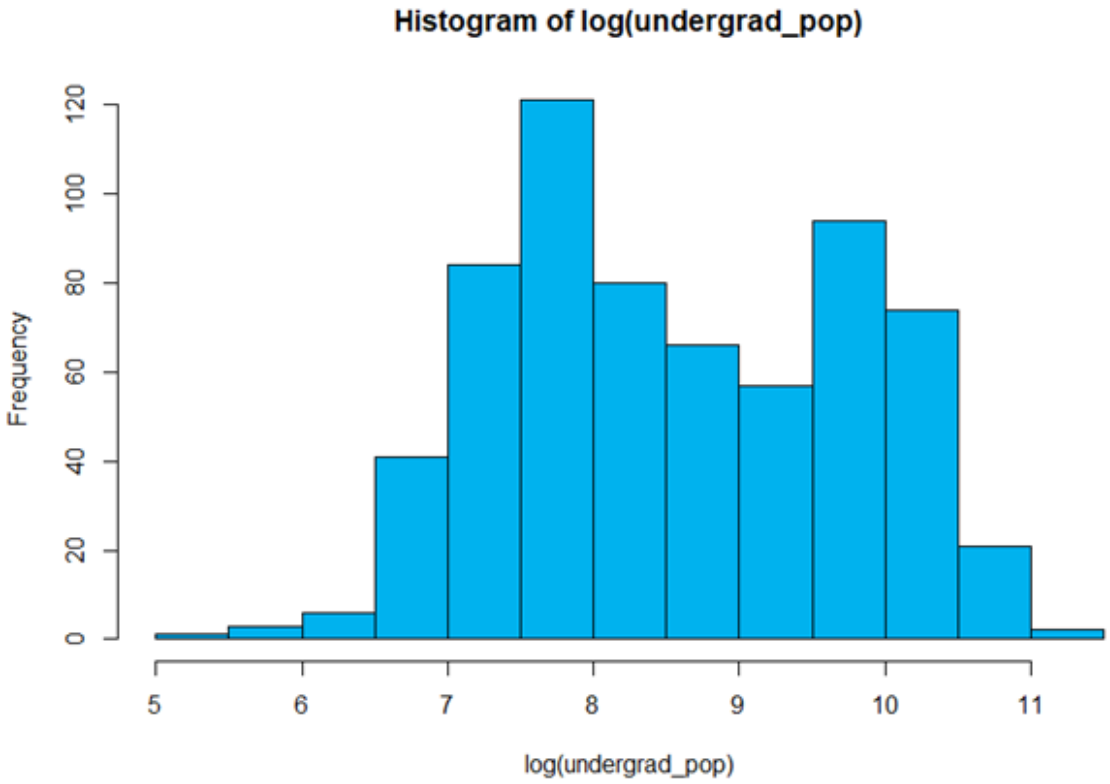

Kernel density of Alumni Salary
Density
0.0e+00
1.0e-05
2.0e-05
3.0e-05
60000
80000
100000
120000
140000
160000
Income



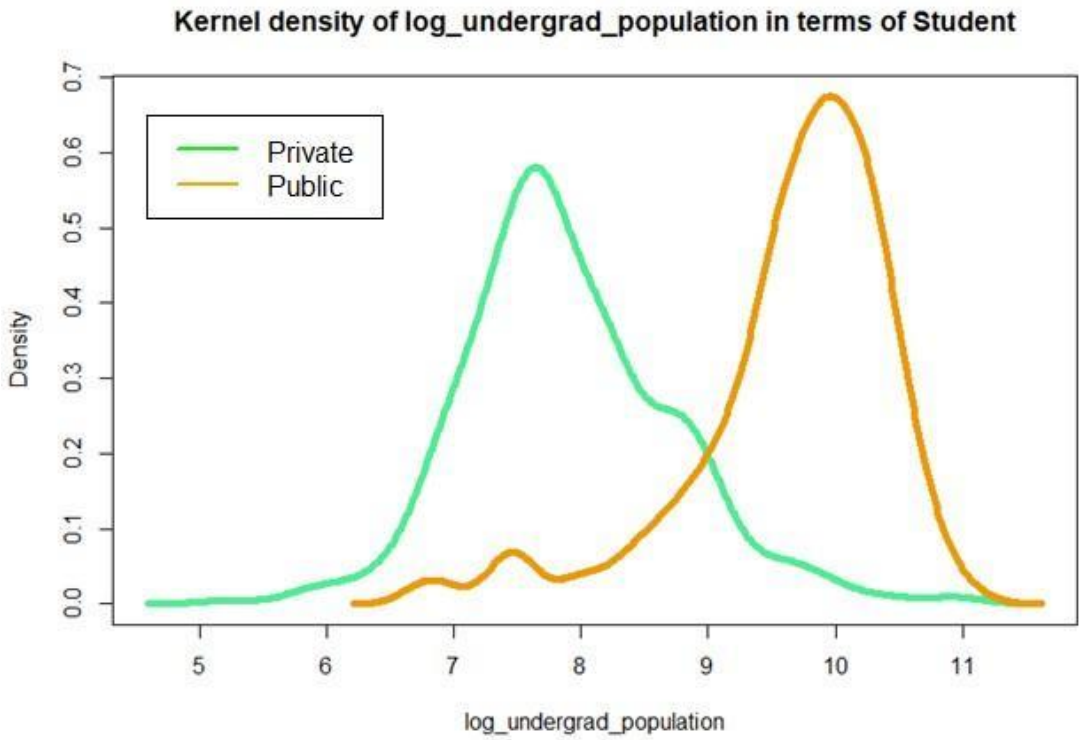


- Variables that seem to follow an exponential distribution. It is then suitable to perform a logarithmic transformation because many times this results in a Gaussian distribution. These are *Undergraduate Population* and *Student Population*.

We now show one of these plots, accompanied by its logarithmic transformation.



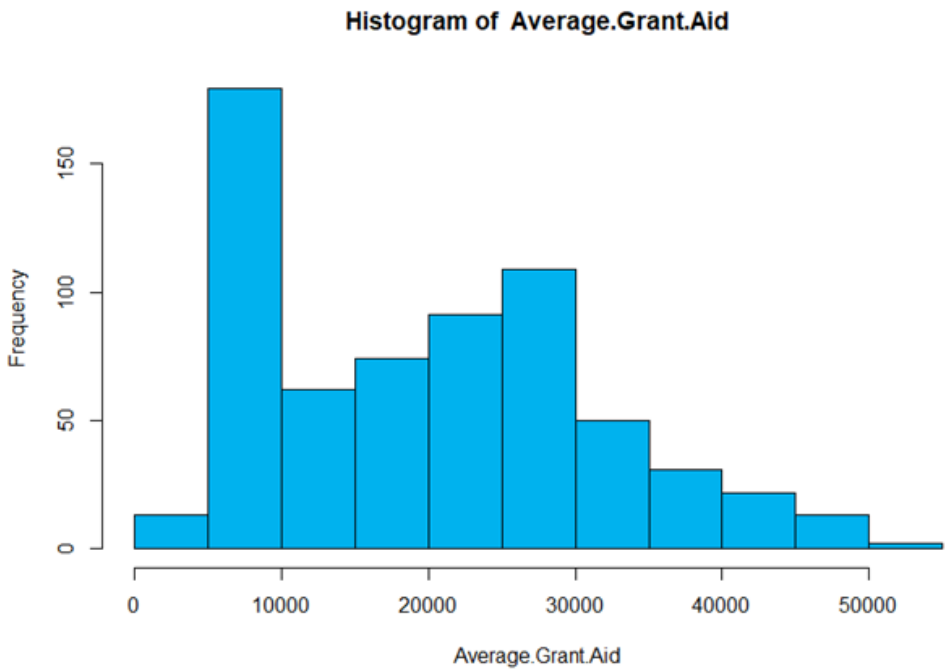
It is worth noting that with the transformation applied, there seems to be two distinguished groups under this distribution. The same thing happened with *Student Population* and its logarithm. If we plot the kernel density of this distribution distinguishing between public and private universities, we get the following result:



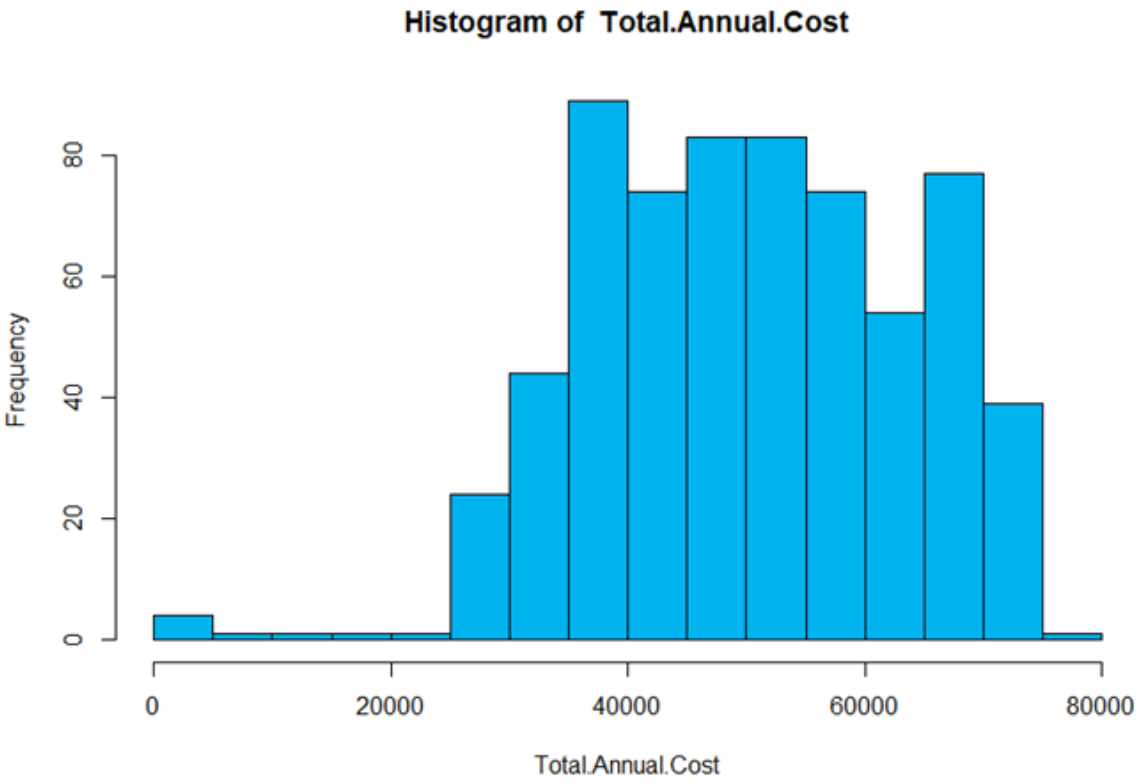
Therefore, we can affirm that these distributions seem significantly different from each other. This means that the **logarithm of the undergraduate population** in a university is a good indicator of whether the university is public or private (low values would indicate private, and high values public university).

We will take note of this and keep it in mind in the rest of the study.

- Variables that present two apparently distinguished groups: *Average Grant Aid*, *Total Annual Cost*. These are the corresponding plots:



In the first case, we see a very accumulated proportion of observations between the 5000 and 10000 indexes. We could presume that this belongs to the public universities, therefore making this variable useful for **distinguishing** between both groups; we will test this later.



In the second case, there is a very large tail below the 25000 position. On the other side of the distribution, we don't see two clear modes, as one could expect given that private universities tend to be much more expensive than public ones.

This could be the case because there are also expensive public universities, and not-so-expensive private universities, resulting in a somewhat uniform distribution around the 35000 to 70000 total annual cost.

- Give answer to some presumed observations about the differences between private and public universities, such as:

    a) Are private schools generally ranked higher?

    b) Are private schools generally more expensive?

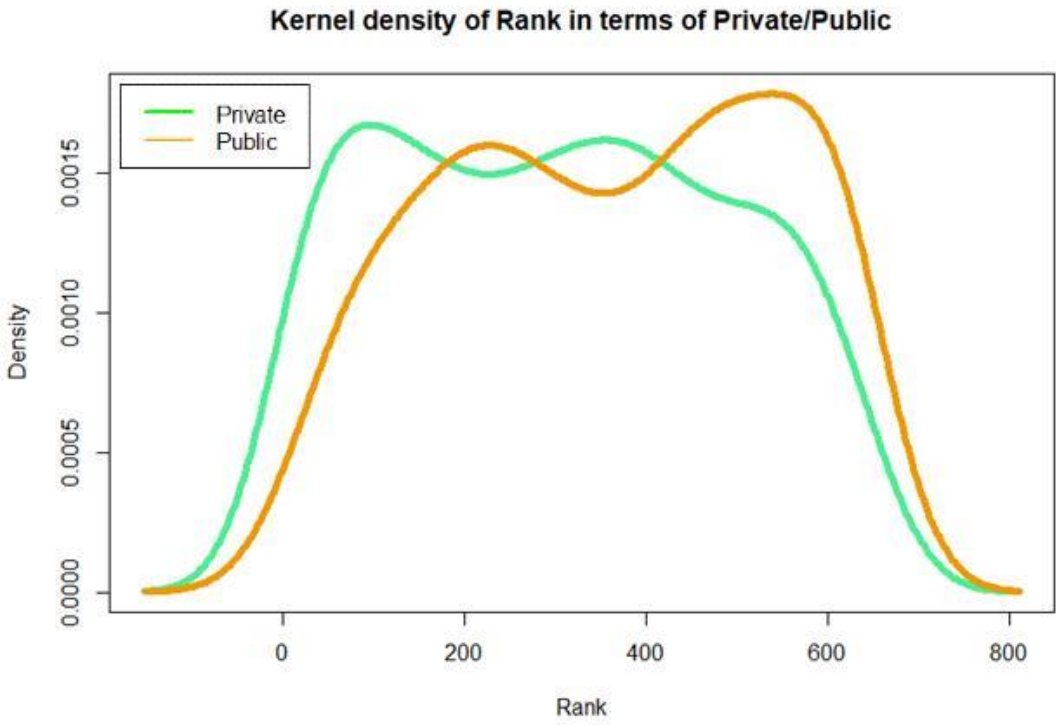
    c) Do private schools tend to have lower acceptance rates?

    d) Did their students get higher scores on the tests that got them there (SAT/ACT)?

To give an intuitive answer to these questions, we will show the kernel distributions of these variables differentiating between private and public schools (green is private,

orange is public). We will try to see whether their distributions are significantly different or not.

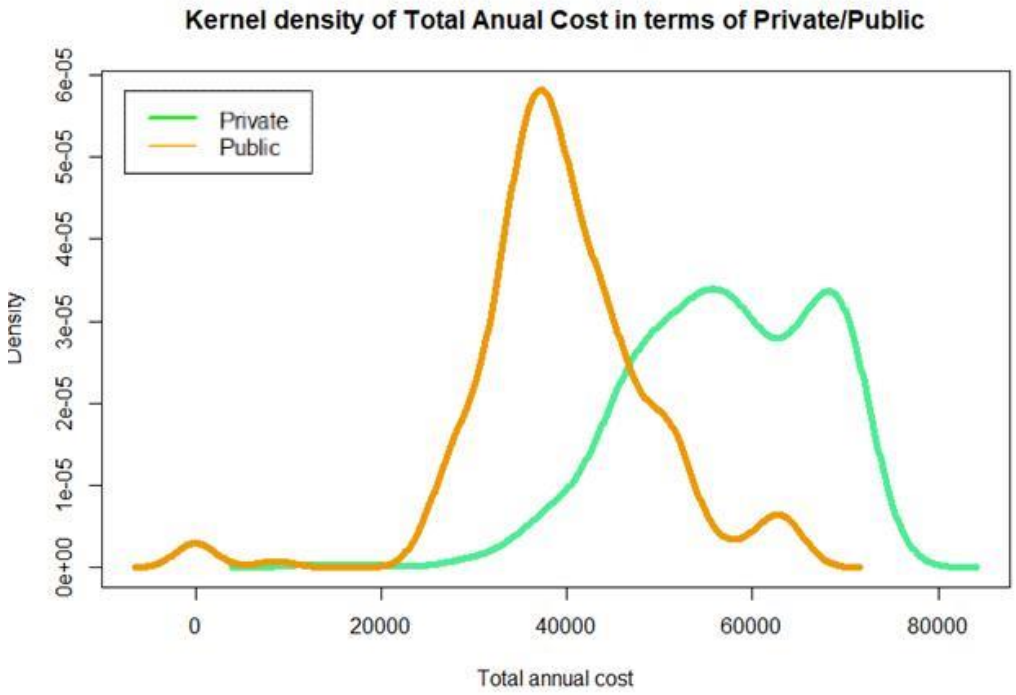
a) Private schools are ranked higher



We see that in fact, there are generally more private universities in the higher ranks, and more public ones in the lowest ranks. However, for universities ranked among the average, there doesn't seem to be a lot of differences.

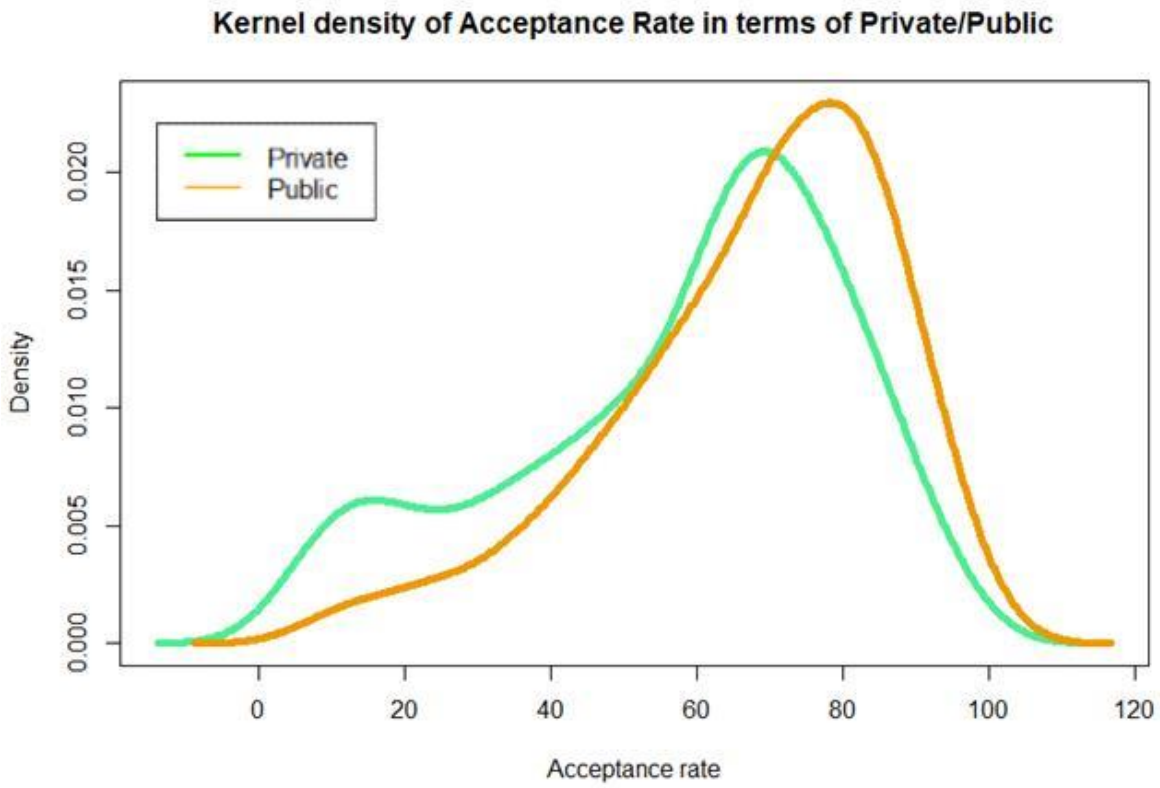
Something else worth noting is that the proportion of public schools in the lower ranks of universities is higher than the proportion of private universities in the higher ranks.

b) Private schools are more expensive



This plot clearly indicates that private schools are more expensive than public ones, and seems like a useful variable to **distinguish** between the two groups.

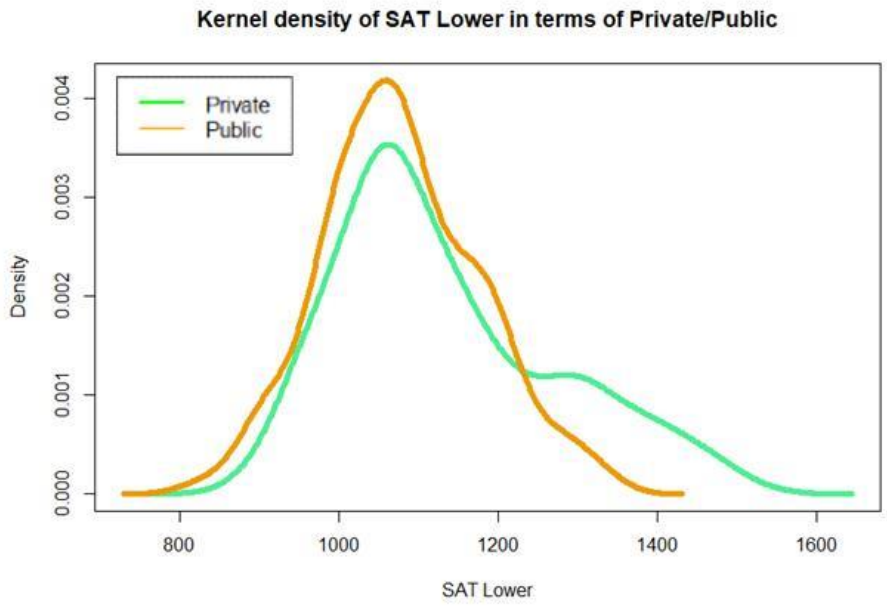
c) It's harder to get into private schools

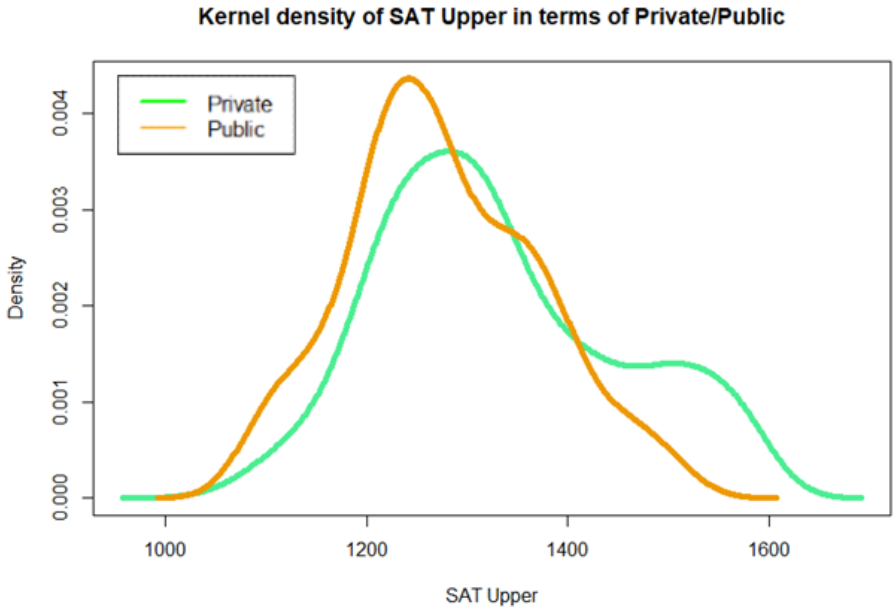


This plot shows that it is in fact generally harder to get into private than public schools. However, this also shows how the proportion of public and private schools with an average acceptance rate (50-60%) is similar.

We can also see that there seems to be a group of private schools that have a low acceptance rate, around 5-20%.

d) Students at private schools did better on the exams that got them there

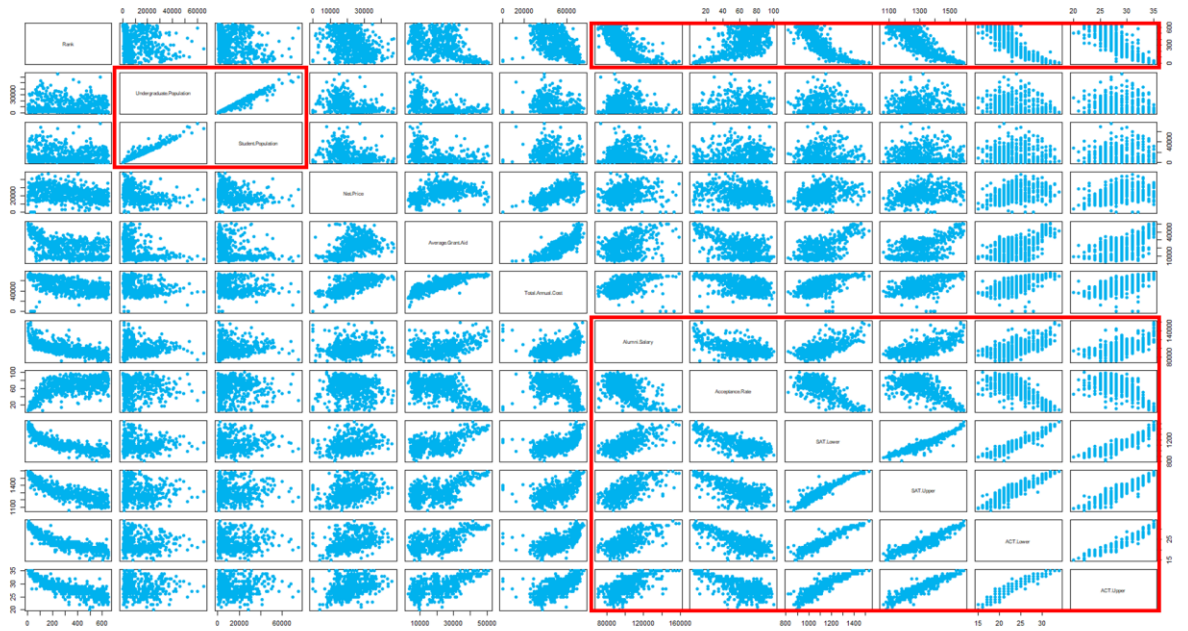




In both plots, we see that the higher scores are more abundant in the private schools, as it was thought of.

We have successfully confirmed through the kernel density plots the assumptions we made about the different distributions. However, this does not indicate that these variables are the most informative to distinguish between the groups, at least not all of them. The only thing we can see with certain confidence that can distinguish between groups is the **total annual cost**.

We will now take a look at the scatter plot matrix with the whole data and try to find interesting relationships between the variables. The remarkable observations are shown in red.

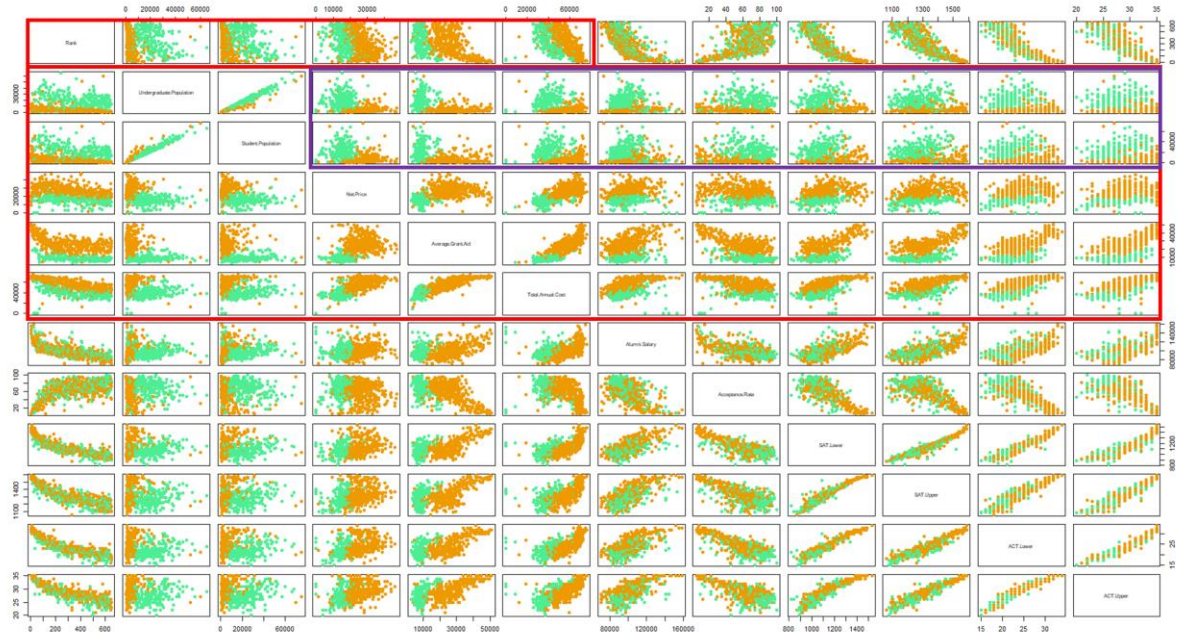


If we zoom in, we can see a perfect linear relationship between the number of undergraduates and the total number of students (top left red box). This means that the proportion of total and new students is the same across all universities; which makes sense, considering that generally speaking, people tend to stay the same number of years in university among different universities.

Additionally, we can remark on a non-linear relationship between the rank and the last six variables, which are the *Alumni Salary*, *Acceptance Rate* and four other variables regarding the scores that students got in the SAT and ACT tests (top right red box). We can highlight here, as can be expected, that better ranked colleges present higher scores in the entrance exams, the salary of their students seems to be higher and those colleges have a lower acceptance rate.

Finally, it is worth noting the bottom right red box, which encapsulates other linear relationships. These relationships occur between the variables *Alumni Salary*, *Acceptance Rate*, and the four 'scoring' variables. The relationships between the first two variables and the rest are not as strong as between the other four, maybe due to the high variability that *Alumni Salary* and *Acceptance Rate* present. On the other hand, it makes perfect sense that the scoring variables would be highly correlated, as it indicates that colleges that tend to have a lower average score in these exams for the 25% worst results will also tend to have a lower average score for the 25% best results. Similarly, for universities with better scores.

Now, by observing the scatter plot matrix separating public from private schools (orange is private, green is public), we can see that in some cases (remarked in red), just with two variables, we are able to clearly identify the two groups.



We can firstly notice something related to what we previously stated, that the **rank** and the **undergraduate population** are variables that distinguish between private and public schools in a somewhat neat way.

We also discover other pairs of variables that, when put in relation to each other, are able to also differentiate between these groups. These pairs consist of *Student population* and *Net price* with every other variable.

Finally, we can also remark that variables that showed signs of being able to discriminate between private and public schools (such as *Average grant aid* or *Total annual cost*), when put together in relation to every other variable, do a good job at discriminating, as it is reflected by the red box.

Additionally, the purple box remark instances where when we compared the two variables without distinguishing between groups, there was no apparent correlation between them. However, as we differentiate between the groups, we can now see some apparent linear correlation.

For instance, looking at the first purple box, we can see that with both variables (*Undergraduate Population* and *Student Population*) there is now an evident linear relationship (almost constant) between this variable and the others, among the **private** universities.

We will see if these observations hold when computing the corresponding correlation matrices.

Let's try to make sense of these observations now. Why do certain relationships appear among the private universities, but not among the public ones?

In the purple box, we can see that this is the case because the variance of *Undergraduate Population* and *Student Population* is very small in the population of private universities. This makes sense, as one could expect that private universities tend to have a restricted maximum number of admissions that they allow for.

2. Missing data imputation

Before going on with the estimations of the main characteristics, we are going to check if there are any missing values in the dataset.

As we can see in the following summary of the data, 8 variables are presenting missing values, varying between 2 and 99 missing values.

```
      Rank           Name               City              State          Public.Private
 Min.   :  1.0   Length:650         Length:650         Length:650         Length:650
 1st Qu.:163.2   Class :character   Class :character   Class :character   Class :character
 Median :325.5   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :325.5
 3rd Qu.:487.8
 Max.   :650.0

 Undergraduate.Population Student.Population   Net.Price     Average.Grant.Aid Total.Annual.Cost
 Min.   :  185            Min.   :  386       Min.   :    0   Min.   : 2975     Min.   :    0
 1st Qu.: 2020            1st Qu.: 2241       1st Qu.:16410   1st Qu.: 9288     1st Qu.:39917
 Median : 4503            Median : 6269       Median :21989   Median :19605     Median :50265
 Mean   :10003            Mean   :12022       Mean   :22337   Mean   :20031     Mean   :50330
 3rd Qu.:15657            3rd Qu.:17788       3rd Qu.:27581   3rd Qu.:27475     3rd Qu.:60772
 Max.   :65100            Max.   :75044       Max.   :47270   Max.   :50897     Max.   :75735
                                              NA's   :2       NA's   :4
 Alumni.Salary    Acceptance.Rate   SAT.Lower      SAT.Upper      ACT.Lower       ACT.Upper
 Min.   : 70700   Min.   :  5.0   Min.   : 820   Min.   :1060   Min.   :15.00   Min.   :20.00
 1st Qu.: 88600   1st Qu.: 48.0   1st Qu.:1020   1st Qu.:1230   1st Qu.:21.00   1st Qu.:26.00
 Median : 96400   Median : 67.0   Median :1080   Median :1290   Median :22.00   Median :28.00
 Mean   : 98852   Mean   : 61.6   Mean   :1110   Mean   :1308   Mean   :23.28   Mean   :28.44
 3rd Qu.:105600   3rd Qu.: 78.0   3rd Qu.:1180   3rd Qu.:1380   3rd Qu.:25.00   3rd Qu.:31.00
 Max.   :158200   Max.   :100.0   Max.   :1530   Max.   :1590   Max.   :34.00   Max.   :35.00
 NA's   :15       NA's   :2       NA's   :99     NA's   :99     NA's   :97      NA's   :97
   Website
 Length:650
 Class :character
 Mode  :character
```

And as we can see in the next table, there are 14 observations with one missing value, 26 observations with 2 missing values, 1 observation with 3 missing values, 79 observations with 4 missing values and 6 observations with 5 missing values.

```
  0   1   2   3   4   5
524  14  26   1  79   6
```

We will use the Predictive Mean Matching (PMM) method to find appropriate values to replace the data that is missing.

For each missing entry, this method forms a small set of candidate donors (typically with 3, 5 or 10 members) from all complete cases that have predicted values closest to the predicted

value for the missing entry. One donor is randomly drawn from the candidates, and the observed value of the donor is taken to replace the missing value.

In the next two screenshots, we can see an example of the imputation of missing values for the 6 observations with 5 missing values. In the first picture, we can see the missing values, whereas in the second picture, those values have been replaced.

	Alumni.Salary	Acceptance.Rate	SAT.Lower	SAT.Upper	ACT.Lower	ACT.Upper
163	NA	43	NA	NA	NA	NA
551	94600	NA	NA	NA	NA	NA
570	NA	68	NA	NA	NA	NA
604	94000	NA	NA	NA	NA	NA
637	NA	97	NA	NA	NA	NA
649	NA	80	NA	NA	NA	NA

	Alumni.Salary	Acceptance.Rate	SAT.Lower	SAT.Upper	ACT.Lower	ACT.Upper
163	118500	43	1210	1330	27	30
551	94600	62	950	1190	19	26
570	83900	68	1025	1190	20	26
604	94000	80	1070	1270	20	23
637	98700	97	1060	1270	22	26
649	87400	80	1030	1190	21	25

3. Estimating the main characteristics of quantitative variables

- **MEAN VECTORS:**

In the next table, we can see the mean vectors of the quantitative variables for the whole dataset, and also separated for the group of Public colleges and the group of Private colleges.

As we can see, there are some variables that show a clear difference between the mean for Public colleges and for Private colleges. That is the case of the first four variables, *Undergraduate.Population, Student.Population, Net.Price* and *Average.Grant.Aid.* For example, the mean of the variable Undergraduate.Population is nearly 5 times higher for the Public colleges than for the Private ones. It indicates that as we suggested before, these variables present clearly differentiated values for each group, and therefore they can be useful to distinguish between the two groups.

	Mean Vector	Mean Vector (Public colleges)	Mean Vector (Private colleges)
Undergraduate.Population	10002.69	19616.45	3994.10
Student.Population	12022.29	22373.42	5552.84
Net.Price	22315.74	15207.44	26758.42
Average.Grant.Aid	19935.26	8957.70	26796.23
Total.Annual.Cost	50330.18	39526.25	57082.63
Alumni.Salary	98714.62	98050	99130
Acceptance.Rate	61.63	67.50	57.96
SAT.Lower	1108.61	1073.55	1130.52
SAT.Upper	1306.06	1272.56	1327
ACT.Lower	23.27	22.00	24.06
ACT.Upper	28.43	27.31	29.13

- **COVARIANCE MATRIX:**

In the following screenshots, we can see the covariance matrix of the quantitative variables. It can be useful to identify how the trends of two variables are related, and also the variability of a variable with respect to its mean.

	Rank	Undergraduate.Population	Student.Population	Net.Price
Rank	35262.5000	-127565.522	-3.296654e+05	-511029.177
Undergraduate.Population	-127565.5223	124594093.347	1.437543e+08	-38659793.347
Student.Population	-329665.3690	143754345.273	1.735830e+08	-39618060.334
Net.Price	-511029.1772	-38659793.347	-3.961806e+07	68383130.844
Average.Grant.Aid	-897907.7388	-65832519.855	-6.600246e+07	47348744.546
Total.Annual.Cost	-1371519.1926	-51346147.143	-4.663545e+07	78565542.436
Alumni.Salary	-1936516.8721	6357373.379	2.139231e+07	22759207.372
Acceptance.Rate	2594.5092	12952.552	-6.145313e+03	-23720.836
SAT.Lower	-18652.8459	-44489.238	9.037646e+04	318463.948
SAT.Upper	-16498.2797	-53680.347	6.911347e+04	295976.674
ACT.Lower	-560.4715	-3109.807	6.197772e+02	10828.238
ACT.Upper	-452.5978	-2359.368	6.919732e+02	9070.165

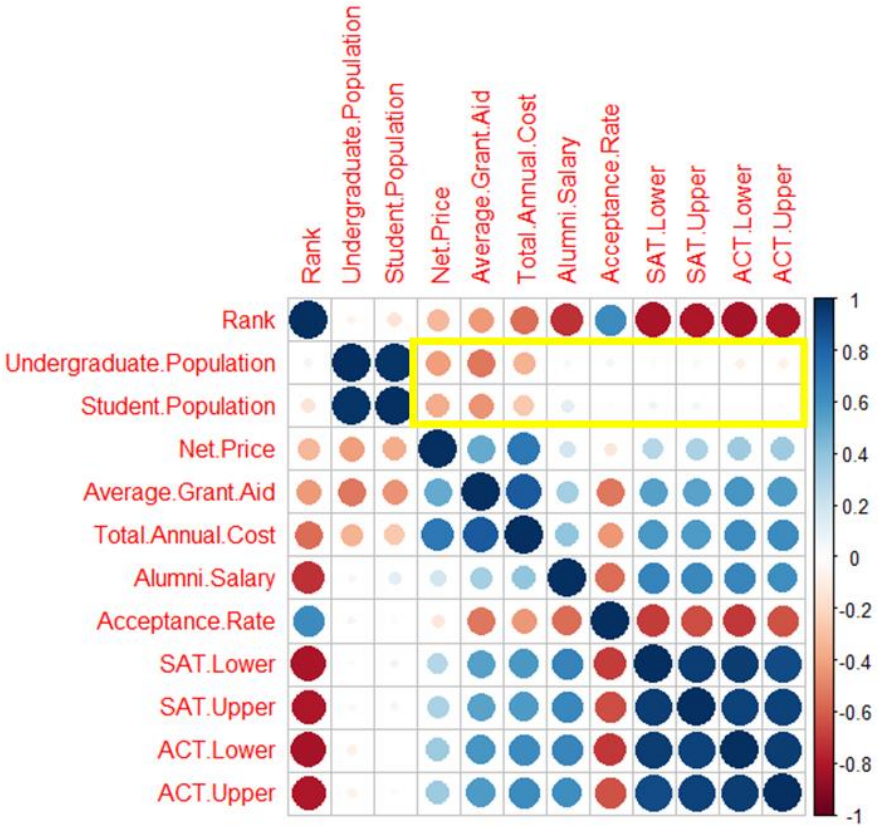
	Average.Grant.Aid	Total.Annual.Cost	Alumni.Salary	Acceptance.Rate	SAT.Lower
Rank	-897907.74	-1371519.19	-1936516.87	2594.50924	-18652.8459
Undergraduate.Population	-65832519.85	-51346147.14	6357373.38	12952.55197	-44489.2375
Student.Population	-66002464.28	-46635447.80	21392306.53	-6145.31332	90376.4589
Net.Price	47348744.55	78565542.44	22759207.37	-23720.83598	318463.9485
Average.Grant.Aid	125623401.92	124562249.57	53117377.29	-129571.97006	751600.0281
Total.Annual.Cost	124562249.57	174849184.53	74758394.35	-127263.88040	948648.8298
Alumni.Salary	53117377.29	74758394.35	204575641.22	-175580.96835	1159869.3564
Acceptance.Rate	-129571.97	-127263.88	-175580.97	484.11734	-1856.7472
SAT.Lower	751600.03	948648.83	1159869.36	-1856.74724	14778.8178
SAT.Upper	667167.06	842015.68	997589.52	-1581.65743	12514.7288
ACT.Lower	23246.27	29594.55	33605.94	-56.25108	407.6718
ACT.Upper	18977.85	24685.75	26122.05	-41.50435	320.8296

	SAT.Upper	ACT.Lower	ACT.Upper
Rank	-16498.2797	-560.471495	-452.597843
Undergraduate.Population	-53680.3473	-3109.806566	-2359.368022
Student.Population	69113.4732	619.777205	691.973166
Net.Price	295976.6745	10828.238070	9070.165225
Average.Grant.Aid	667167.0562	23246.273182	18977.847552
Total.Annual.Cost	842015.6792	29594.550821	24685.753301
Alumni.Salary	997589.5235	33605.942871	26122.045751
Acceptance.Rate	-1581.6574	-56.251080	-41.504350
SAT.Lower	12514.7288	407.671793	320.829608
SAT.Upper	11855.1563	356.084784	296.176176
ACT.Lower	356.0848	12.710973	9.929193
ACT.Upper	296.1762	9.929193	8.692426

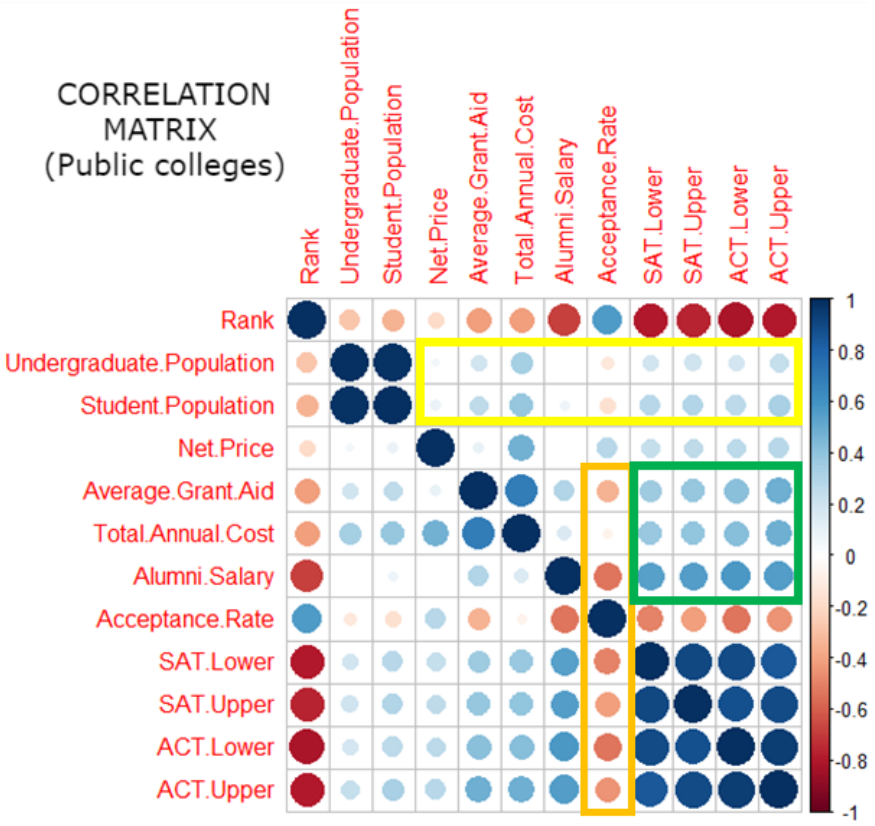
Those pairs of variables that present a high positive value, as it is the case of *Total.Annual.Cost* and *Net.Price*, have a positive relation, and when one of them increases the other one also increases. On the other hand, when two variables present a high negative value, they have a negative relation, and when one of them increases the other one decreases. That is the case, for example, of the variables *Alumni.Salary* and *Rank.*

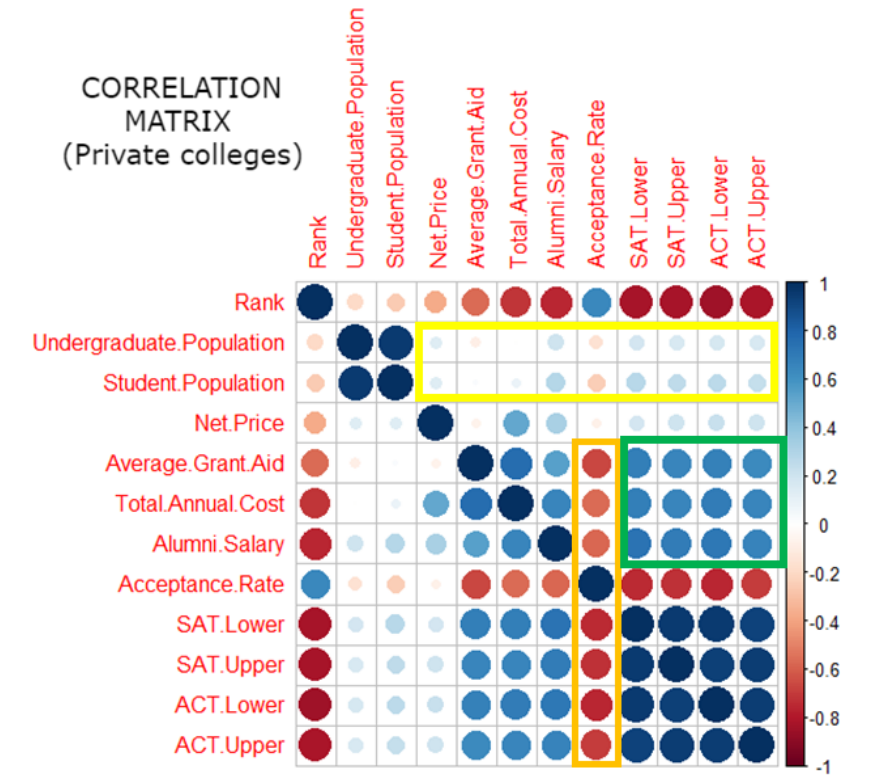
However, with the covariance matrix we cannot compare values and say that two variables have a stronger relation than other two variables; and also, while analyzing the whole population, it may seem that there is a certain kind of relationship between two variables, but if we analyze Private and Public universities separately, it may occur that for each group they present different relations. That is why we will obtain the correlation matrix for each of the groups in the next step.

- **CORRELATION MATRICES:**



Looking at the marked variables, we see that as we suspected, considering the whole population (not distinguishing between groups) the linear correlation between *Undergraduate Population*, *Student Population* and the last 6 variables is almost non-existent.
However, the result for the other correlations seems to be larger than what the plots suggested.
We will now compare these matrices among the private and public universities.





As we can see looking at the yellow box, the magnitudes of these correlations have changed when compared to the previous matrix. In particular, the correlations that were worth mentioning have reduced its value significantly. On the other hand, the correlations between

the population variables and the scoring variables are now a bit more significant than in the previous case.
This shows how one cannot deduce that two variables are linearly related or not from the observations of the whole data. Instead, we could only say so about the entire population; one needs to separate among groups and then make the appropriate conclusions.

We can also remark on two groups of variables for which their linear correlation is much more significant in the private universities population than in the public ones.
For the first group (highlighted in orange), it corresponds to the correlation between the *Acceptance Rate* and the **financial** and **scoring** variables (negative correlation). This means that for the private universities, higher acceptance rates equates to lower financial aid and cost and also to lower scoring on the entry exams.

As for the second group (remarked in green), it regards the correlation between the **financial** and **scoring** variables. The interpretation is in the sense that higher scores on the entrance exams will equate to higher financial aid and that cost is stronger/more present among private universities.

4. Outliers and other characteristics of interest

Using the Minimum covariance determinant (MCD) estimators, we can find robust estimates of the mean, covariance matrix and correlation matrix. With these estimators, we can try to detect potential outliers present in the dataset.

But the MCD estimates should be used with caution because they are mainly appropriate for approximately symmetric data sets, and therefore, we are going to transform the highly non-Gaussian variables to something more symmetric. Logarithmic transformations will be applied to *Undergraduate.Population*, *Student.Population* and *Alumni.Salary*.

We are also going to search for outliers by subgroups because if there are two groups in the data, the number of detected potential outliers will increase as the data will be far from Gaussianity. In this case, the two groups are the public universities and the private universities.

First of all, an estimation of the sample mean is compared with the robust mean estimate (the sample mean of the non-outliers) within each subgroup.
For public universities:

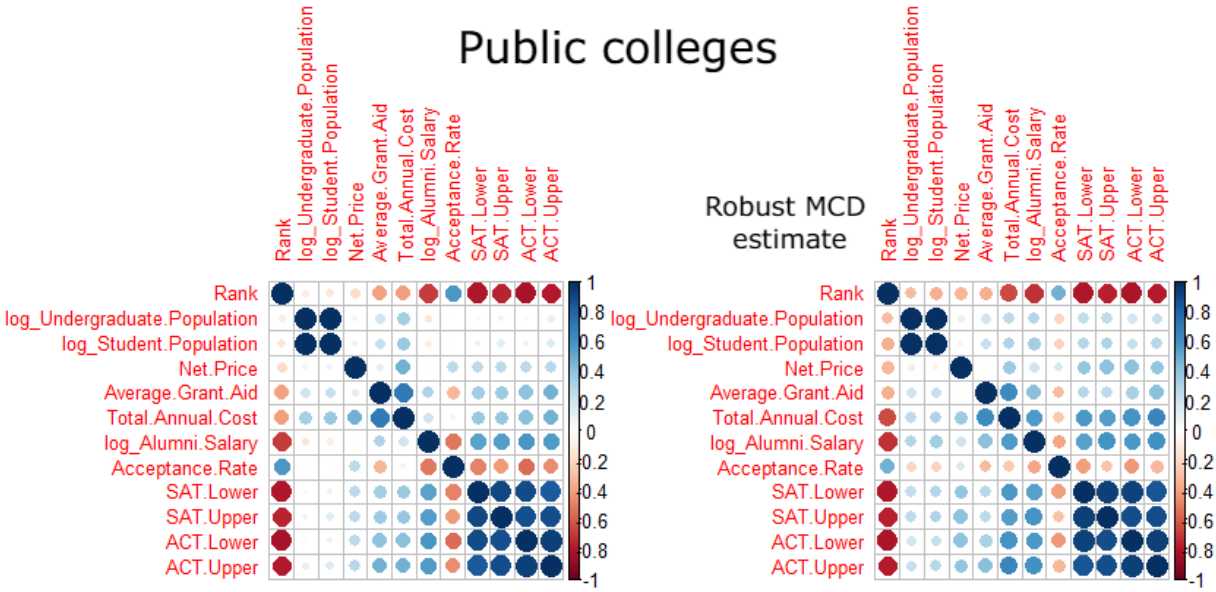
```
> m_pub
                      Rank log_Undergraduate.Population       log_Student.Population
                   360.532                        9.654                        9.771
                 Net.Price            Average.Grant.Aid           Total.Annual.Cost
                 15207.440                     8957.704                     39526.252
           log_Alumni.Salary               Acceptance.Rate                   SAT.Lower
                    11.486                       67.496                      1073.548
                 SAT.Upper                    ACT.Lower                    ACT.Upper
                  1272.564                       22.004                       27.312
> m_MCD_pub
                      Rank log_Undergraduate.Population       log_Student.Population
                   381.238                        9.801                        9.908
                 Net.Price            Average.Grant.Aid           Total.Annual.Cost
                 15476.173                     8562.126                     39497.935
           log_Alumni.Salary               Acceptance.Rate                   SAT.Lower
                    11.461                       70.215                      1064.075
                 SAT.Upper                    ACT.Lower                    ACT.Upper
                  1260.056                       21.650                       26.981
```

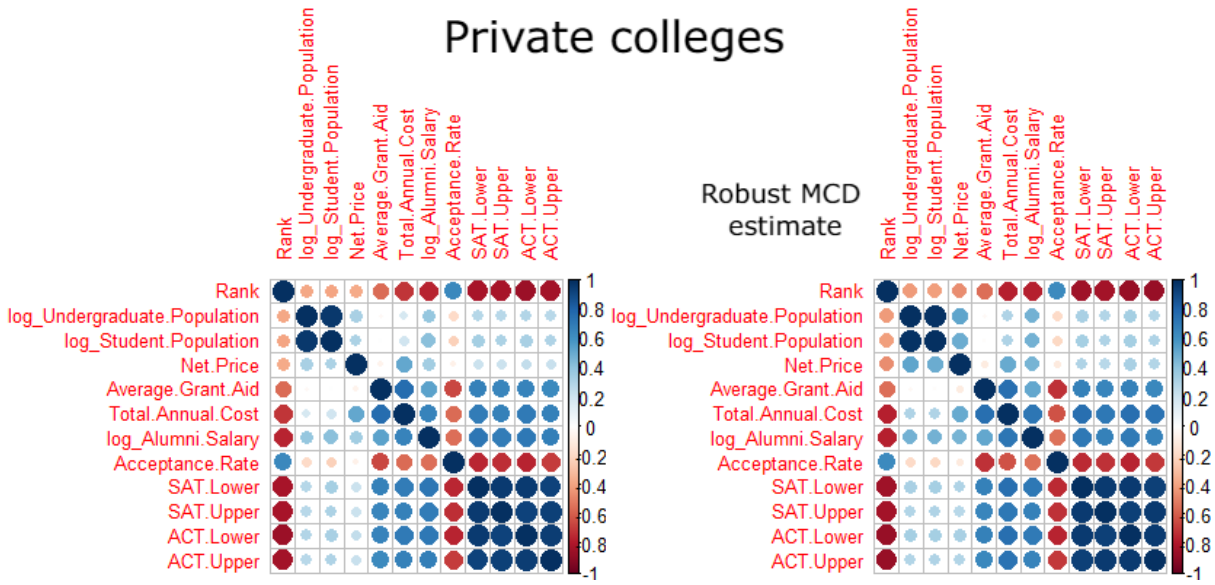
And for private universities:

```
> m_priv
                      Rank log_Undergraduate.Population       log_Student.Population
                   303.605                        7.897                        8.115
                 Net.Price            Average.Grant.Aid           Total.Annual.Cost
                 26758.425                    26796.228                     57082.628
           log_Alumni.Salary               Acceptance.Rate                   SAT.Lower
                    11.493                       57.958                      1130.523
                 SAT.Upper                    ACT.Lower                    ACT.Upper
                  1327.000                       24.058                       29.130
> m_MCD_priv
                      Rank log_Undergraduate.Population       log_Student.Population
                   304.503                        7.883                        8.068
                 Net.Price            Average.Grant.Aid           Total.Annual.Cost
                 26963.409                    27008.658                     57541.690
           log_Alumni.Salary               Acceptance.Rate                   SAT.Lower
                    11.485                       59.652                      1125.953
                 SAT.Upper                    ACT.Lower                    ACT.Upper
                  1323.316                       23.918                       29.038
```

In both groups we can appreciate some slight differences between the means of the whole data and the means of the non-outliers, but there are no important variations.

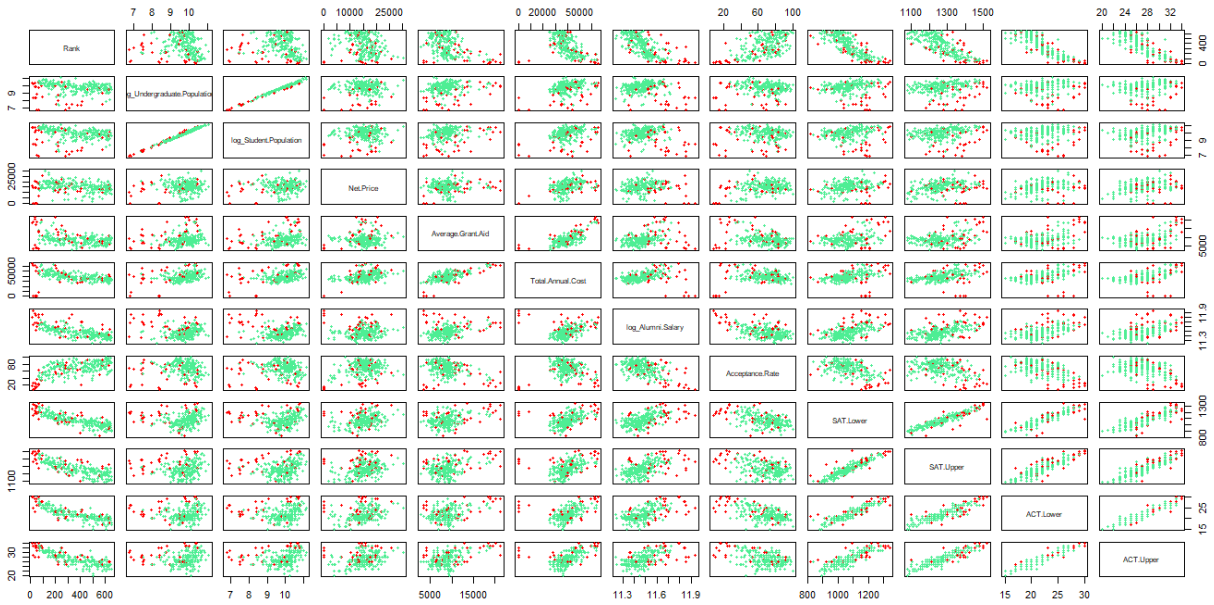
Comparing the correlation matrices in the same way, it seems that the outliers don't hide any strong correlations, just some slight correlations for public colleges involving *Total.Annual.Cost* variable and for private colleges between populations and *Net.price* as we can see in the next screenshots:



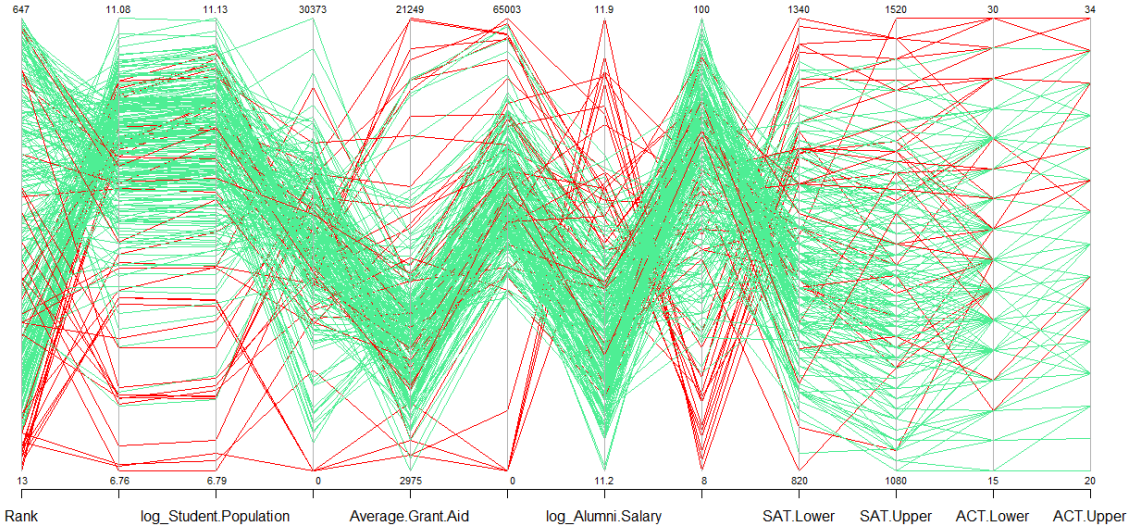


Finally, with the next scatter plot matrices, Parallel coordinates plots and Andrew's plots, we can appreciate the points that have been skipped to compute the MCD estimators (red points) and thus, the estimated outliers of each subgroup. (Zoom in to better appreciate the red points)

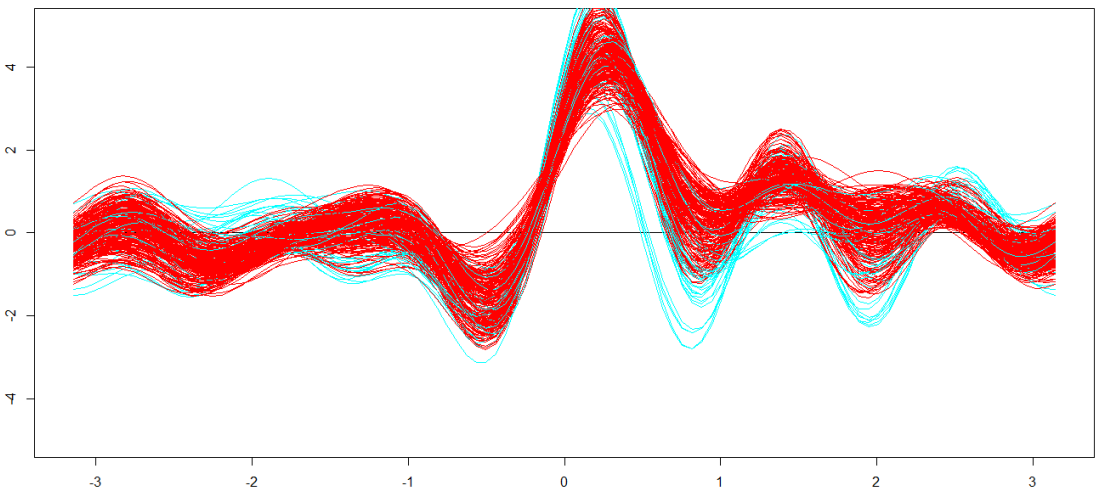
For public universities:



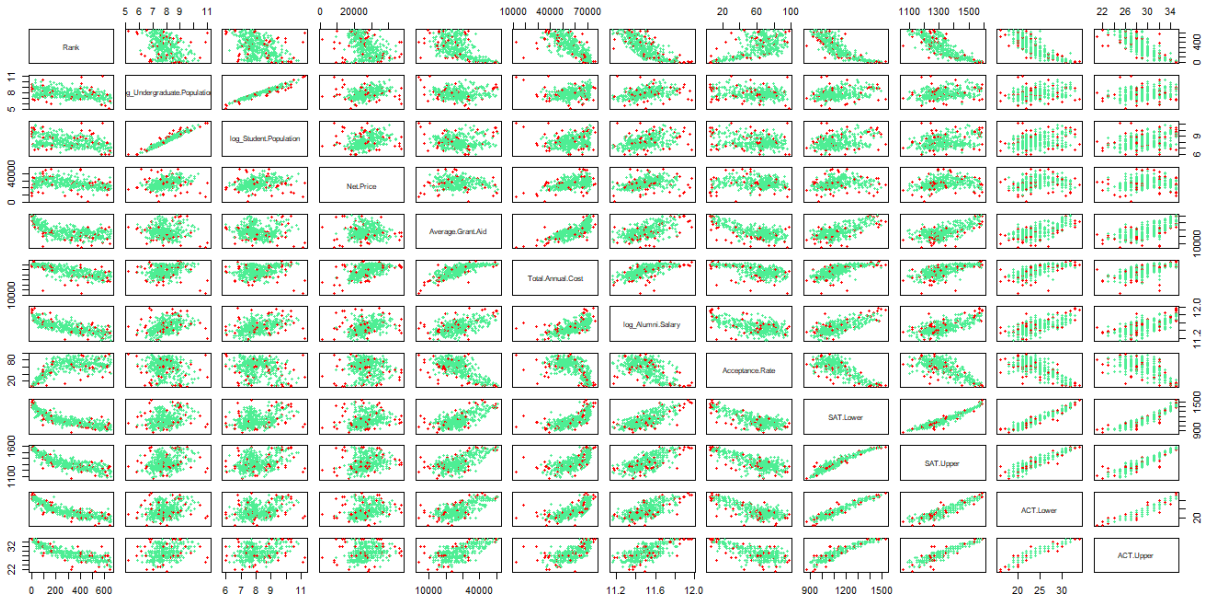
PCP (public)



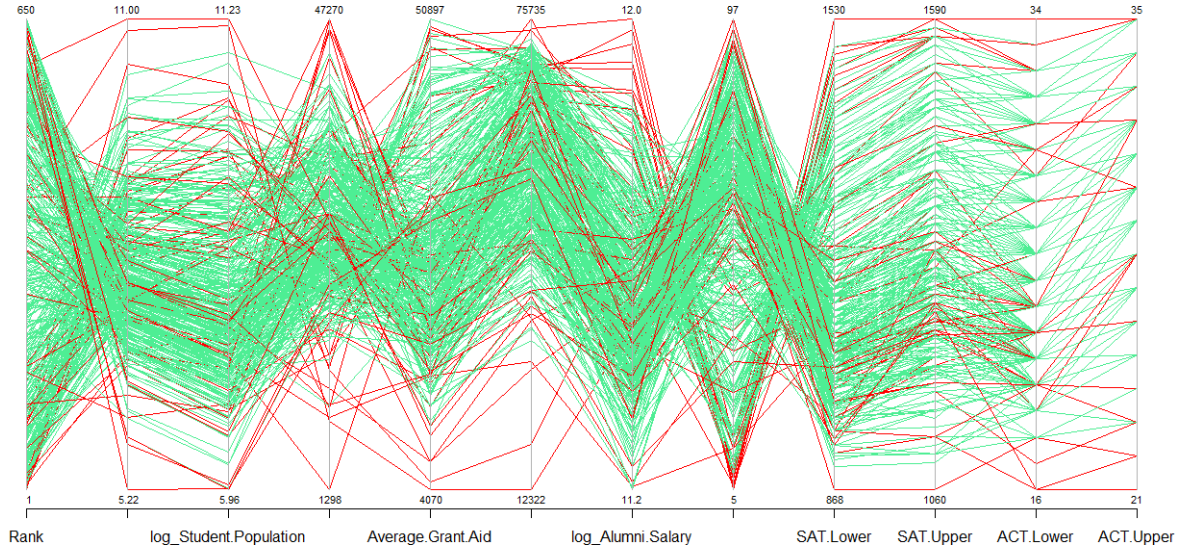
Andrews' Plot (public)



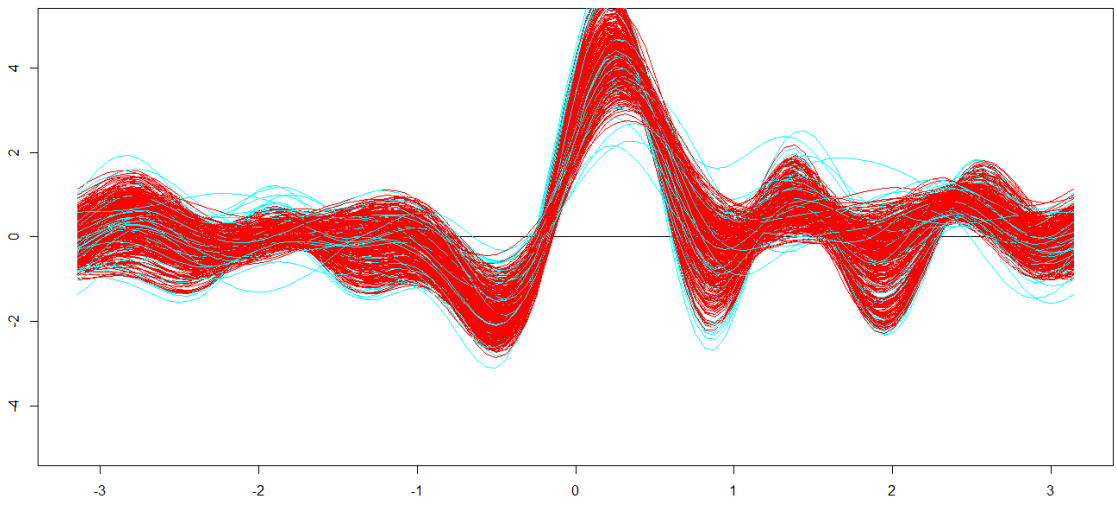
For private universities:



PCP (private)



Andrews' Plot (private)



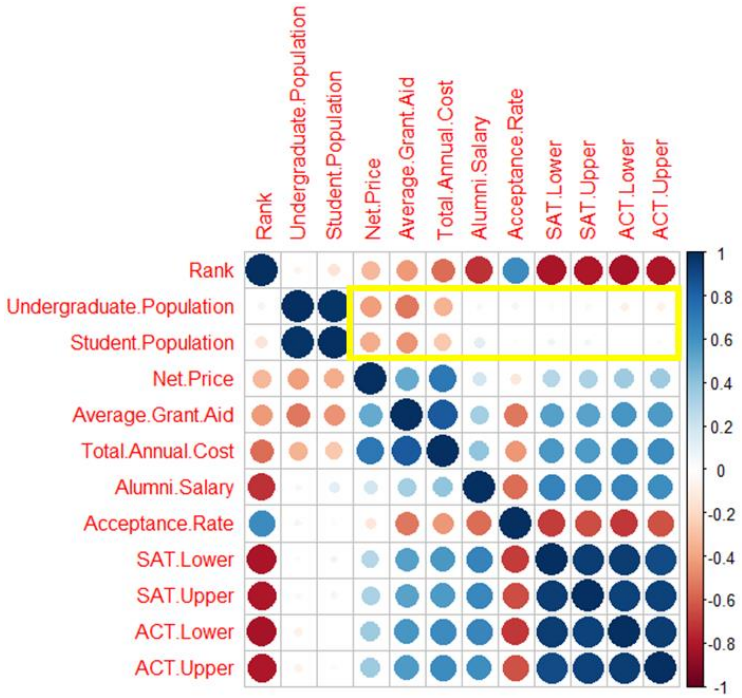
It seems that there are slightly more outliers affecting the public universities. It is also easy to appreciate that most of the observations labeled as outliers by the MCD estimators (red lines in PCP and Blue lines in Andrew's plot), behave somewhat differently from the non-outliers.

5. Dimension reduction techniques

We will now perform the two main dimension reduction techniques (PCA and ICA) with the intention of projecting the whole set of variables into a smaller set (two or three dimensions) that are able to explain most of the variability in the data and also allow us to differentiate between the private and public universities. Furthermore, we will look for outliers among these variables.

These are the following observations we will take into account prior to performing these methodologies:

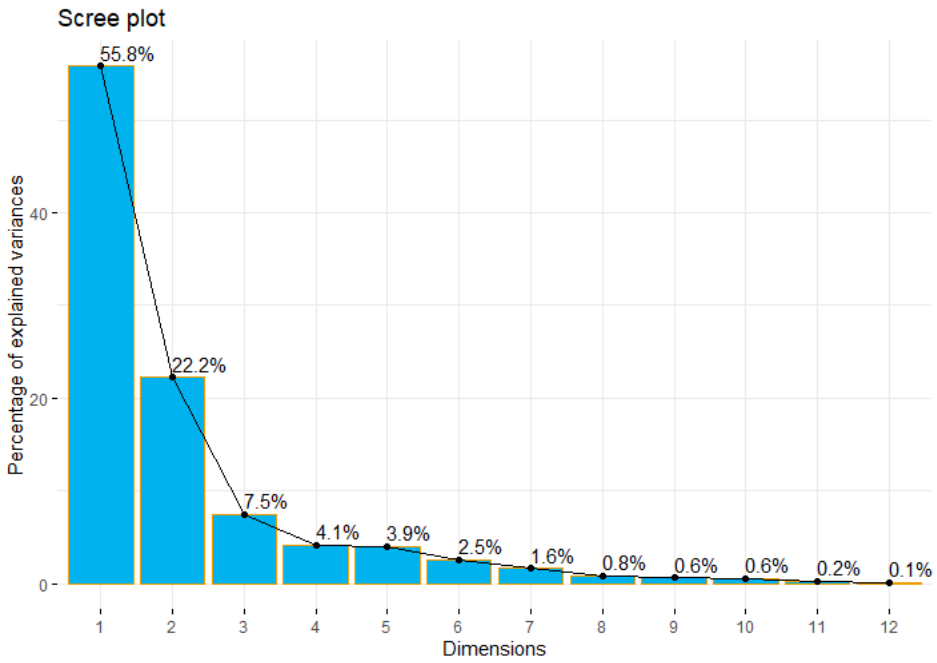
- Transformations which made variables appear to be more Gaussian-like will be applied and included. These are *log_undergrad_population* (heavily skewed), *log_student_population* (exponential) and *log_alumni_salary* (exponential).
- Certain variables seem to already distinguish pretty well between groups on their own: *log_undergrad_population, Total annual cost, Rank, Average grant aid*. Therefore, if we obtain components with some of these variables present in a significant way, we could expect that these components will also perform well in separating private and public universities.
- Study of the correlation matrix:



Looking at the correlations, we can presume that we could have a component regarding the exam scores (SAT/ACT Lower/Upper), another one representing the financial resources, and perhaps one involving the population in a university.
Additionally, as two of the variables that on their own are able to separate between groups are financial variables, this financial component could be the one that is also able to do so.

5.1 Principal Component Analysis

Let's begin the analysis by analysing a reasonable number of components to be studied, in terms of the cumulative amount of variance explained.
Below is the scree plot for the percentage of variance explained for each component, and also the corresponding table.



	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	6.69283938	55.77366152	55.77366
Dim.2	2.66546535	22.21221123	77.98587
Dim.3	0.89697561	7.47479671	85.46067
Dim.4	0.49168801	4.09740011	89.55807
Dim.5	0.47274114	3.93950951	93.49758
Dim.6	0.30463823	2.53865188	96.03623
Dim.7	0.19301701	1.60847504	97.64471
Dim.8	0.10155270	0.84627249	98.49098
Dim.9	0.07281817	0.60681812	99.09780
Dim.10	0.06874833	0.57290278	99.67070
Dim.11	0.02808534	0.23404448	99.90474
Dim.12	0.01143074	0.09525613	100.00000

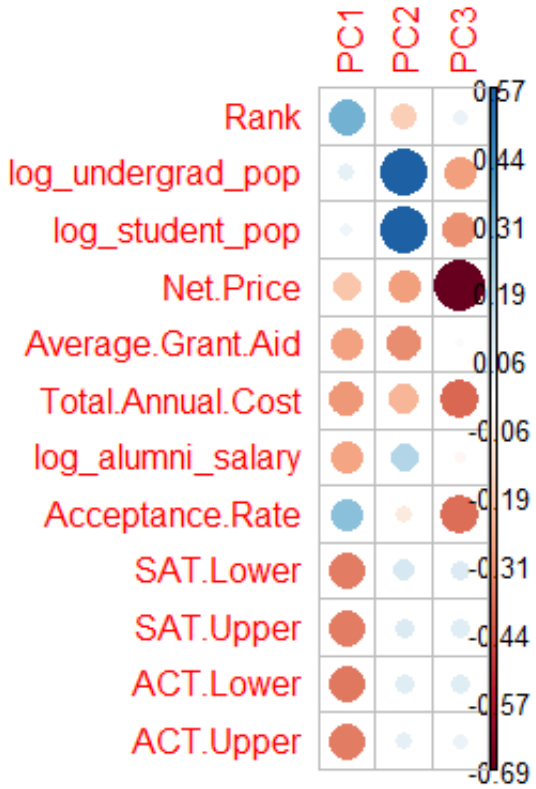
Looking at these figures, we observe that we need two PCs to have 78.03% of the total variability of the data set. With three PCs we are able to explain 85.44%. We decided to stay with three principal components.

This means we are able to still extract 85.44% of the information with only 25% of the variables.

Now, let's study what these components represent. After this, we will see how well they are able to distinguish between groups and identify outliers.
The following is the **loading matrix** corresponding to the first three PCs:

	PC1	PC2	PC3
Rank	0.33207632	-0.16975610	0.05827317
log_undergrad_pop	0.07032664	0.56641437	-0.28676603
log_student_pop	0.04593191	0.56473375	-0.31182421
Net.Price	-0.19021446	-0.28956436	-0.69184979
Average.Grant.Aid	-0.28145411	-0.32217205	0.01543664
Total.Annual.Cost	-0.29867254	-0.23159457	-0.39436509
log_alumni_salary	-0.27628305	0.20628117	-0.02925118
Acceptance.Rate	0.28680920	-0.08160911	-0.38697746
SAT.Lower	-0.35904872	0.12190298	0.10606128
SAT.Upper	-0.35671625	0.10638086	0.09429402
ACT.Lower	-0.36611046	0.09222225	0.09023999
ACT.Upper	-0.35954877	0.07912584	0.06185543

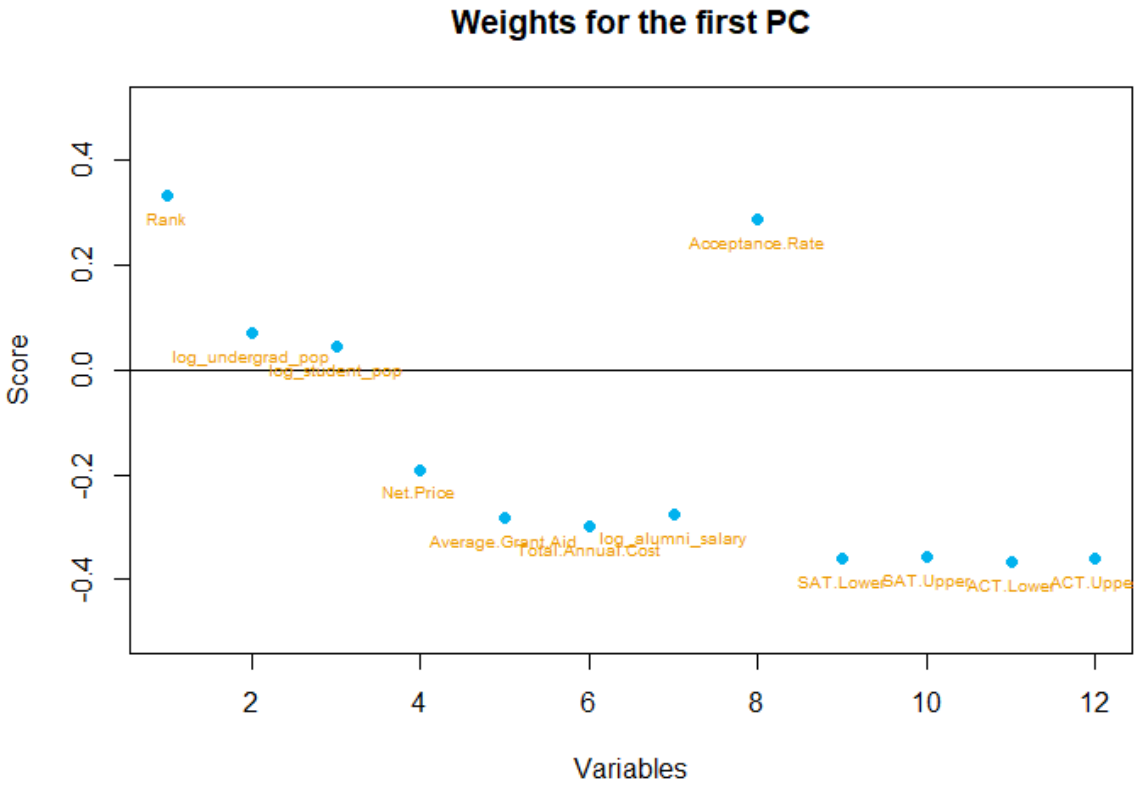
To get a more visual interpretation of this table, we plot the following figure (with the parameter *is.corr=FALSE*):

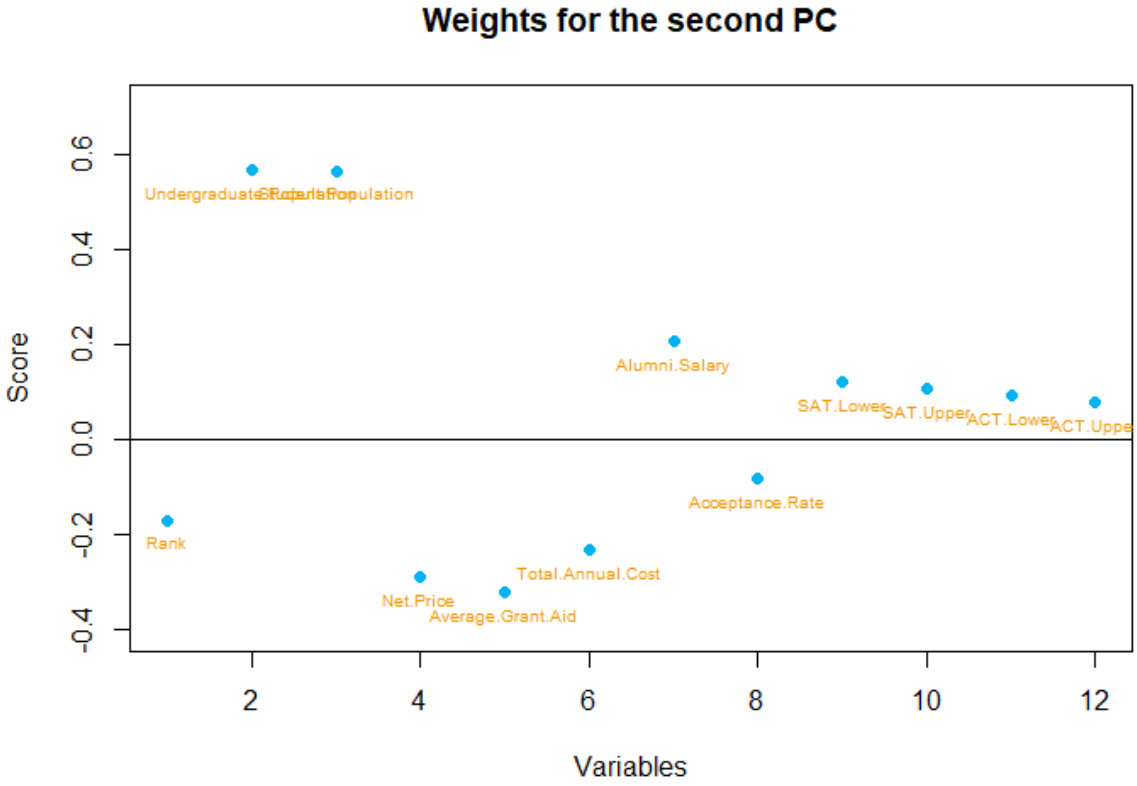


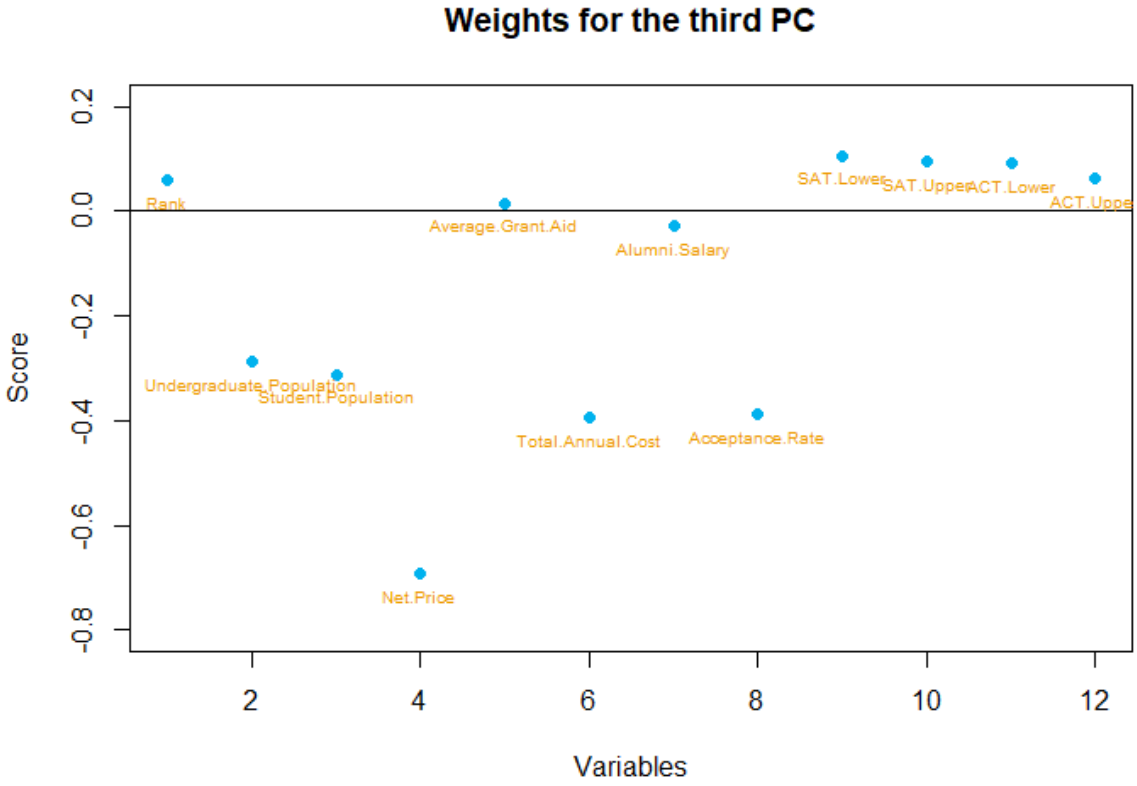
First, the second and third components have a somewhat clear interpretation. The second one has the highest loading values in the two variables regarding the **university population**. As we previously saw, private universities tend to have fewer students. Therefore, this principal component could be useful in order to distinguish between private and public universities.

As for the third PC, it has a very large value in *Net Price* and mild ones in *Total annual cost* and *Acceptance Rate*. We can infer this component refers to the general financial cost of attending each university.
Finally, the first component does not have a clear interpretation. This is often the case in PCA as the first component is always the one that explains the most variability. That is why it often has middle-range values across multiple variables, unless there are a few variables that explain most of the variability (which is not the case in this data).
However, we can observe that this component is coherent with our previous analysis. This is for two reasons: first, all of the exam scoring variables are put in the same component, with similar weights. Secondly, we see that the weights for these variables are the opposite as the *Rank* and *Acceptance rate* variables. This means that this component is able to capture the fact that for higher scores in the entrance exams, these universities will tend to have a lower (better) rank and also lower acceptance rates, which again, is something we mentioned previously.
Considering all of this, we can infer that this component reflects whether the university is good or not, regardless of it being private or public. This is the case because the most prevalent variables in this component are the rank and the scoring exam variables, which indicate how requested these universities are.

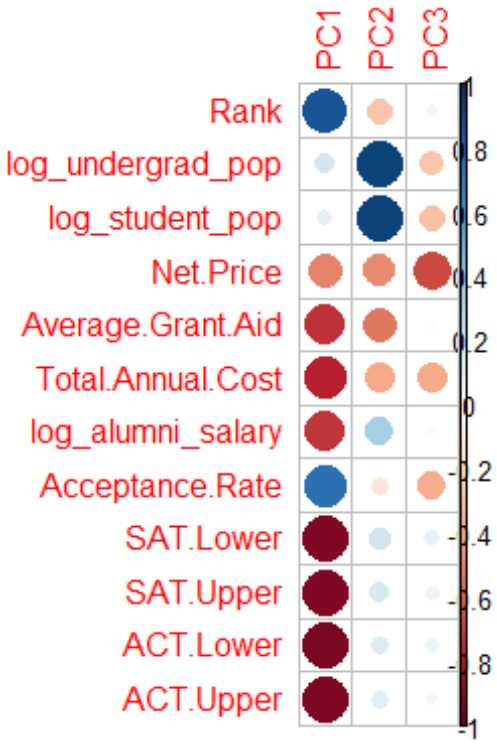
We also plotted the weights in the following way, which interpretation is similar to the previous one:







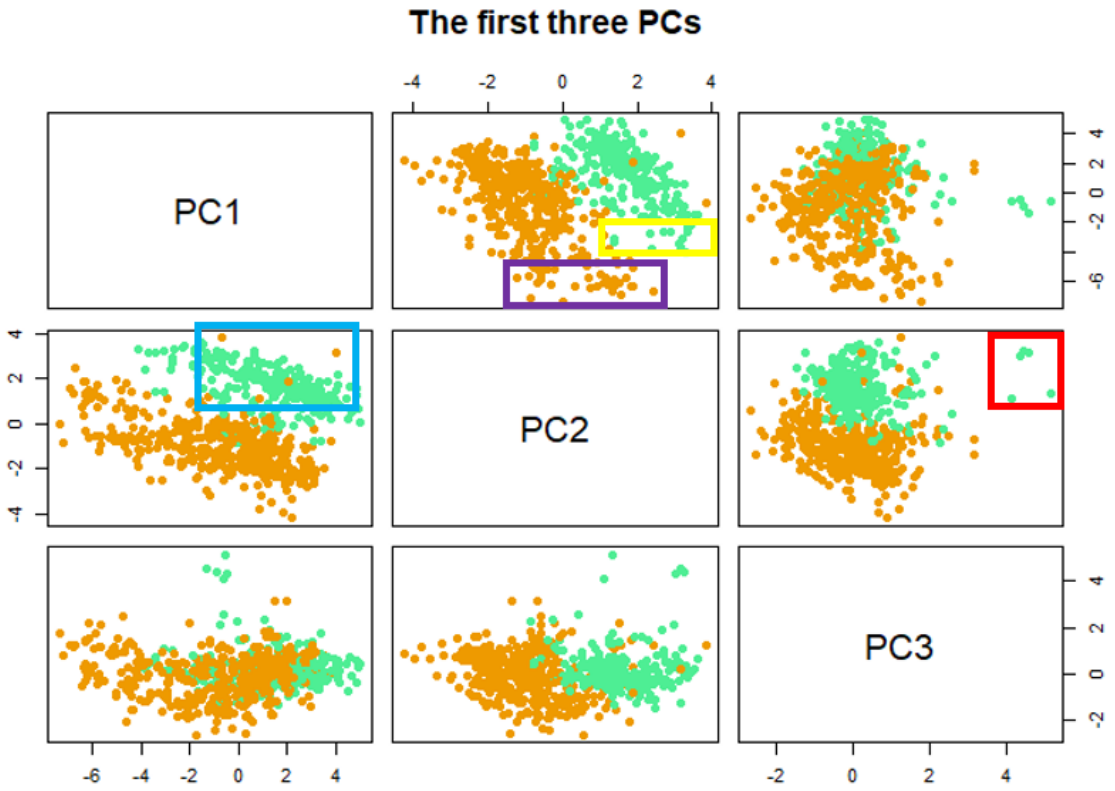
Now, let's look at the correlation matrix plot between the PCs and the variables themselves:



From this plot, we can make a couple of interesting observations:

- **High correlation** between both the *Rank* and the exam scoring variables with the first principal component. If we take a look at the scatter plot matrix between these groups, we see a clear linear relationship between them. Therefore, it is expected that if a principal component was able to capture the variability for these variables, both groups of variables would have to be highly correlated with the PC.
- However, observing the scatter plot matrix differentiating between groups, we see that the plots regarding these two groups of variables together are not able to differentiate well between private and public universities. Therefore, we can conclude that this component will probably not be adequate to distinguish between groups.

We will now begin with the analysis of how well these PCs are able to differentiate between groups and identify outliers.
This is the scatter plot matrix for the PCs scores.



As we can see, the second component is able to separate very well between private and public universities (green is public, orange is private).
As for the remarked squares, the purple one shows the subset of private universities which are easily identified by how good they are (according to the first PC). In other words, the best/most requested universities are private. On the other hand, the yellow square represents the best public universities.

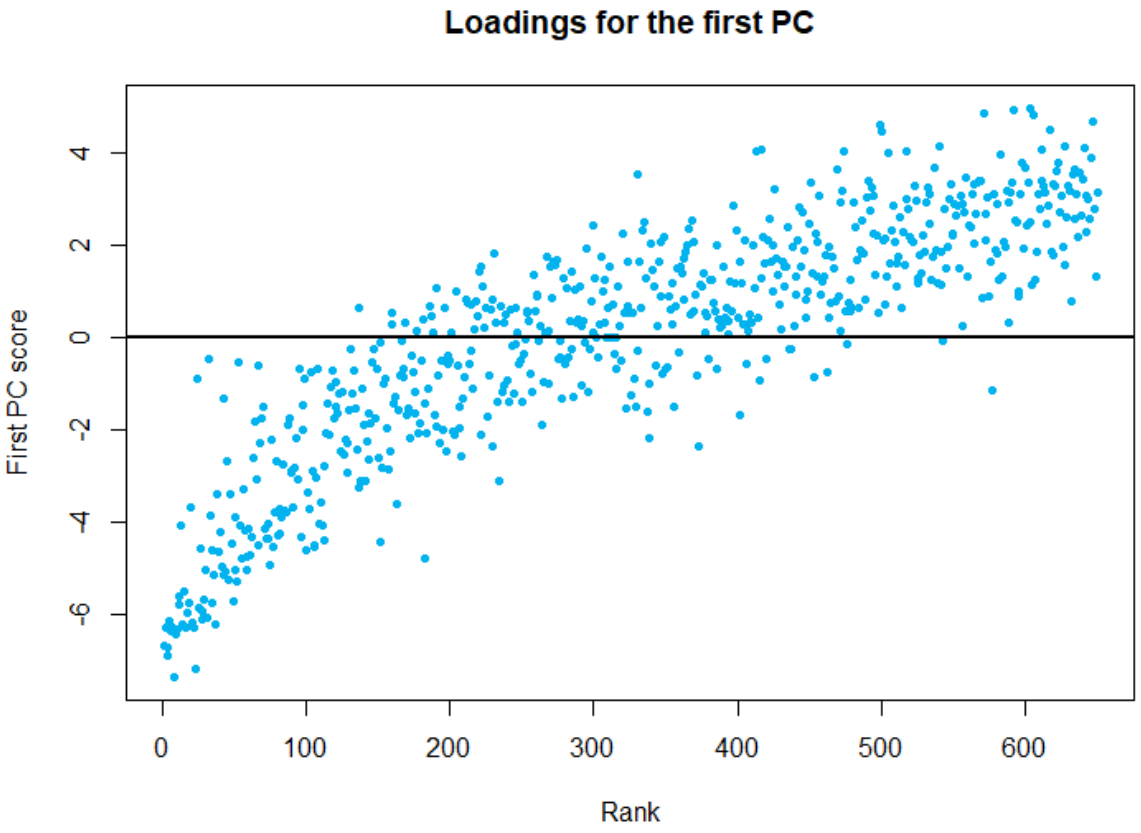
Regarding the identification of outliers, one could say that the points inside the red box are outliers, since they supposedly represent public universities that are much more expensive than private ones (assuming that the third PC represents the general financial cost).
However, if we look at the top five most expensive universities (according to *Net price*, which was the most prevalent variable in the third PC), we see that they are all private!

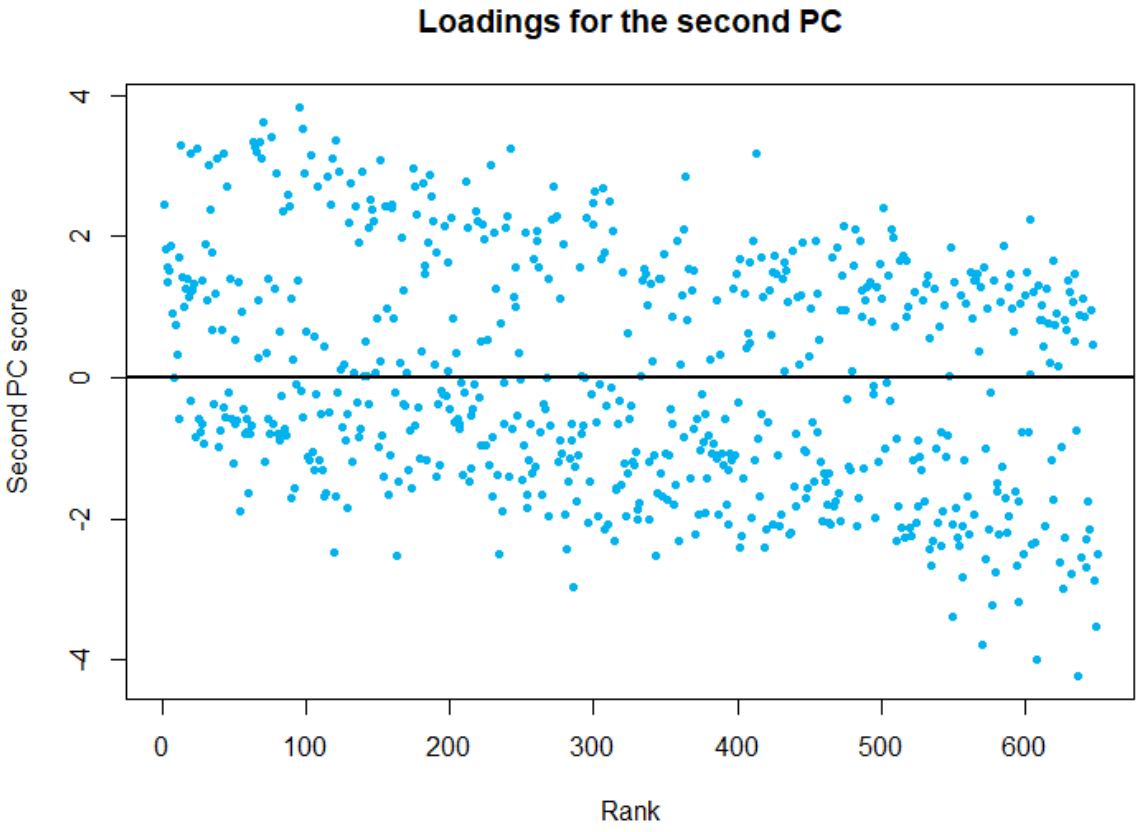
```
> X_imp[which(X_imp$Net.Price>sort(X_imp$Net.Price, decreasing=TRUE)[6]),
c('Rank','Net.Price','Public.Private')]
    Rank Net.Price Public.Private
112  112     46277        Private
163  163     47116        Private
328  328     43969        Private
402  402     46101        Private
404  404     47270        Private
```

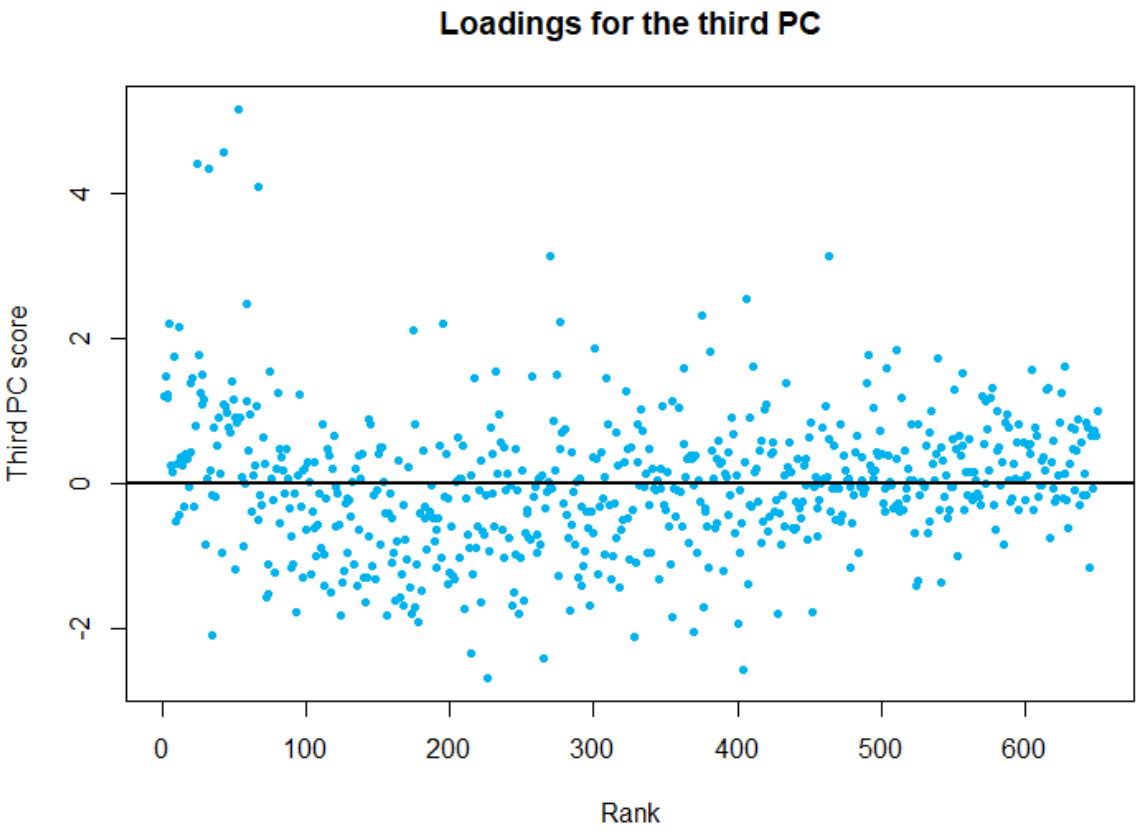
These observations help us better understand that the third PC is not as obvious as the financial cost of the university.

Now, the points inside the light-blue box could in fact be considered as outliers, since they represent private universities with a worse rank than usual. Remarkably, these points also tend to have a higher student population than usual, which confirms our interpretation of the second principal component.

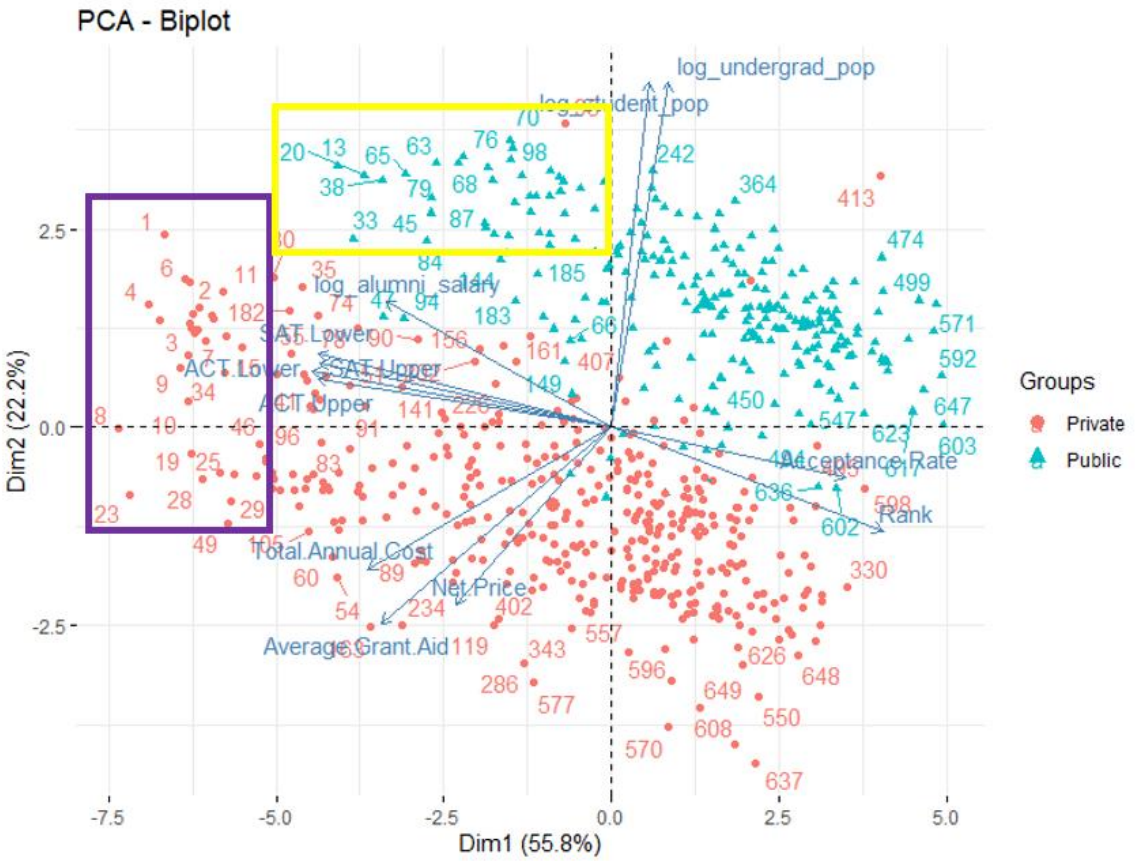
Continuing with the identification of outliers, in these plots we can observe that only the third component is able to identify some points as extreme, although their interpretation is not clear. Additionally, the first figure shows an observation we previously made: the high linear relationship between the *Rank* and the first principal component. As for the second plot, we can intuitively see the differentiation between two groups, which we now know correspond in fact to public and private universities.




Loadings for the second PC
Second PC score
4
2
0
-2
-4
0
100
200
300
400
500
600
Rank


Loadings for the third PC
Third PC score
4
2
0
-2
0
100
200
300
400
500
600
Rank

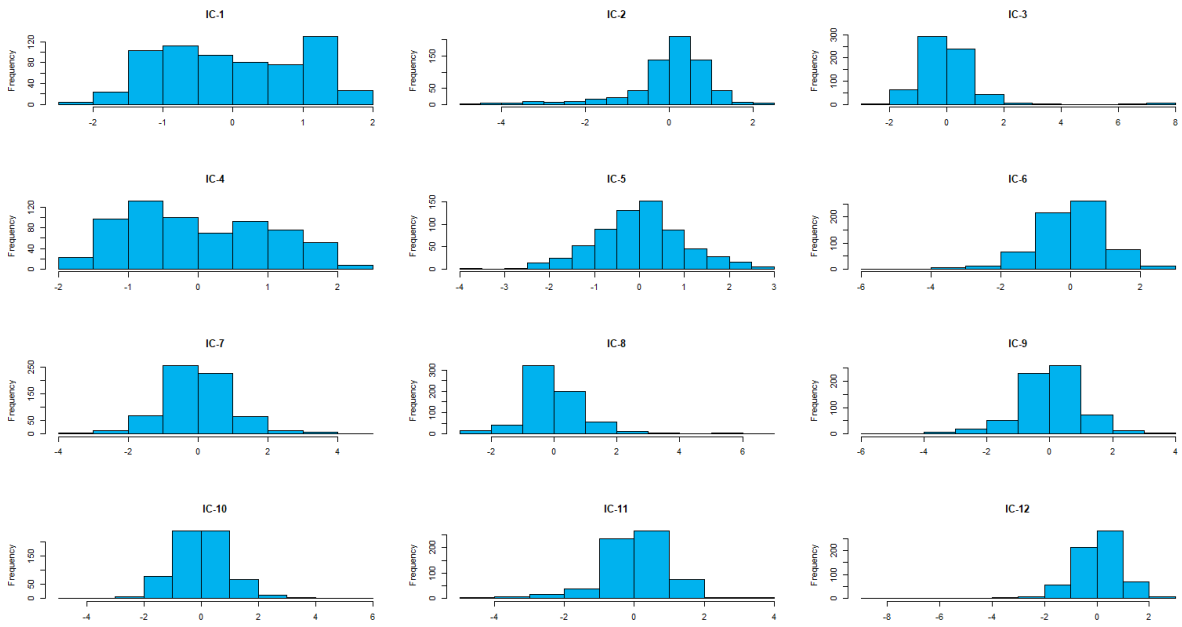
We conclude this section with the biplot in terms of the first two components, developed using the *factoextra* package.



This plot shows the observations we previously made in a compact manner. Both boxes are equivalent to the ones we highlighted in the same colors (best private and public universities). Additionally, we are able to see the importance of each variable with regards to each principal component, as we discussed before.

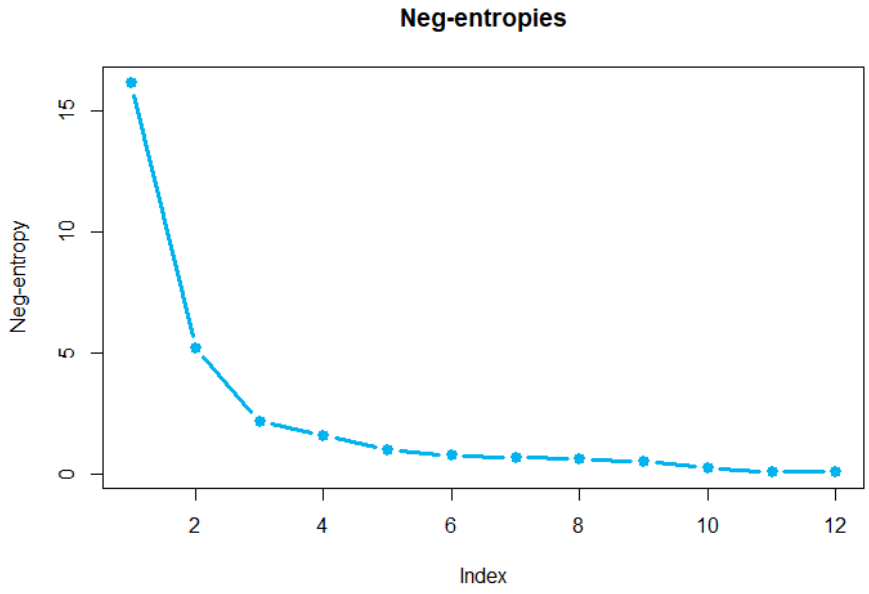
5.2 Independent Component Analysis

Let's see what results does ICA provide us with. First, we will take a look at the histograms of the scores for all the components.



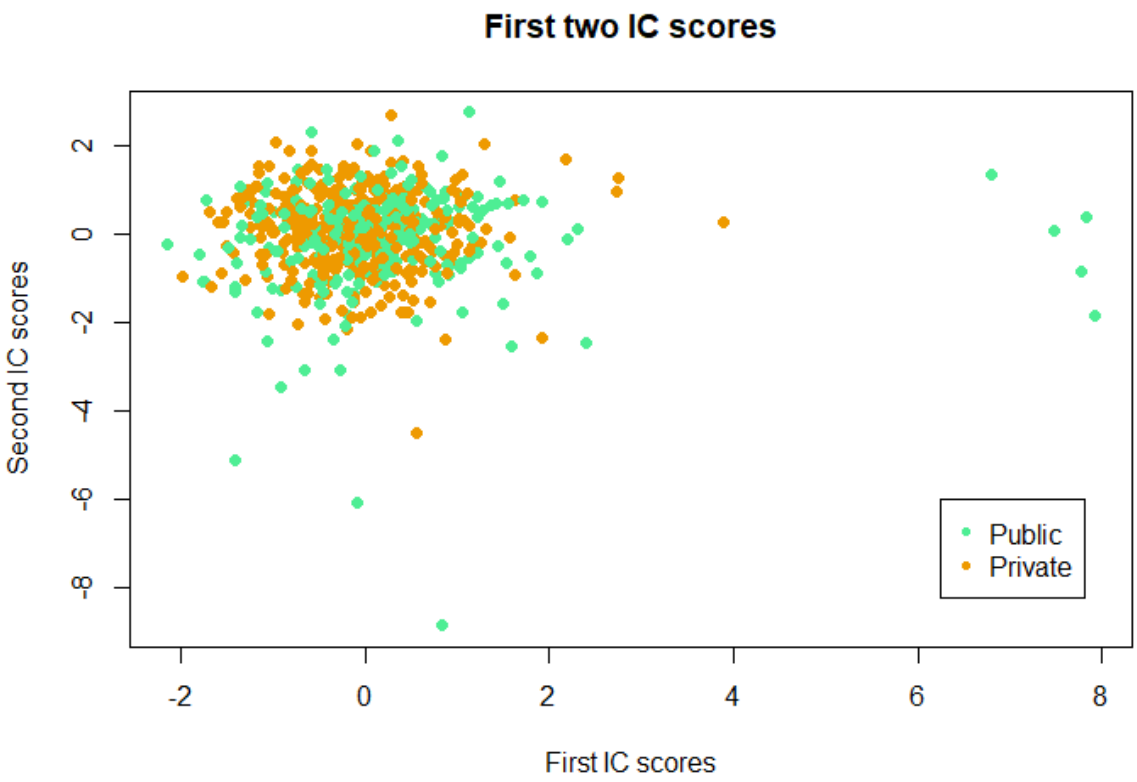
We expect to get non-Gaussianity, as the neg-entropy is maximized. In some cases, we see more evidence of non-Gaussianity than others. For instance, IC number 5 is clearly Gaussian-like, but IC-1 and IC-4 are not. As for the others, it's hard to tell. In most cases, the tails are too small to be Gaussian.

Below is the plot of neg-entropies for all the components, sorted in decreasing order.



We also previously computed the ICs according to this order. Therefore, we can say that the third IC has a much larger neg-entropy than the rest of the ICs. The next one in order, the twelfth one, also has high values of neg-entropy, as it is close to 5 and for Gaussian data, we expect a score of 1/16 (close to 0).

Now, we can look for the presence of outliers according to the two largest ICs in terms of entropy.



We see some clear outliers for the largest values of the first IC, and other not so apparent outliers for the lowest values of the second IC. These points refer to the following universities:

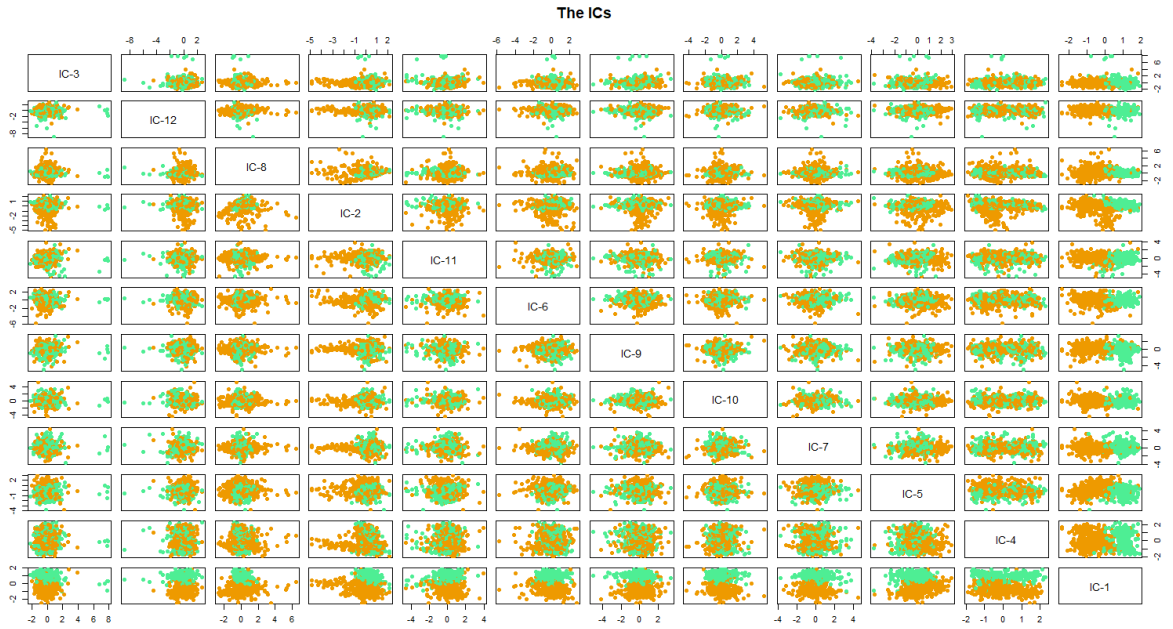
```
> data[which(Z_ic_imp[,1]>6),c('Rank','Name')]
    Rank                                    Name
24    24              United States Naval Academy
32    32           United States Military Academy
43    43         United States Air Force Academy
53    53      United States Coast Guard Academy
66    66 United States Merchant Marine Academy
```

Interestingly, we see the first IC is able to identify non-academic public institutions, regarding several military entities.

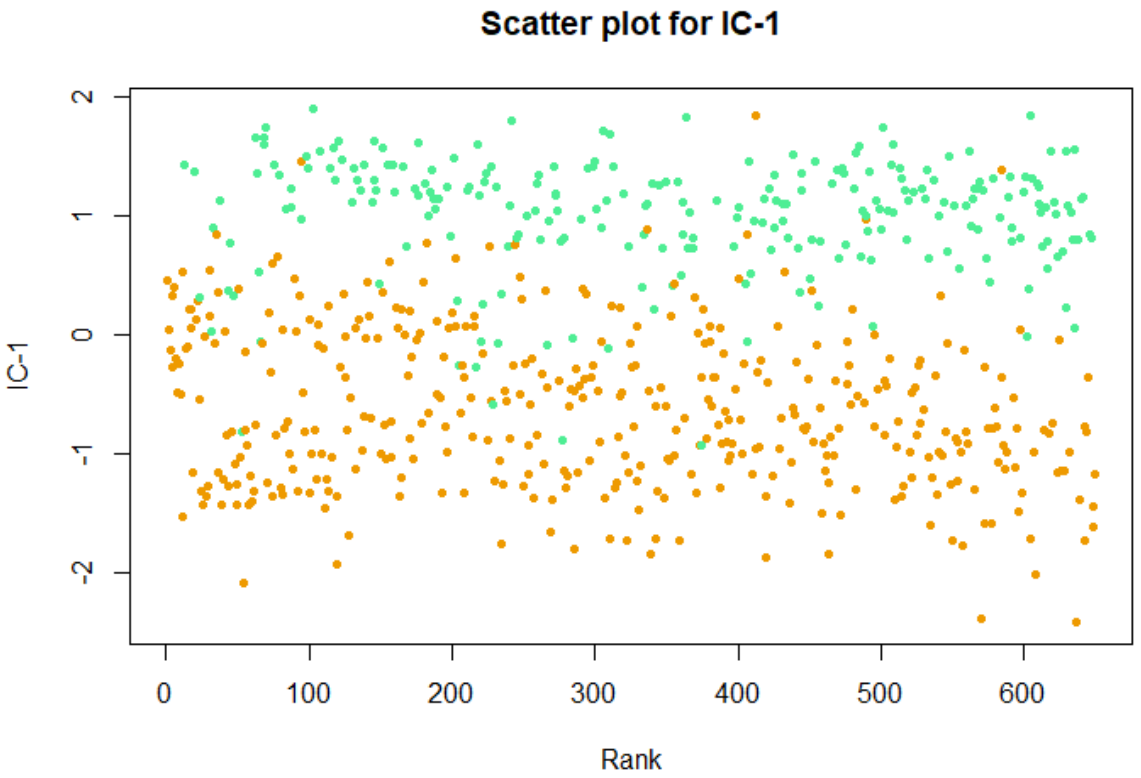
As for the other outliers, they refer to these universities:

```
> data[which(Z_ic_imp[,2]<(-5)),c('Rank','Name')]
     Rank                                      Name
65     65            Georgia Institute of Technology
555   555         University of Louisiana, Lafayette
629   629 Southern Illinois University Carbondale
```

We will now display the scatter plot matrix between all the ICs, ordered by neg-entropy:

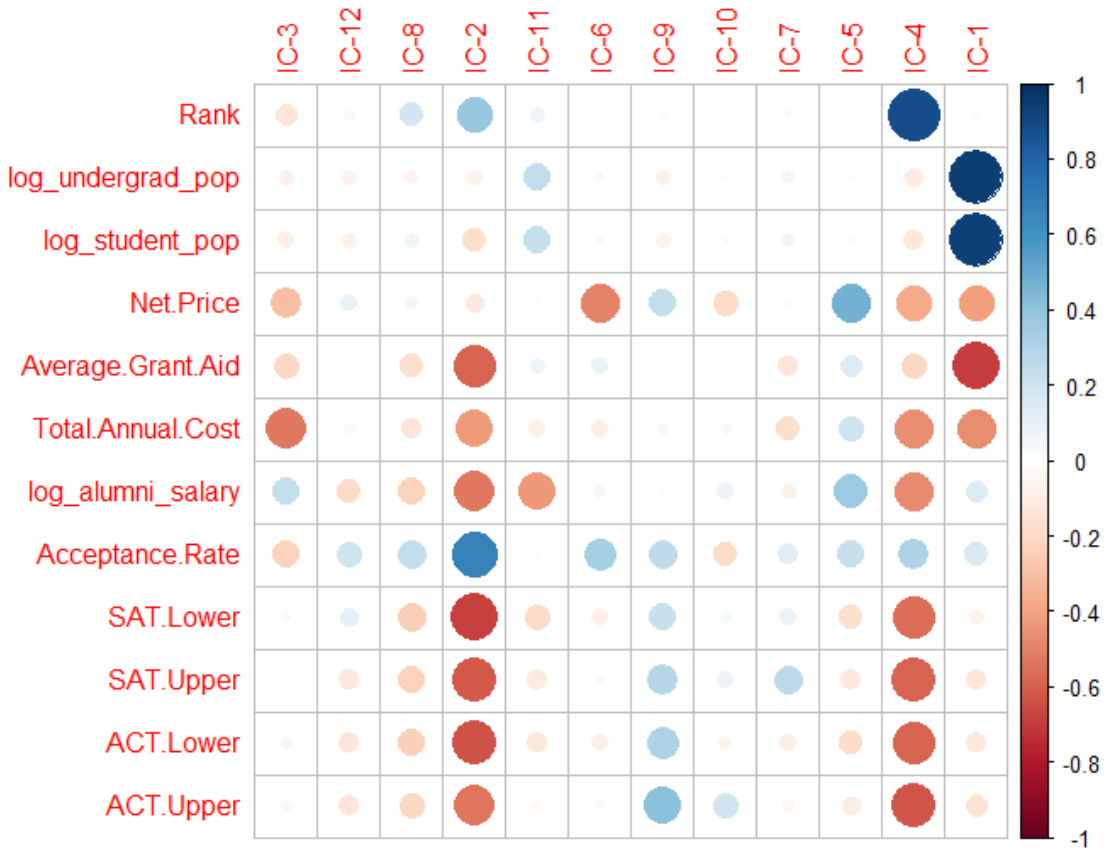


Regarding the presence of outliers, we can see they are apparent in the ICs with highest neg-entropy.
As we can see, the last IC (the one with least neg-entropy, number one), is able to distinguish very well between the groups across all pairs of plots.



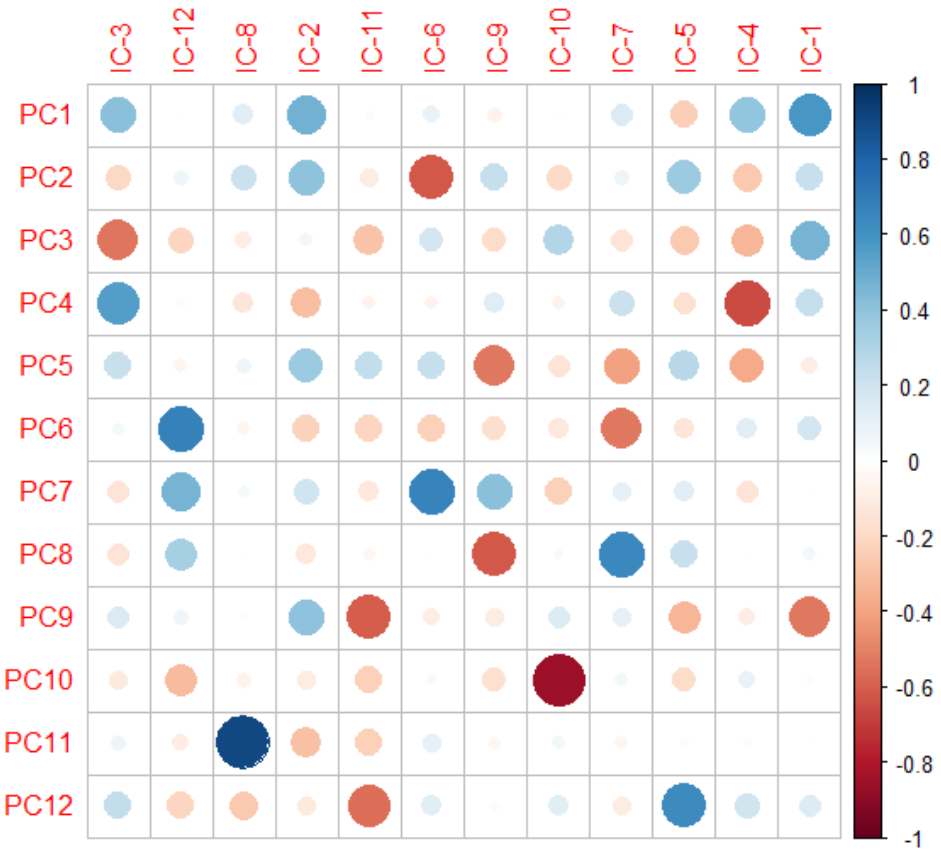
Looking at its scatter plot, we can see a clear separation between private and public universities, similarly to what we had for the second principal component.

Finally, we will look at the correlations between the ICs and both the variables and PCs.



The relationships in general are not very strong.
However, we can remark the following statements:

- IC-2 has high correlation with all exam scoring variables, and some financial ones.
- IC-4 is quite similar to the first PC, with medium-high correlation with *Rank* and the exam variables.
- IC-1 seems to resemble the second PC, which similarly, was able to distinguish very well between groups.

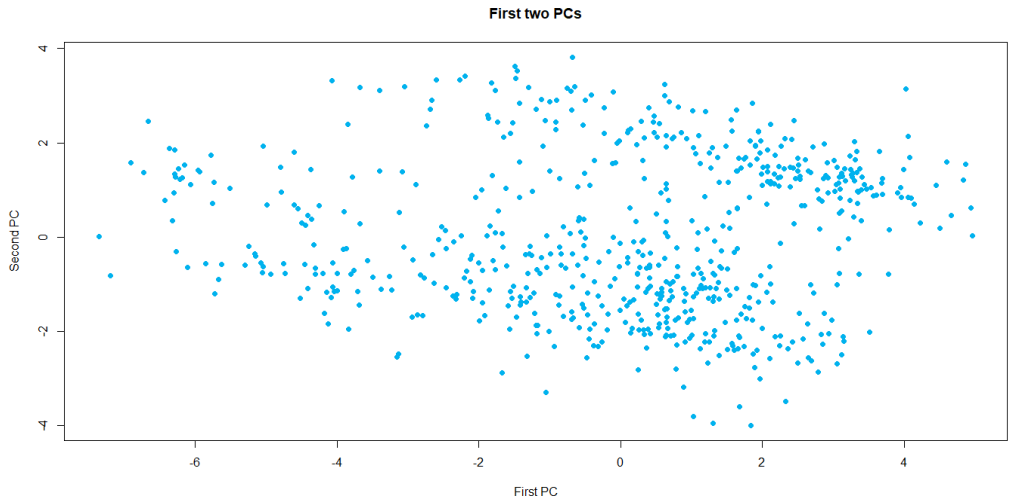


In this case, we see that in fact, the observations we previously made do not seem to hold. That is, the first PC is not highly linearly related to IC-4, and the same for PC-2 and IC-1. The only explanation we can give for this is that they are related, but **not in a linear way**.
The linear relationships that are worth noting are between PC-11 and IC-8 and PC-10 with IC-10.

6. Unsupervised classification

The main goal of the Unsupervised classification, also known as *clustering*, is to group the objects in a multidimensional dataset into different homogeneous groups. We will see in this section how to estimate the best number of clusters *K*, different procedures to obtain these clusters and we will analyze the apparent differences between those clusters.

Firstly, we start plotting the first two principal components of the dataset:



Looking at the plot above, we can appreciate at least two or more groups the data could be grouped into.

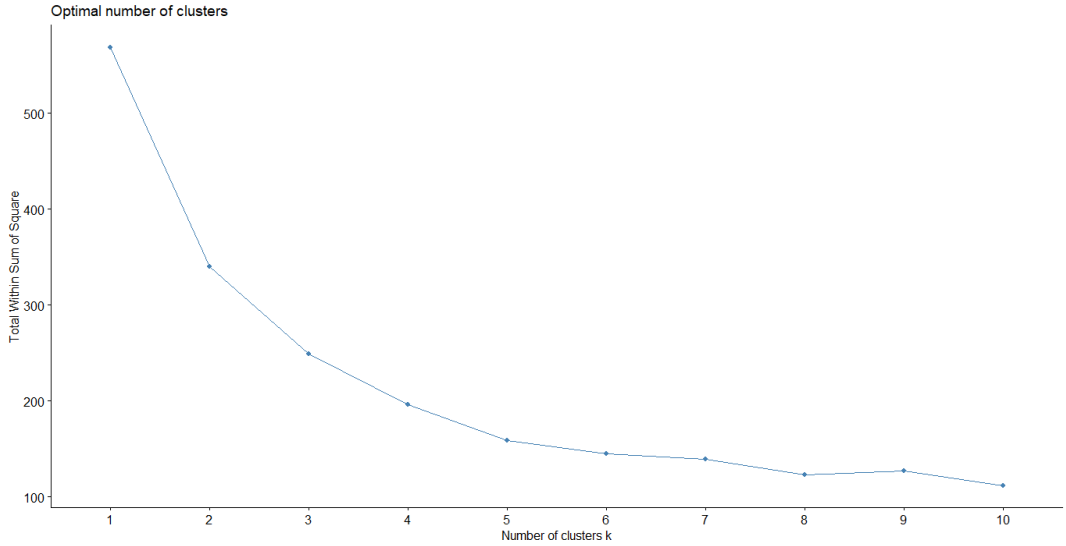
We will use and compare in this section 3 different clustering procedures: Partitional clustering, Hierarchical clustering and Model-based clustering:

1. PARTITIONAL CLUSTERING

This procedure starts from an initial cluster definition and proceeds by exchanging elements between clusters until an appropriate cluster structure is found.

First of all, we are going to estimate the optimal number of clusters *K* to divide the data into. For that, we will use three different methods: WSS (within-cluster sums of squares), average silhouette and Gap statistic.

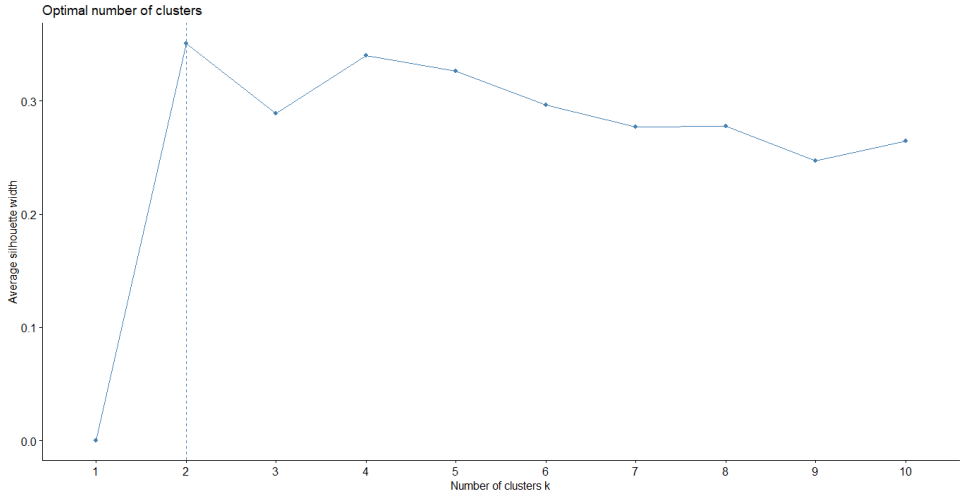
- ➢ WSS (Within-cluster sums of squares):



With this method, the optimal number of clusters is the one at which the WSS starts to stabilize. But looking at the plot above, we cannot select exactly one number of clusters as optimal, so we could think approximately of values between K=2 and K=5.

➢ Average silhouette:

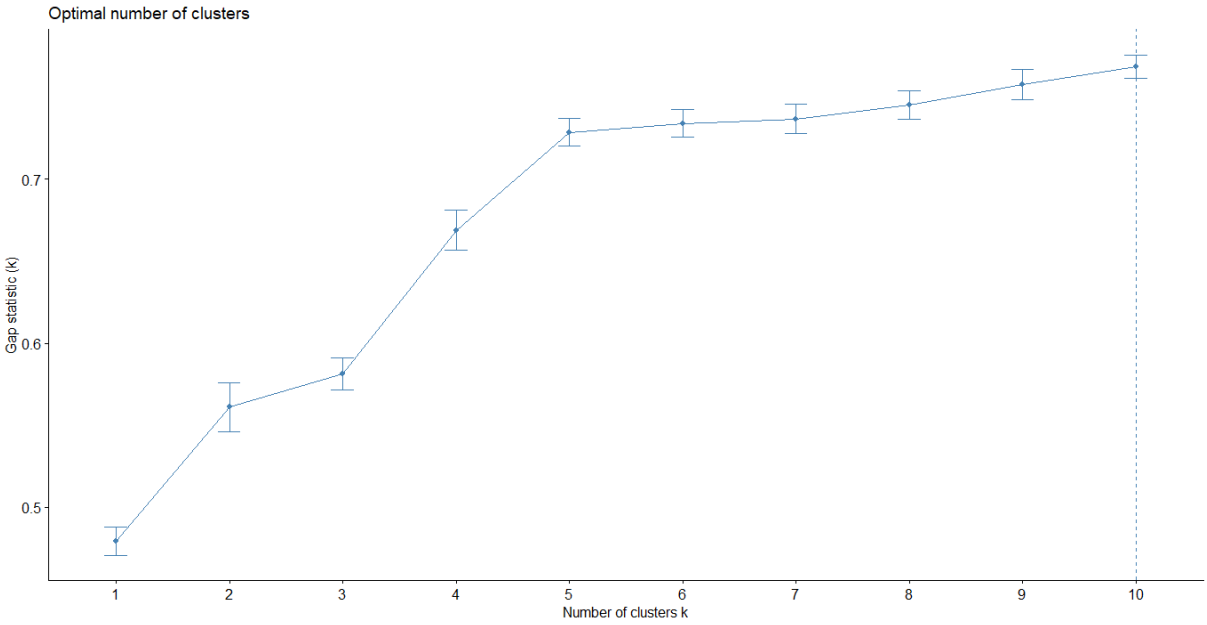
With this method, the more positive the average silhouette, the better is the configuration.



So, looking at the plot, we can see that in this case, K=2 seems to be the optimal number of clusters, but with K=4 and K=5 very close.

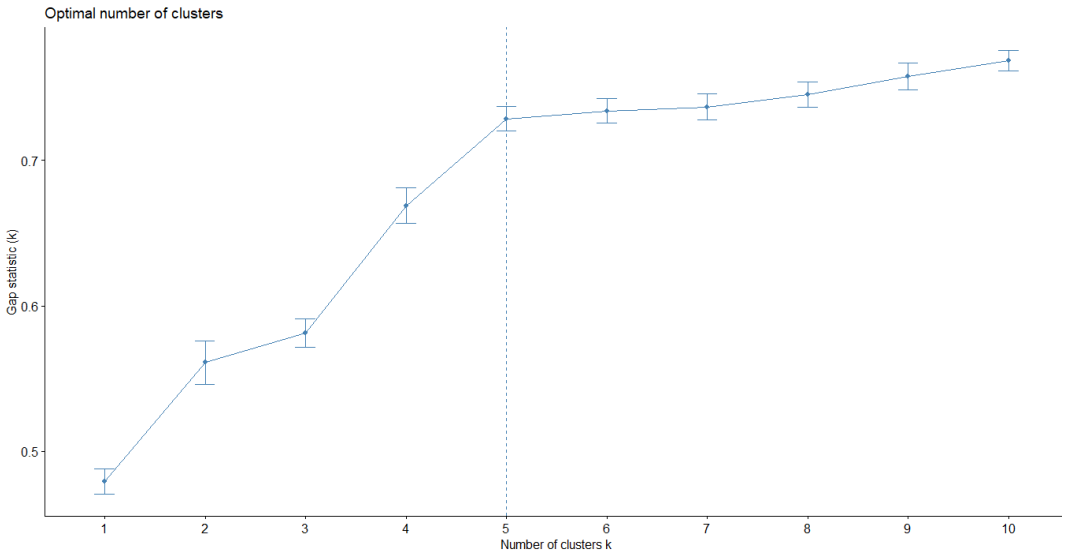
➢ Gap statistic:

Finally, with the Gap statistic method we look for the number of clusters *K* that corresponds to the first local maximum of the Gap statistic.



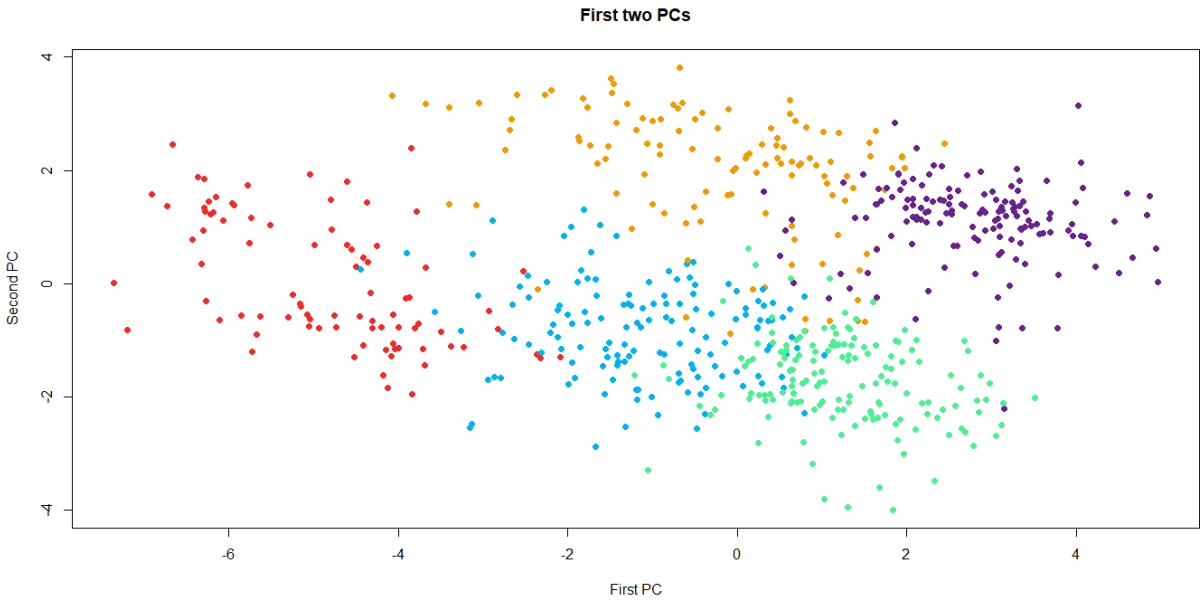
In this case, we can't see any local maximum in the plot, but for K=2 and K=5 we can appreciate a slight variation of the trend that may seem like slight peaks, especially for K=5.

To help decide between K=2 and K=5, we also tried computing the gap statistic with a different method called "*Tibs2001SEmax"* proposed in 2001 by Tibshirani et al. obtaining the following plot:

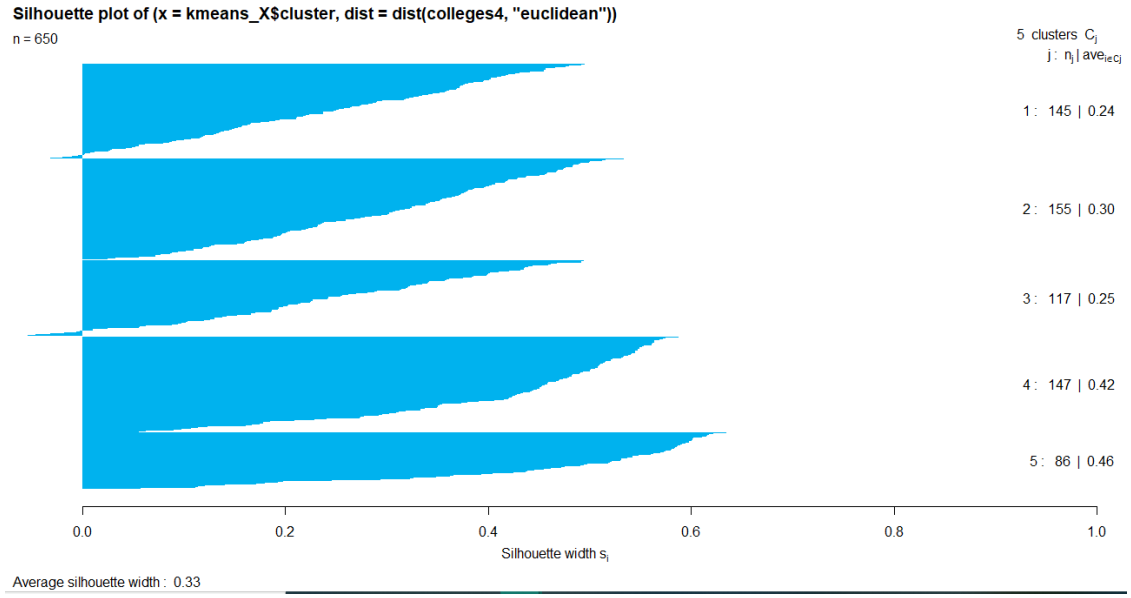


Now, by using this method, we can see that the number of clusters K=5 seems to be confirmed as the optimal one.

We proceed now to cluster the data with the **K-means algorithm** and **K=5** obtaining the following plots for the first two principal components and the silhouette values:



We can easily appreciate the five different clusters obtained. Later on, the apparent differences between the clusters will be analyzed.

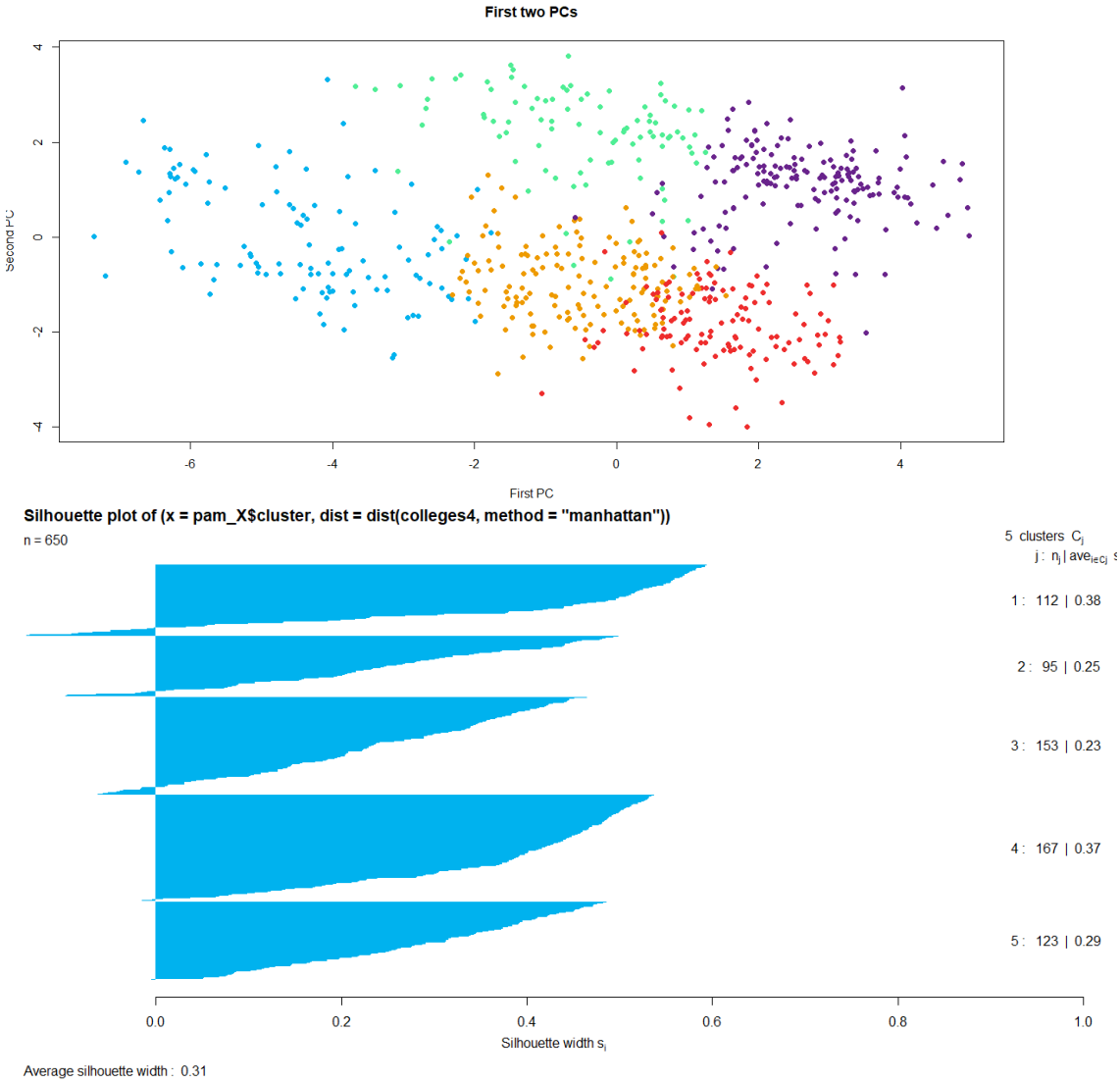


Looking at the silhouette plot, we can see that this is a valid configuration and the average silhouette width is 0.33.

Having selected the optimal number of clusters we continue the analysis by trying the **K-medoids clustering** (less sensitive to outliers), and two different algorithms to apply it. In this case, the sample mean vector is replaced with a central observation of the cluster, called the medoid.

- PAM algorithm (Partitioning around medoids):

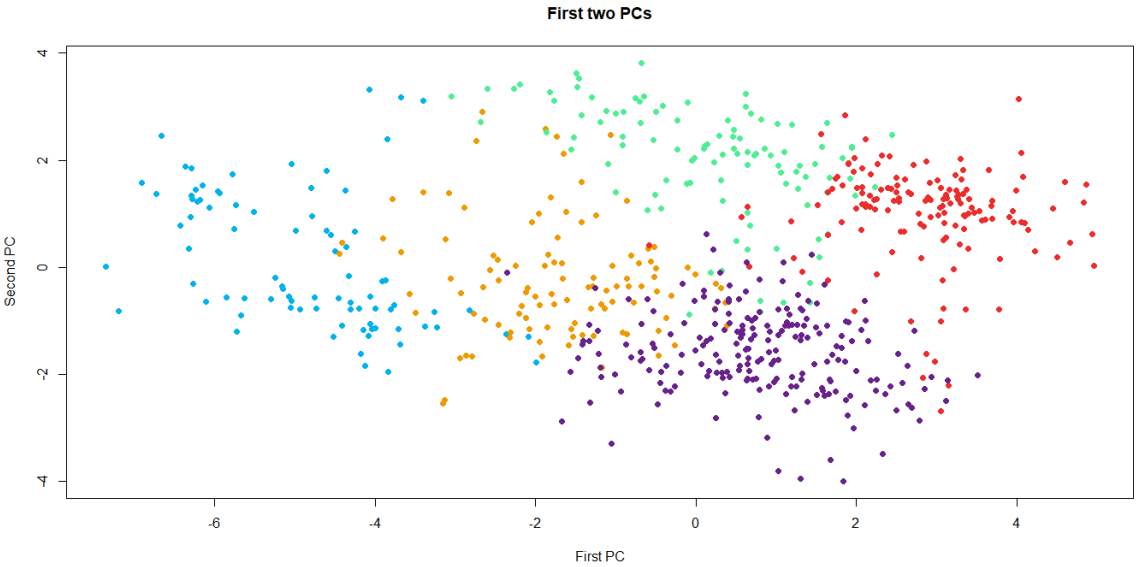
Applying the PAM algorithm we obtained the following plots.

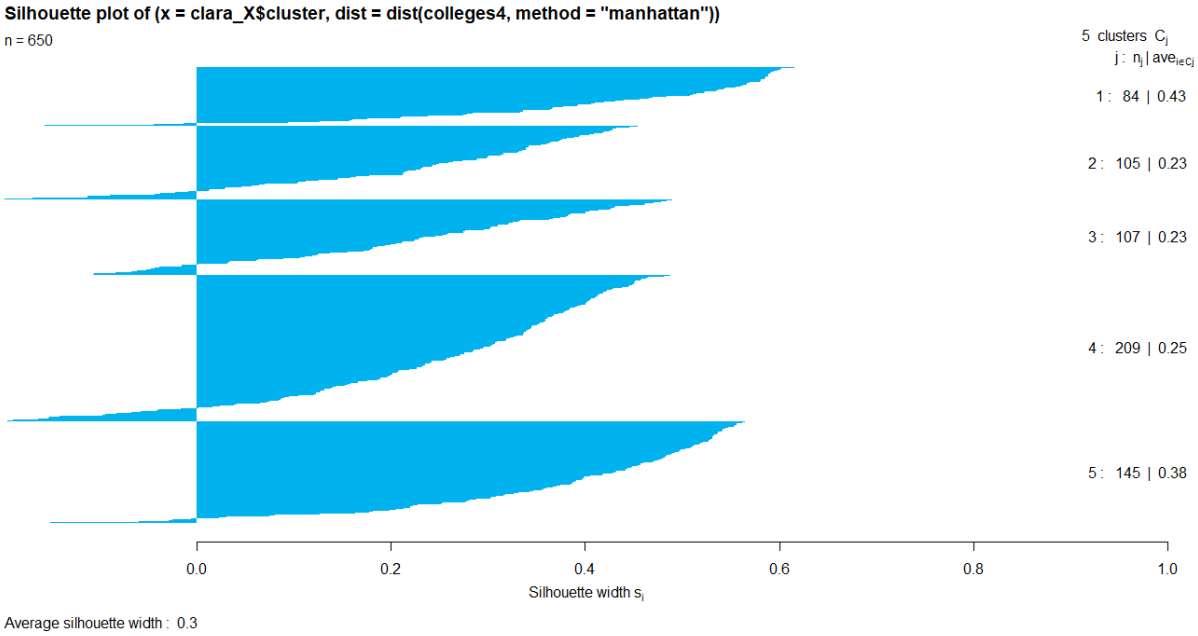


With similar results to the obtained from the K-means algorithm but now with apparently some more misclassified observations.

- Clustering for Large Applications (CLARA) algorithm:

Applying the CLARA algorithm we obtained the following plots, with results similar to before but in this case, as we can see in the silhouette plot, with even more apparently misclassified observations.





So, the K-means algorithm seems to be the best partitional clustering method, with an average silhouette width of 0.33.

The unscaled sample mean vectors of the clusters are the following:

	1	2	3	4	5
Rank	189.009	504.748	57.814	236.476	490.335
log_Undergraduate.Population	9.644	9.545	8.238	8.061	7.507
log_Student.Population	9.792	9.649	8.558	8.265	7.680
Net.Price	16206.744	14631.000	27934.488	30611.897	23336.729
Average.Grant.Aid	10299.137	8079.054	38809.477	24662.876	23558.484
Total.Annual.Cost	43167.590	35892.442	69305.767	59746.034	50092.555
log_Alumni.Salary	11.557	11.416	11.662	11.524	11.383
Acceptance.Rate	57.111	75.714	22.860	65.366	69.684
SAT.Lower	1147.547	1011.912	1324.640	1129.552	1031.471
SAT.Upper	1339.376	1216.483	1490.581	1328.034	1242.942
ACT.Lower	24.256	20.177	29.605	24.152	21.110
ACT.Upper	29.393	25.646	33.047	29.407	26.871

Looking at these vectors we can try to analyze the meaning and differences of each cluster. For example, cluster 3 presents high scoring variables, low acceptance rates and higher ranks, so it may be grouping the best ranked and more difficult to access universities. On the other hand, cluster number 2 seems to represent the opposite, universities with high acceptance rates, high populations, low scoring variables, low price and low ranks. Cluster number 4 can be representing universities with high costs but also high acceptance rates. Cluster number 5 is similar to cluster number 2 but in this case with higher costs and less population.

Comparing these clusters with the categorical variable of interest *Public.Private*, we can see that clusters 4 and 5 include all public universities, and private universities are divided mainly between clusters 1, 2 and 3.

```
           1   2   3   4   5
Private 155 145  85  10   5
Public    0   0   1 137 112
```

So we could conclude that cluster 4 includes public universities with highest ranks and scoring variables and cluster 5 includes public universities with lowest ranks and lowest scoring variables. And on the other hand, cluster 3 could represent best private universities, cluster 2 worst private universities, and cluster 1 could represent private universities with the highest population.

2. HIERARCHICAL CLUSTERING

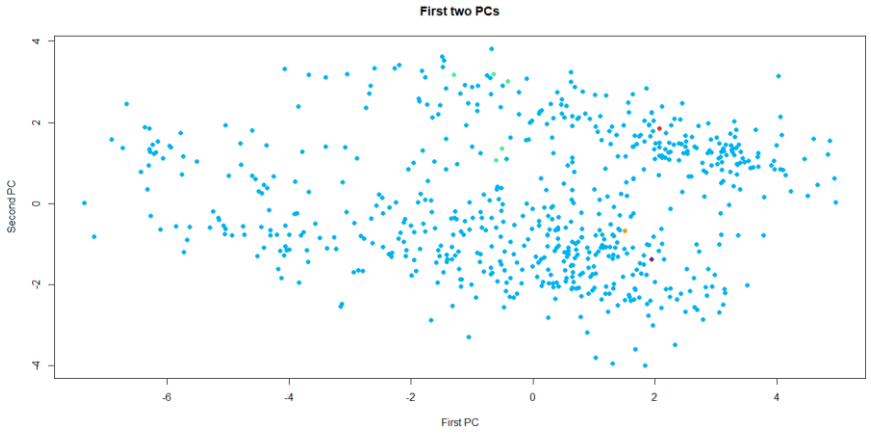
a) Agglomerative algorithms:

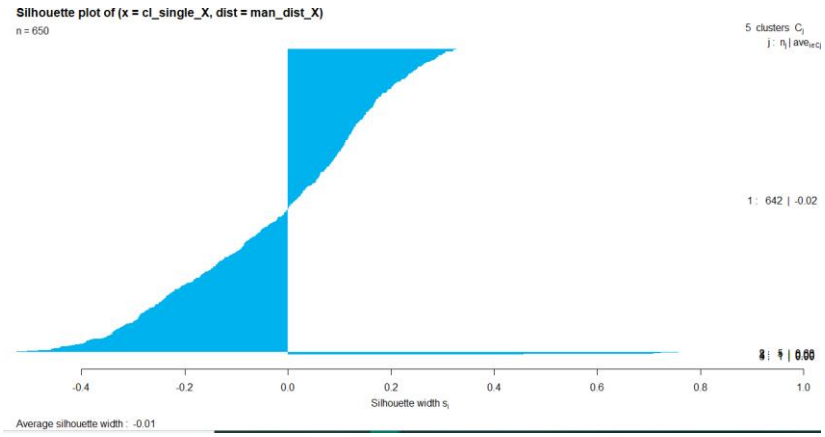
Start with clusters containing a single observation and continue merging clusters. Distances between clusters can be computed using the next linkage methods:

i) Single linkage

```
  1   2   3   4   5
642   5   1   1   1
```

As we can appreciate in the assignment, there is a very long group and isolated observations taking K=5. Looking at the plots from below, we can see that this is a very bad solution.

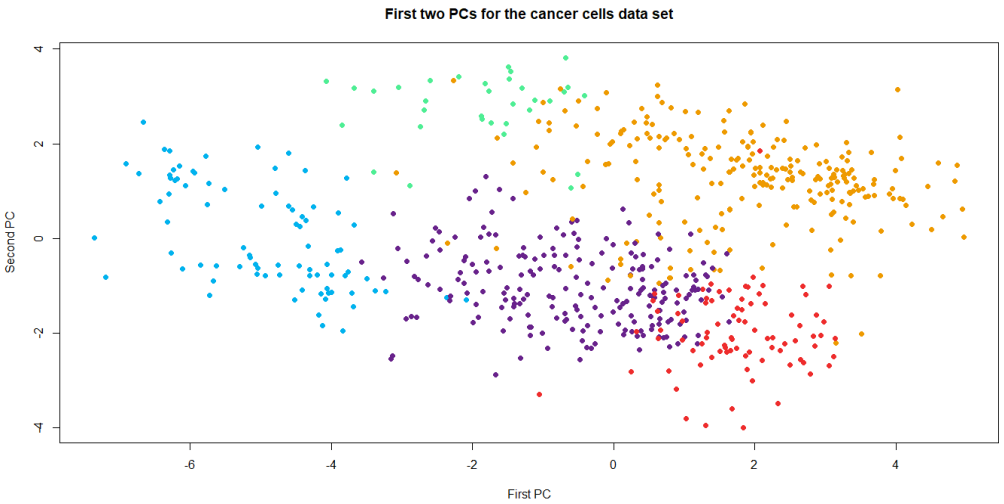


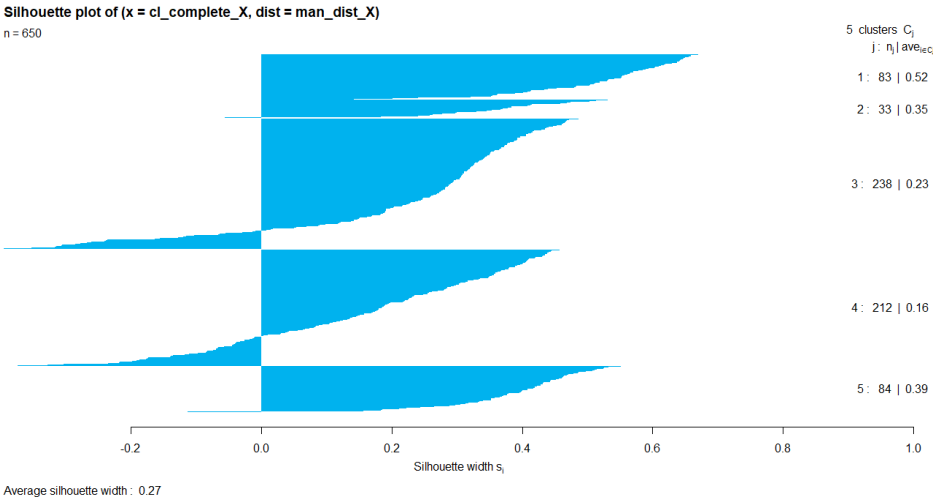


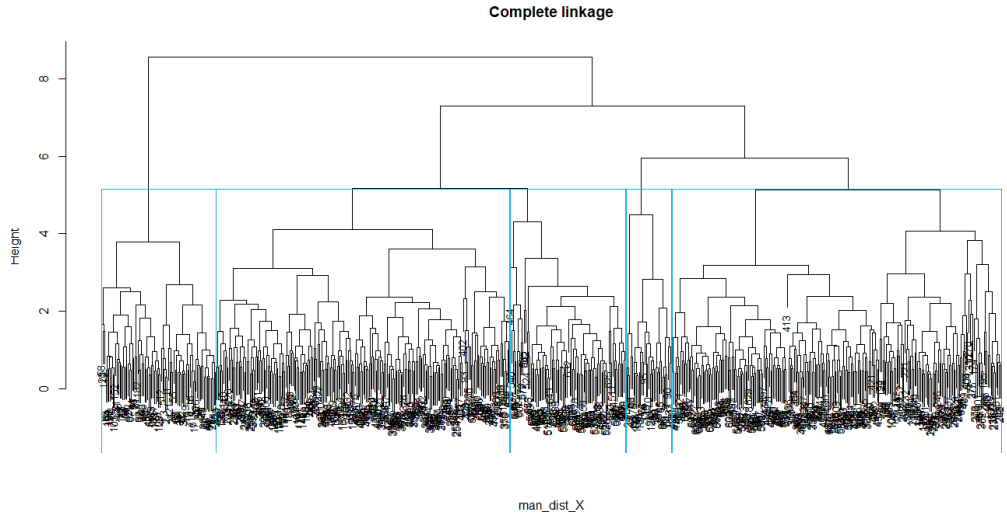
ii) Complete linkage

```
 1  2  3  4  5
83 33 238 212 84
```

In this case, there are several groups of moderate sizes, as we can see in the PCAs plot and the dendrogram from below, but even so, the solution is not better than the solutions from partitional methods.

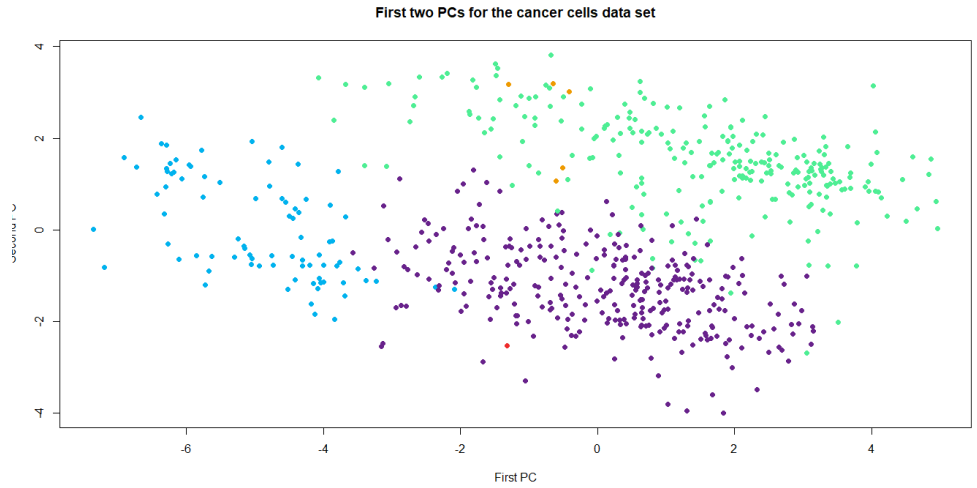






iii) Average linkage

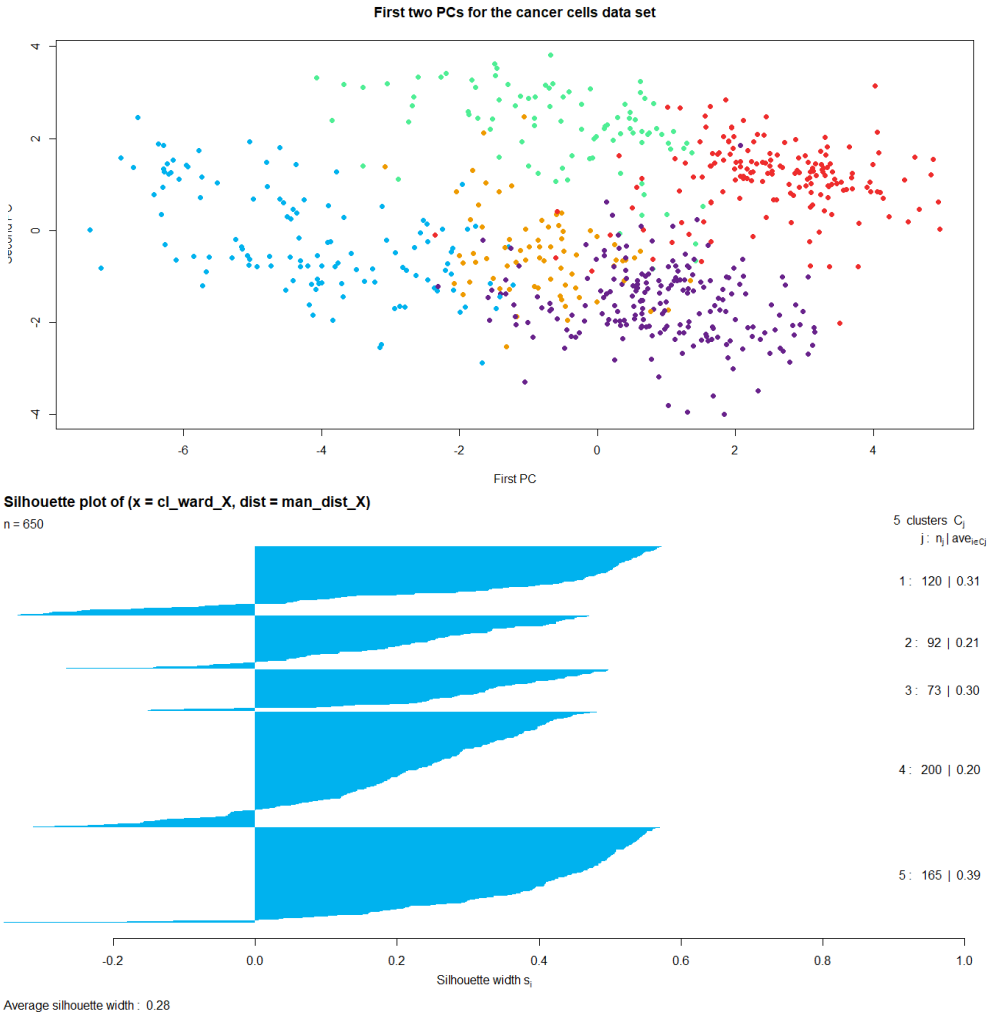
```
  1   2   3   4   5
 83 256   5 305   1
```



With the average linkage, the solution is not good either, with some large groups and other very small groups.

iv) Ward linkage

```
  1   2   3   4   5
120  92  73 200 165
```



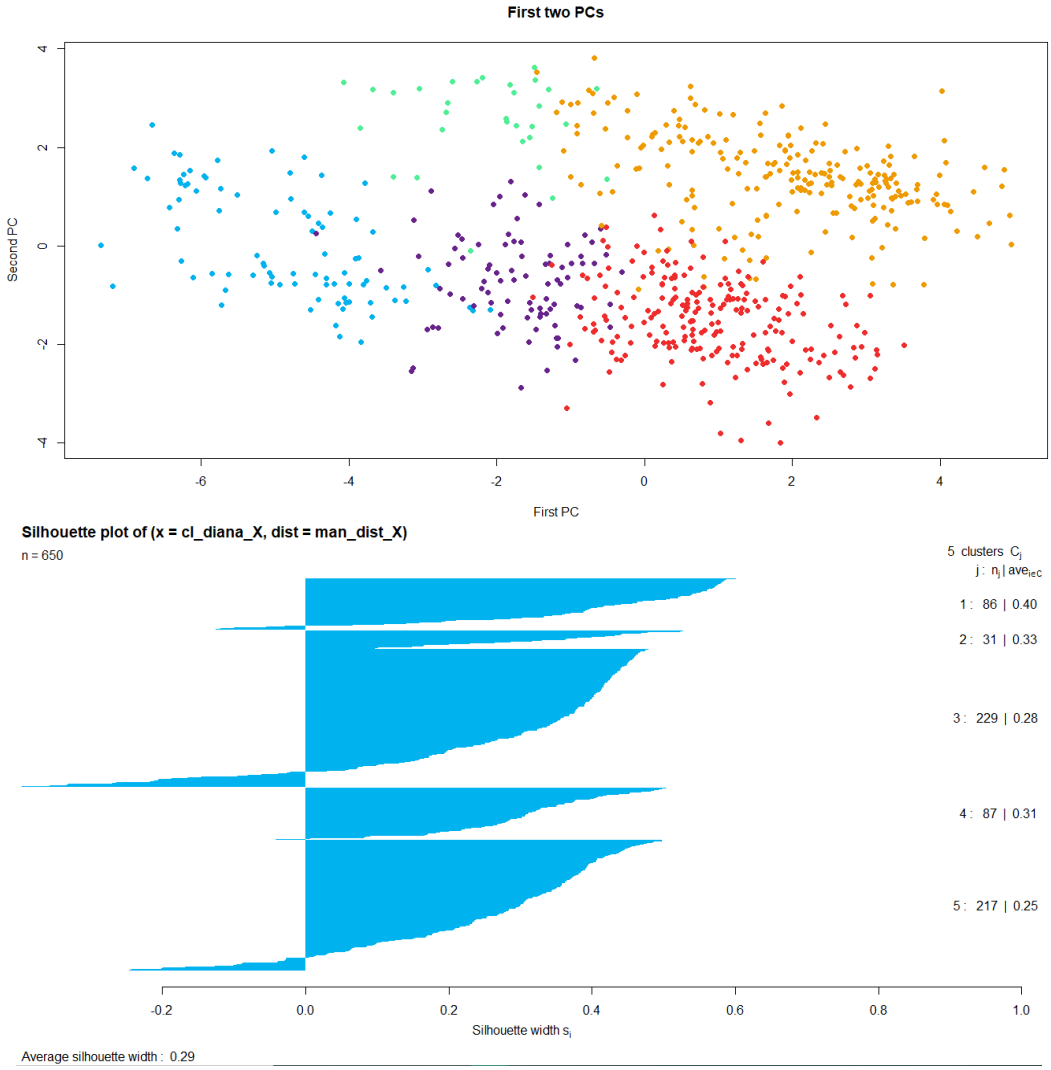
Even with moderate groups in this case, there are some misclassified observations, and the solution is not as good as the one from the partitional clustering.

b) Divisive algorithms:

We will continue trying and comparing different methods, in this case divisive methods, consisting in starting with a single cluster containing all the observations and continuing splitting clusters.

i) Divisive Analysis Clustering (DIANA):

```
  1   2   3   4   5
 86  31 229  87 217
```

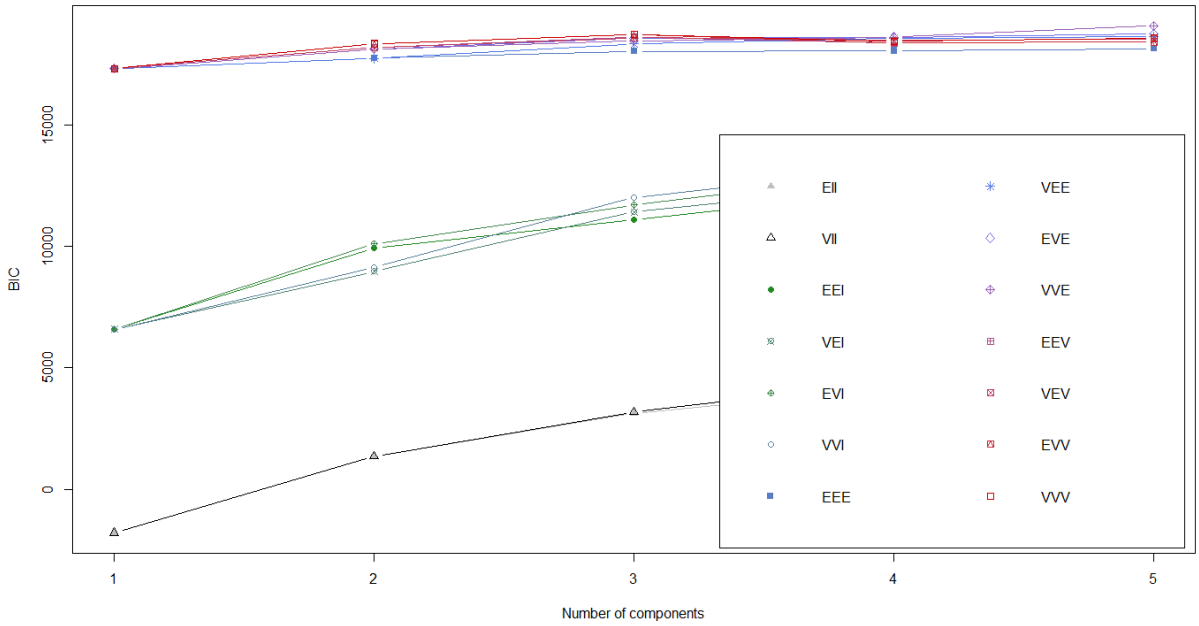


With some misclassified observations, and several moderate groups again, but the result is not as good as the one from the partitional method.

3. MODEL-BASED CLUSTERING

In this probabilistic method, it is assumed that the multivariate random variable has a mixture distribution, the mixture is fitted to the data, and observations are assigned to the mixture populations with the Bayes Theorem.

For each possible value of *K* and every configuration of Σk we apply the Bayesian Information Criterion (BIC) obtaining the following best results:




```
Top 3 models based on the BIC criterion:
VVE,5 EVE,5 VVV,3
19103 18766 18763
```

Thus, the model has 5 clusters with the covariance matrices ellipsoidal and with equal orientation. We will see later that sometimes the suboptimal options are better and we will analyze the configuration for the K=2 number of clusters.

Continuing with K=5:

```
Clustering table:
  1   2   3   4   5
 73  95 139 159 184
```

The estimated mixing probabilities are the following:

```
0.1113 0.1545 0.2171 0.2389 0.2782
```

And these are the estimated sample mean vectors unscaled for each cluster:

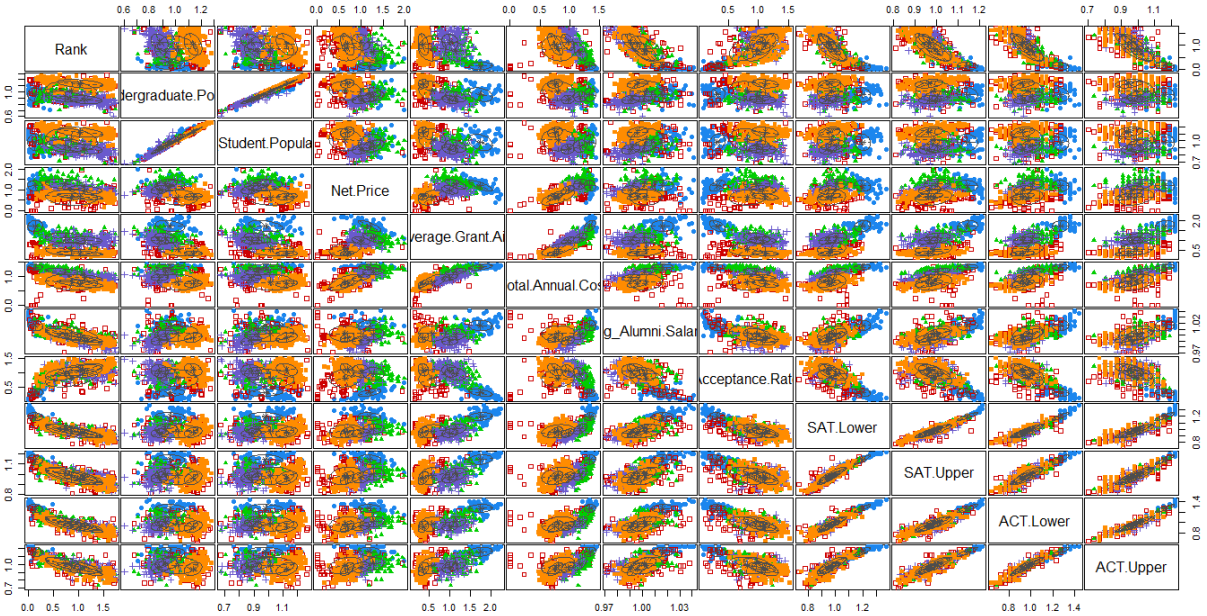
	[,1]	[,2]	[,3]	[,4]	[,5]
Rank	53.644	323.837	276.918	425.012	387.614
log_Undergraduate.Population	8.290	8.809	8.014	7.589	9.836
log_Student.Population	8.661	8.970	8.176	7.797	9.937
Net.Price	28466.722	15179.601	30419.058	23720.487	16290.516
Average.Grant.Aid	39760.061	13246.426	26099.079	22977.287	8300.824
Total.Annual.Cost	69402.557	41820.199	61046.598	50140.972	39230.069
log_Alumni.Salary	11.685	11.533	11.518	11.391	11.452
Acceptance.Rate	20.760	53.804	63.985	70.136	73.167
SAT.Lower	1336.718	1084.385	1123.331	1053.524	1066.645
SAT.Upper	1498.022	1291.883	1313.460	1268.878	1263.325
ACT.Lower	29.962	22.853	23.839	21.804	21.631
ACT.Upper	33.195	28.240	28.892	27.609	26.977

In this case, we can interpret the clusters in the next way. Cluster 1 seems to include best ranked universities with lowest acceptance rates. Cluster 5 seems to include cheapest universities with highest acceptance rates. Cluster 4 could represent universities with low population, maybe because they are the worst ranked ones.

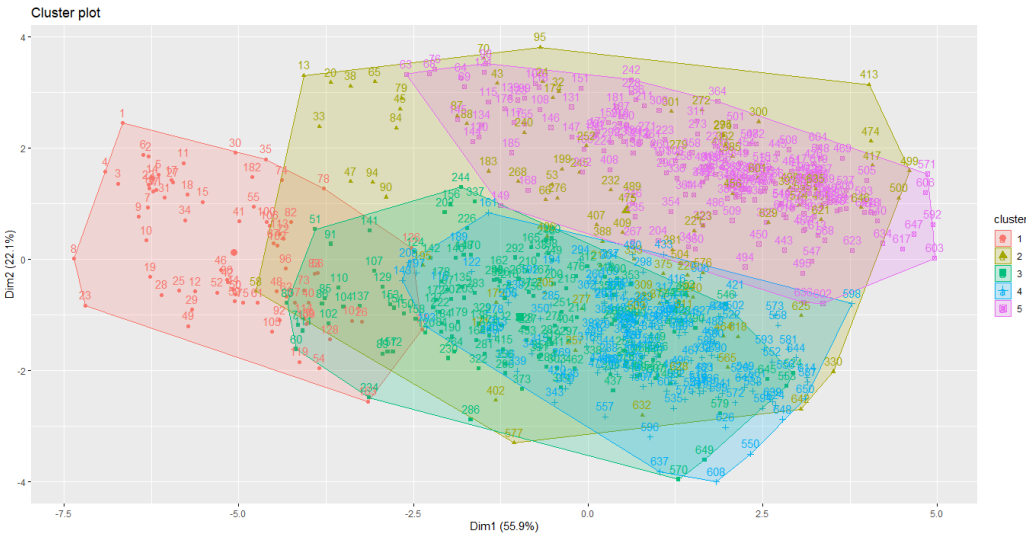
Comparing these clusters with the categorical variable of interest, we can see that clusters 2 and 5 include all the public universities, differentiating mainly by population and acceptance rate. And cluster 1 includes, as we suspected, the best universities that tend to be private.

	1	2	3	4	5
Private	73	28	139	159	1
Public	0	67	0	0	183

Plotting the clusters over the different scatter plots we obtain the following:



Very difficult to distinguish anything. Looking at the first two PCs on the plot below , we can differentiate better the clusters but we can see that they are somehow overlapping.



As we said before, sometimes the suboptimal options of the BIC criterion are better. And as we could see in the graph of the BIC graph, the values for K equals 2, 3 or 4 were very close to the BIC values for K=5. So we are going to try forcing K=2. In this case the configuration

would be VVV, i.e. the covariance matrices are ellipsoidal, with varying volume and orientation.

Looking at the sample mean vector of the cluster obtained we can differentiate two clusters. Cluster 1 with less population, higher prices and scores and less acceptance rate that could represent private universities as we saw in previous sections, and cluster 2 representing universities with higher populations, lower prices and scores, and higher acceptance rates, that might be referring to public universities.

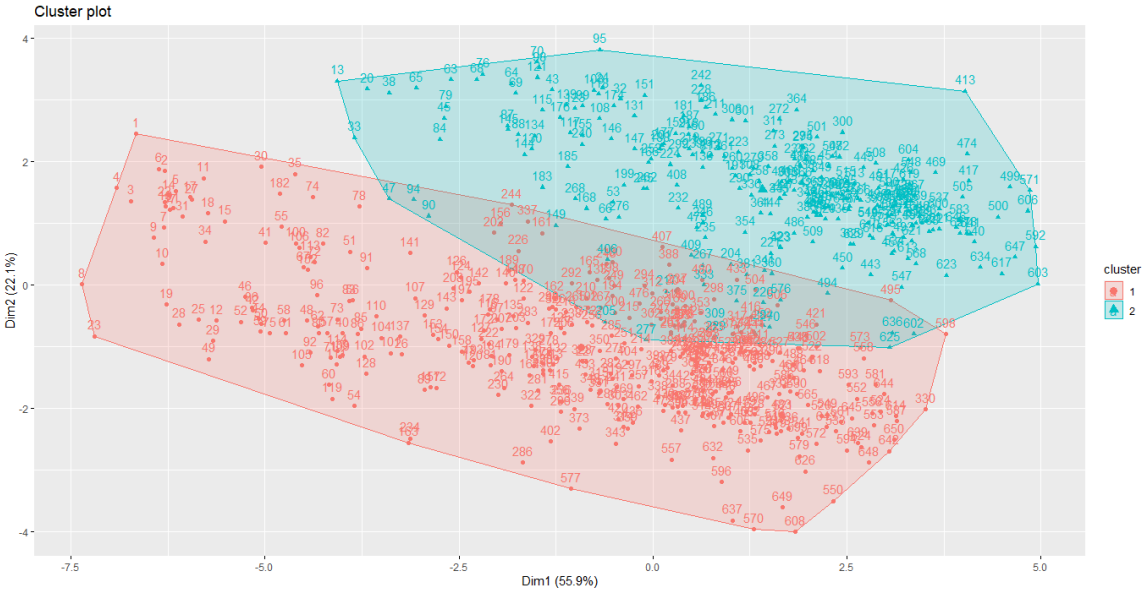
	[,1]	[,2]
Rank	302.037	361.149
log_Undergraduate.Population	7.871	9.640
log_Student.Population	8.090	9.759
Net.Price	26976.443	15234.486
Average.Grant.Aid	27131.915	9000.998
Total.Annual.Cost	57574.141	39324.040
log_Alumni.Salary	11.495	11.483
Acceptance.Rate	57.990	67.150
SAT.Lower	1131.503	1073.825
SAT.Upper	1328.253	1272.349
ACT.Lower	24.085	22.027
ACT.Upper	29.150	27.338

To check our assumptions we are going to compare these clusters with the categorical variable of interest:

	1	2
Private	392	8
Public	0	250

And confirming our assumptions, these clusters divide almost perfectly the dataset into two groups, the first one for private universities and the second one for public ones.

We can see the clusters well differentiated in the next plot with the first two PCs.



7. Supervised classification

The goal of supervised classification is to assign a new object to a class from a given set of classes based on the attribute values of this object and on a training set.

There are so many different methods available to perform supervised classification. There is no method which is better than the other for all the datasets so it is important to determine which classifier obtains better results for a certain data set. For that we need some measure of the quality of the results given by the different classifiers to determine which classifier gives the best performance.

We want our classifier to work with responses that haven't been used to estimate the parameters, so we divide the data matrix X and the response vector Y into two samples, one for estimating the parameters called training sample and one for measuring the classifier called test sample.

We will use 70% of our data for the training sample and the remaining 30% for the test sample. Therefore, we will have 454 rows for training and 196 for testing of the total of 650 rows.

For measuring the quality of our classifier we will use the Test Error Rate (TER). The classifier will have a good performance if TER is close to 0 and a bad performance if TER has a high value.

As it was mentioned above there is no method that is the best for all possible data sets but it is possible to prove that under a probabilistic framework the expected TER is minimized with the Bayes rule method.

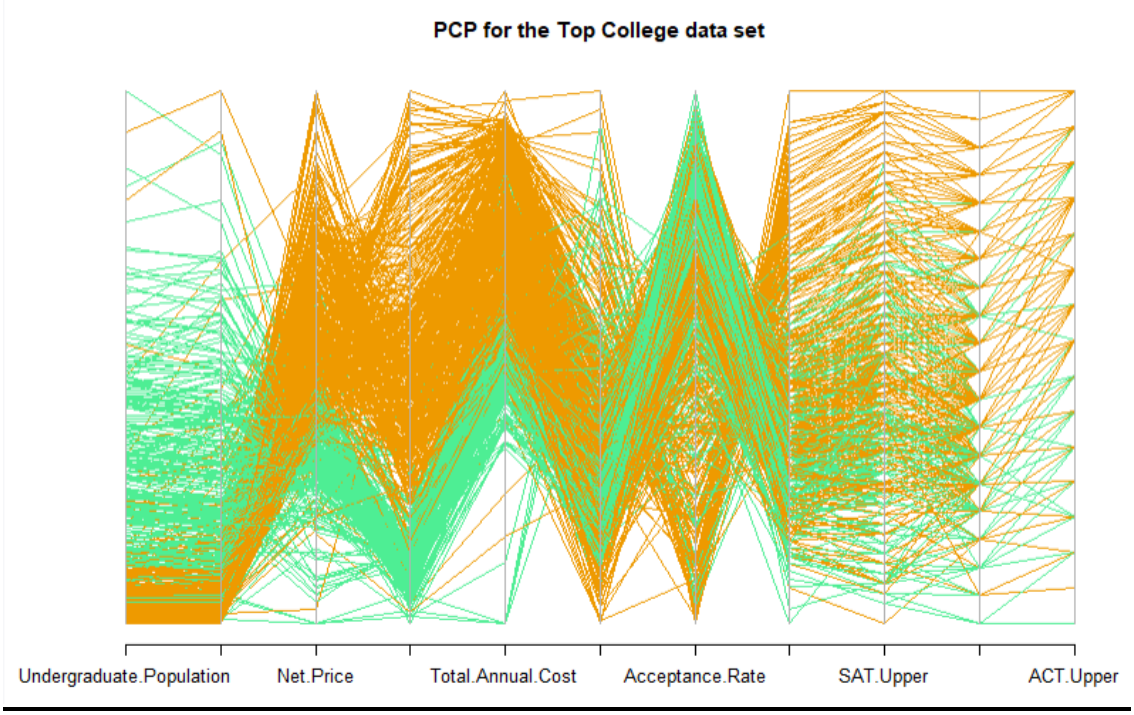
Then we would like to use the supervised classification with the Buyes rule method, to use it we will have to estimate the conditional probabilities in some way. There are different ways to estimate conditional probabilities which lead to different classifiers.

These ways are:

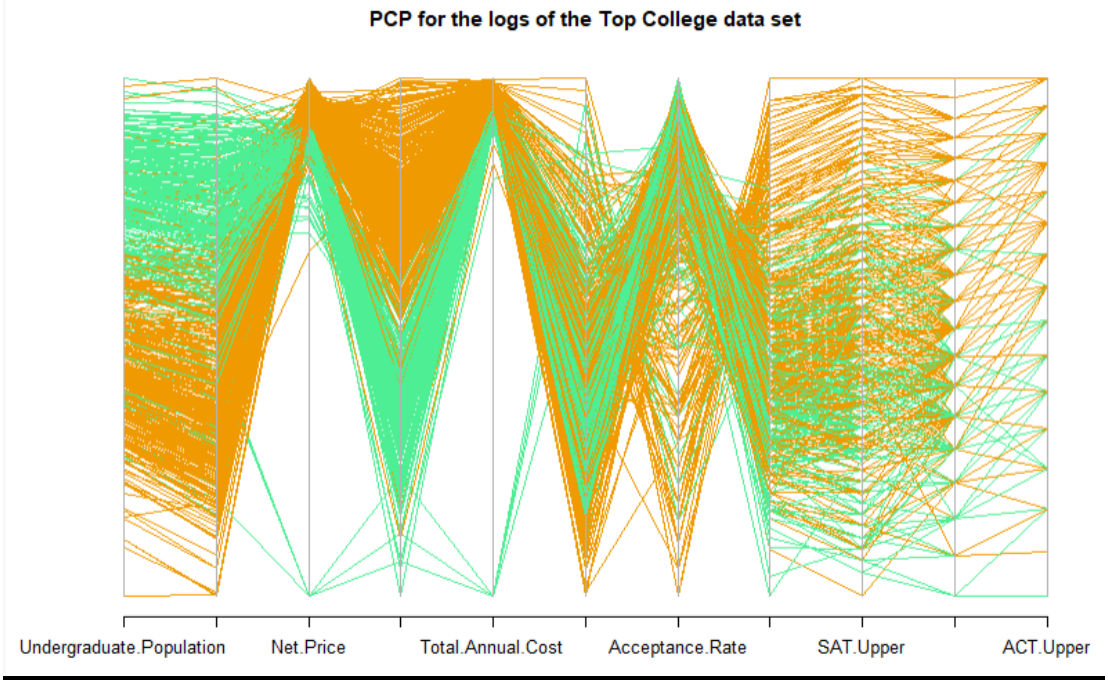
- k-nearest neighbors (kNN)
- Methods based on the Bayes Theorem:
    - Linear Discriminant Analysis (LDA).
    - Quadratic Discriminant Analysis (QDA).
    - Naive Bayes (NB).
- Logistic regression (LR).

A previous step that I have done for all the methods is selecting what variables I will include in the supervised classification. One of the criteria to know what variables are uninteresting is discarding predictors which do not seem to separate the classes.

To see that we perform a PCA for the variables in our dataset



And the log PCP



We could see clearly that the last four variables can´t be distinguished between groups, we could say something similar of the variable between the Total Annual Cost and the Acceptance

Rate the Alumni salary variable but it is no so clear in this case, so we are going to do another test for checking if we should include that variable.

Another possibility for checking which variables I should include is performing a supervised classification with the single variables and discard those with large TER. I obtained the following TER for each variable.

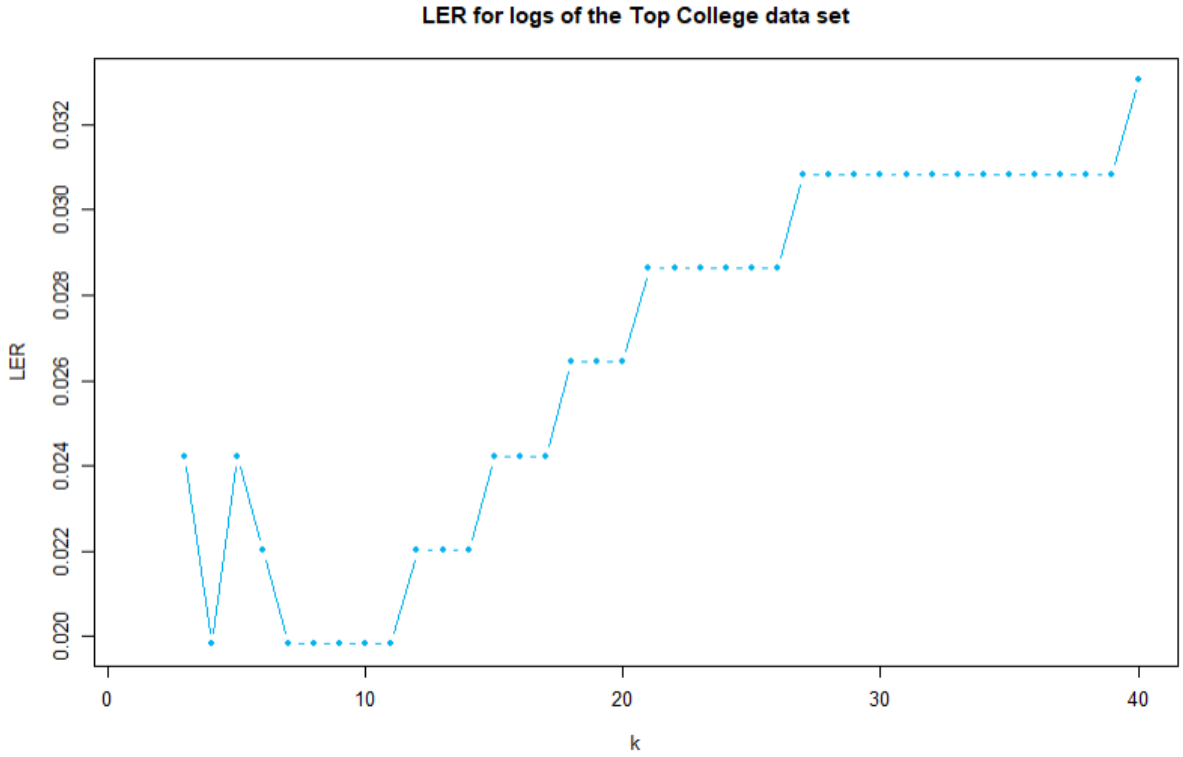
Acceptance Rate → 0.3214
Alumni Salary → 0.4847
Total Annual Cost → 0.1531
Average Grant Aid → 0.04592
Net Price → 0.1531
Student Population → 0.2194
Undergraduate Population → 0.1378

I discarded the variable Acceptance Rate and Alumni salary because they are the ones with the greatest TER. I finally performed the supervised classification with the variables Total Annual Cost, Average Grant Aid, Net Price, Student Population and Undergraduate Population.

1. K Nearest Neighbors (KNN)

This method classifies a new element represented by a vector to the most common class among its k nearest neighbors.

Firstly, we have to select one value for the k, to know which is the optimal value we use the leave-one-out cross validation (LOOCV) with the different observations in the training sample. LOOCV allow us to classify all the observations for the training sample using the algorithm KNN for the differents values of k from 1 to a certain value which in this case is 40.



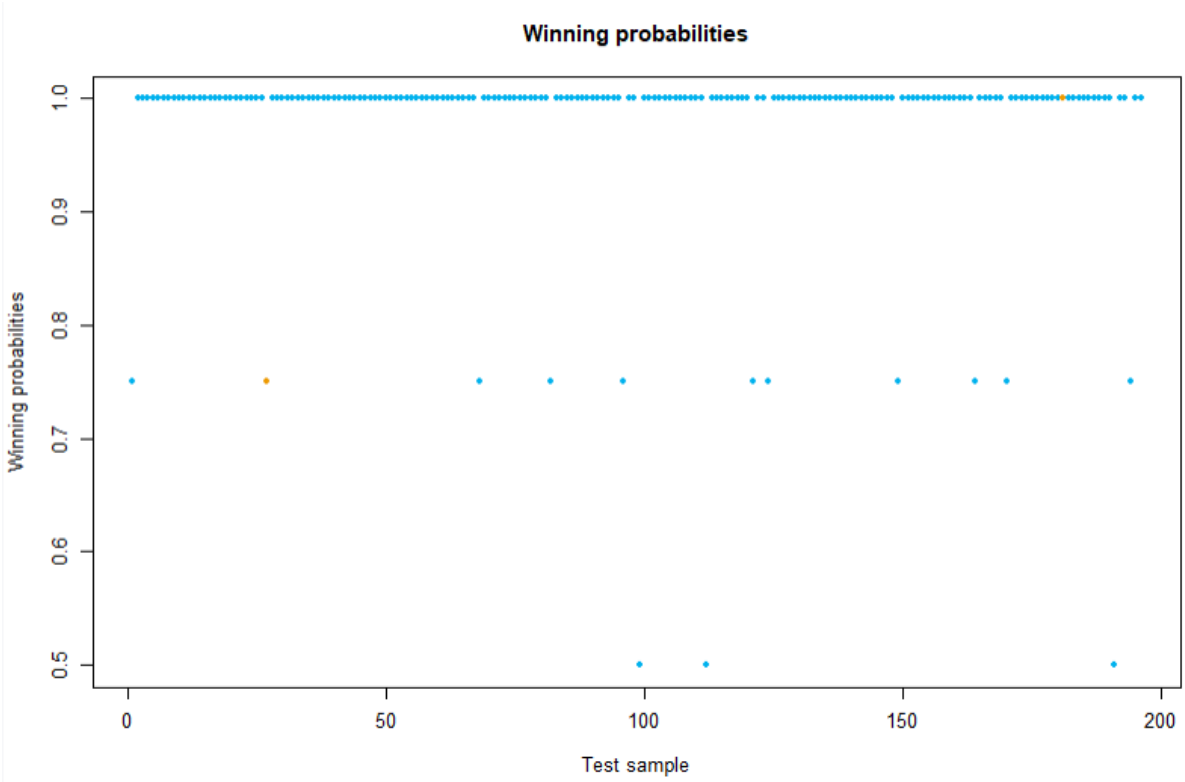
In this graph we could see that there are different values for k in which we obtain the minimum value for LOOCV, these values are 4,7,8,9,10 and 11. We will choose the value 4 because it gives us the same performance with a minor quantity of neighbours.

The next step is training the model with the value 4 for k, after training and testing it with their corresponding samples.

We can see the results of our prediction in this confusion matrix, it shows that for the class Private 126 observations were classified as Private correctly and 0 as Public incorrectly and for the Public class that 66 were classified correctly and 2 incorrectly.

Y_test/knn_Y_test	Private	Public
Private	126	2
Public	0	66

The obtained TER was 0.0102, therefore this classifier has a good performance. We can see in the following graph the probabilities of the winning group for the test sample, we can see the classes predicted correctly with the blue color and the classes incorrectly predicted with the orange color. We can clearly see how almost the total of the classes have been predicted correctly.



2. The Bayes Theorem

In this theorem K is the probability of a randomly chosen observation of the variable x como from the class k. There are different assumptions on the densities which lead to different methods based on the Bayes Theorem.

Another thing to consider is that here unlike the KNN algorithm we won´t discard any variable so we will use the 11 variables and not 5 variables like in KNN,

A. Linear Discriminant Analysis (LDA)

In this method we assume that there are multivariate Gaussian with a common covariance matrix and different mean vectors.

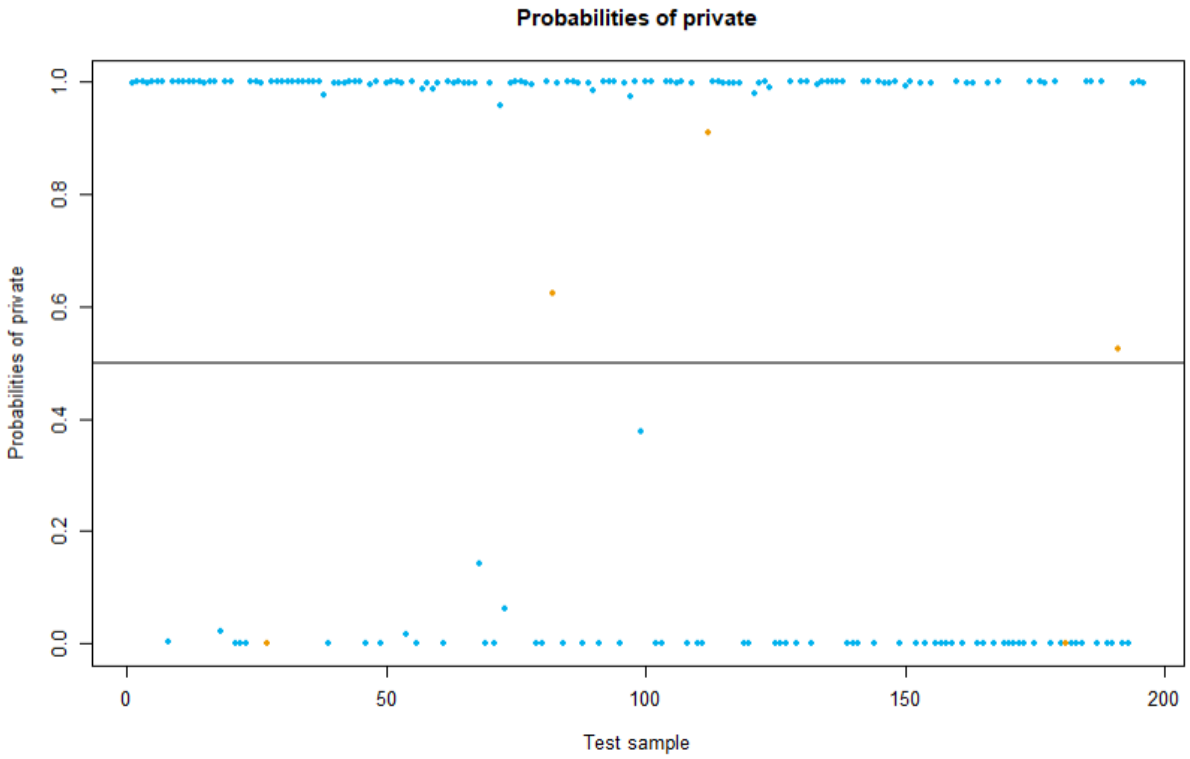
We trained a model with the LDA method and we obtained the following confusion matrix.

Y_test/knn_Y_test	Private	Public
Private	126	2
Public	3	65

The obtained TER was 0.02551, therefore this classifier has a good performance but a worse performance than the KNN method which has a lower TER.

This is normal because our data is not appropriate for LDA (and all the Bayes Theorem methods) because the variables are non-Gaussian, the only variable that have
a behavior that can be described as Gaussian is the Net Price, the other variable doesn´t present a Gaussian distribution. Despite this the obtained TER although higher than KNN is pretty good for an LDA.

We can see in this graph how there are so many successes compared to the mistakes predicting the classes.



B. Quadratic Discriminant Analysis (QDA)

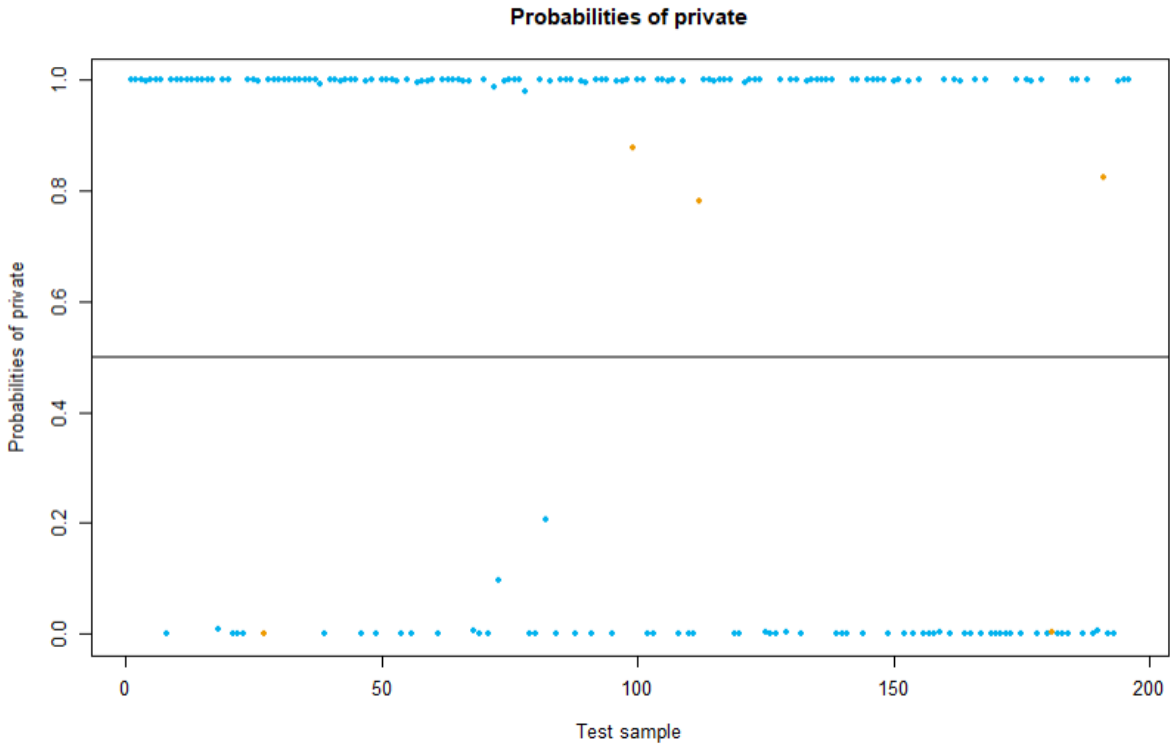
In this method we assume that there are multivariate Gaussian with mean vectors and covariance matrix. We trained a model with the QDA method and we obtained the following confusion matrix.

Y_test/knn_Y_test	Private	Public
Private	126	2
Public	3	65

The obtained TER was 0.02551, therefore this classifier has a good performance but a worse performance than the KNN method which has a lower TER and the same performance as the LDA method.

In this method, like in the LDA method we obtain a good performance despite the non-gaussianity of most of the variables.

We can see in this graph that the obtained mistakes are the same that we can see in the LDA method.



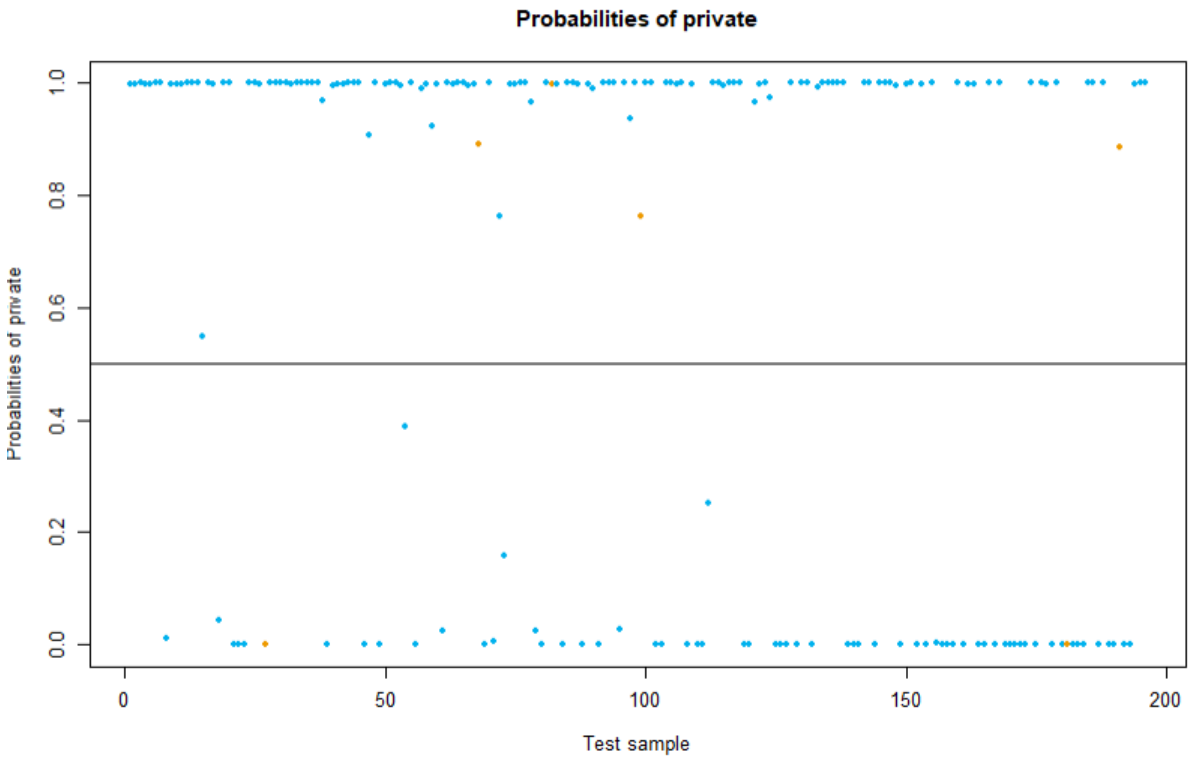
C. Naive Bayes (NB)

We assume that the predictors for this method are independent variables. We trained a model with the NB method and we obtained the following confusion matrix.

Y_test/knn_Y_test	Private	Public
Private	126	2
Public	4	64

The obtained TER was 0.03061, we can see that the obtained TER is much larger than the obtained for the LDA and the QDA method, this is because of the effect of the dependency between the variables and their non-Gaussianity.

We can see in this graph how the number of mistakes is greater than the previous methods.



3. LOGISTIC REGRESSION (LR)

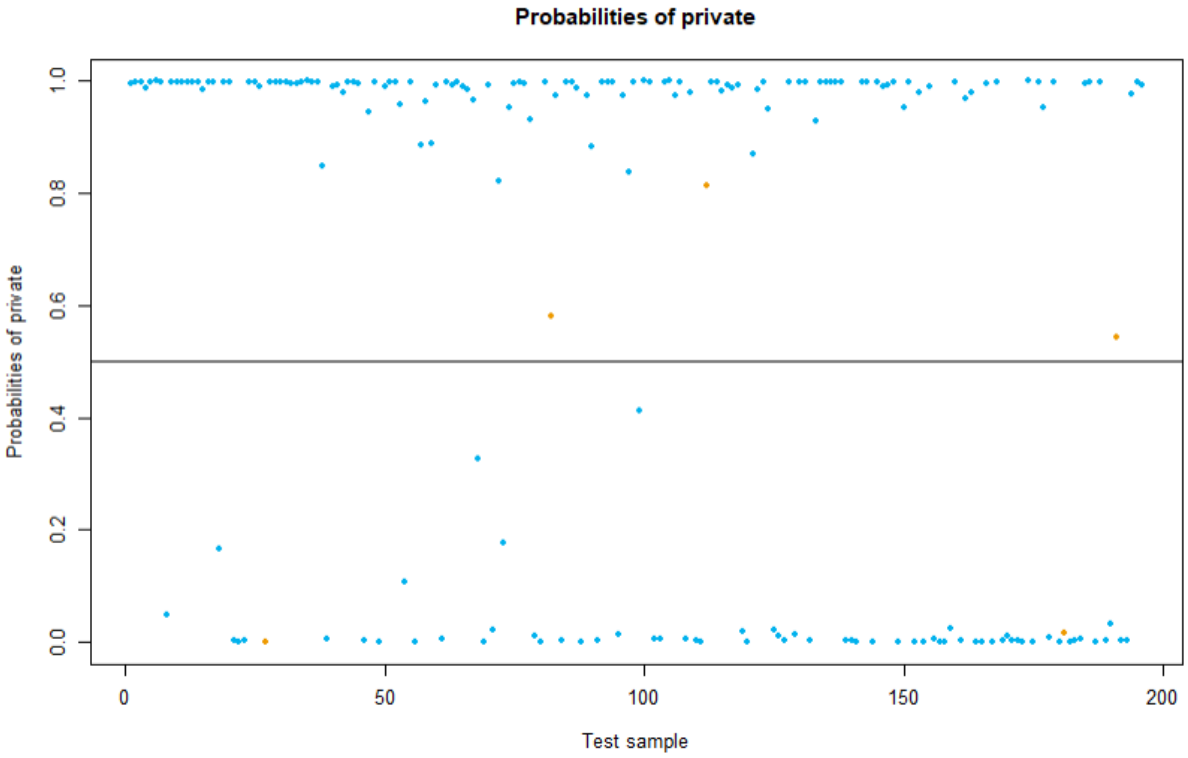
In this method we use a regression model to predict the value of y. This method predicts a dependent variable by analyzing the relationship between one or more existing independent variables.

We trained a model with the Logistic regression model and we obtained the following confusion matrix.

Y_test/knn_Y_test	Private	Public
Private	126	2
Public	3	64

As we could conclude seeing the confusion matrix the obtained TER was 0.02551, therefore this classifier has a good performance but a worse performance than the KNN method. It also has the same performance as the LDA and QDA methods based on the Bayes Theorem.

We could see in this graph that we have five mistakes, the same amount that we could find in the LDA and QDA methods.



4. CONCLUSIONS

We can conclude that all the methods allow us to predict between the classes private and public in a precise way in our dataset. Despite this, not all the methods obtain the same performance. The Naive Bayes method is the one which obtains the worst performance due to the dependency of the variables and their non gaussianity. After that we have the LDA, QDA and LR methods which obtains the same performance. Finally we have the KNN method which is the combination with the best performance among the examined ones.