US Colleges Analysis Report

Detailed explanation on GitHub: https://github.com/ajeyvishnu/US_Colleges_Analysis

Dataset:

- The dataset is obtained from Kaggle: https://www.kaggle.com/datasets/yashgpt/us-college-data
- It has 777 colleges as rows and 18 columns defining various aspects of each college.

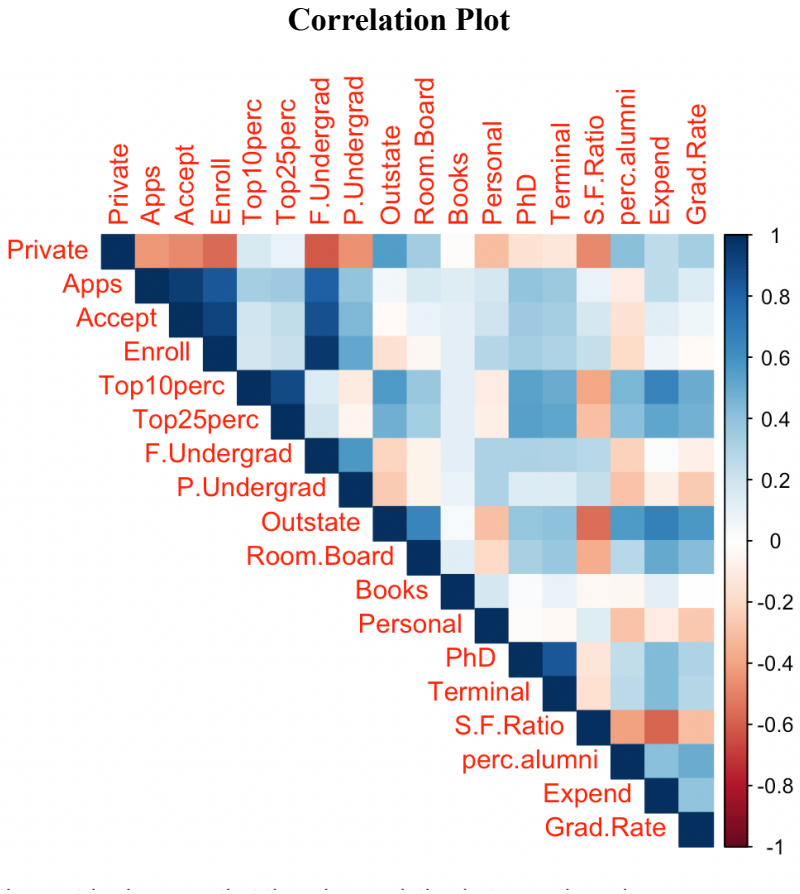
Data Dictionary:

- Group: A categorical variable indicating whether the college is public or private. 1 indicates a private college and 0 indicates a public college.
- Private: A binary variable indicating whether the college is private or not. 1 indicates a private college and 0 indicates a public college.
- Apps: The total number of applications received by the college.
- Accept: The total number of applications accepted by the college.
- Enroll: The total number of students enrolled in the college.
- Top10perc: The percentage of new students who ranked in the top 10% of their high school class.
- Top25perc: The percentage of new students who ranked in the top 25% of their high school class.
- F.Undergrad: The total number of full-time undergraduate students enrolled in the college.
- P.Undergrad: The total number of part-time undergraduate students enrolled in the college.
- Outstate: The out-of-state tuition fee for the college.
- Room.Board: The cost of room and board for the college.
- Books: The estimated cost of books and supplies for a year of study.
- Personal: The estimated personal expenses for a year of study.
- PhD: The percentage of faculty members who hold a PhD degree.
- Terminal: The percentage of faculty members who hold a terminal degree in their field of study.
- S.F.Ratio: The student-to-faculty ratio for the college.
- perc.alumni: The percentage of alumni who donate to the college.
- Expend: The instructional expenditure per student at the college.
- Grad.Rate: The graduation rate of the college as a percentage.

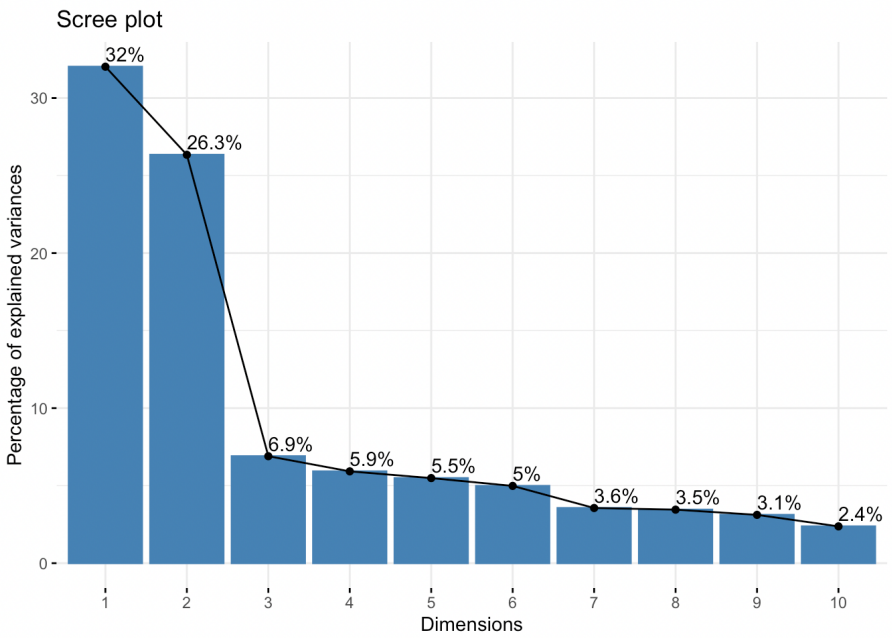
QUESTIONS:

1. **Based on the given variables, can we classify the colleges based on their type (Public/Private)?**
2. **Based on the given variables, can we predict the type of college?**

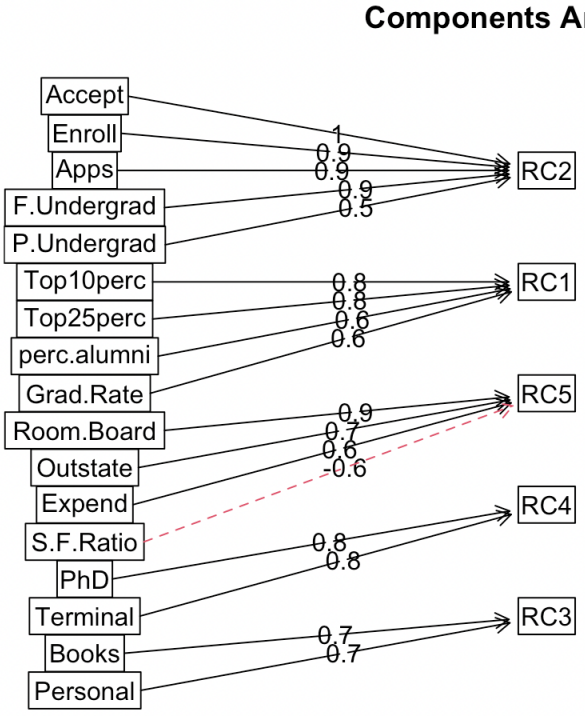
Q1 SOLUTION:



- The correlation plot shows correlation, so we start with Principal Component Analysis (PCA) to reduce the number of variables.



- The scree diagram shows us that the sum of the first two principal components is less than 70%.
- So, we cannot move forward using PCA for column reduction.
- We now move on to check Exploratory Factor Analysis (EFA).



Defining the factors obtained

RC2 - Student numbers

- RC2 pertains the Applications, Acceptances, Enrollments, Number of full time and part time undergraduate students.
- This summarises that RC2 is taking the number of students count in a broader picture.

RC1 - Student merit

- RC1 summarize the percentage of students who ranked 10%, 25% in their high school, the Graduation Rate of the college.
- This shows that this factor is defined by the merit of the student.

RC5 - Educational/College Expenses

- RC5 caters to Outstate tuition expenses, Room and Board expenses, and the instructional expenses per each student.
- So, this can be defined as the educational expenses incurred by student.

RC4 - Faculty qualification

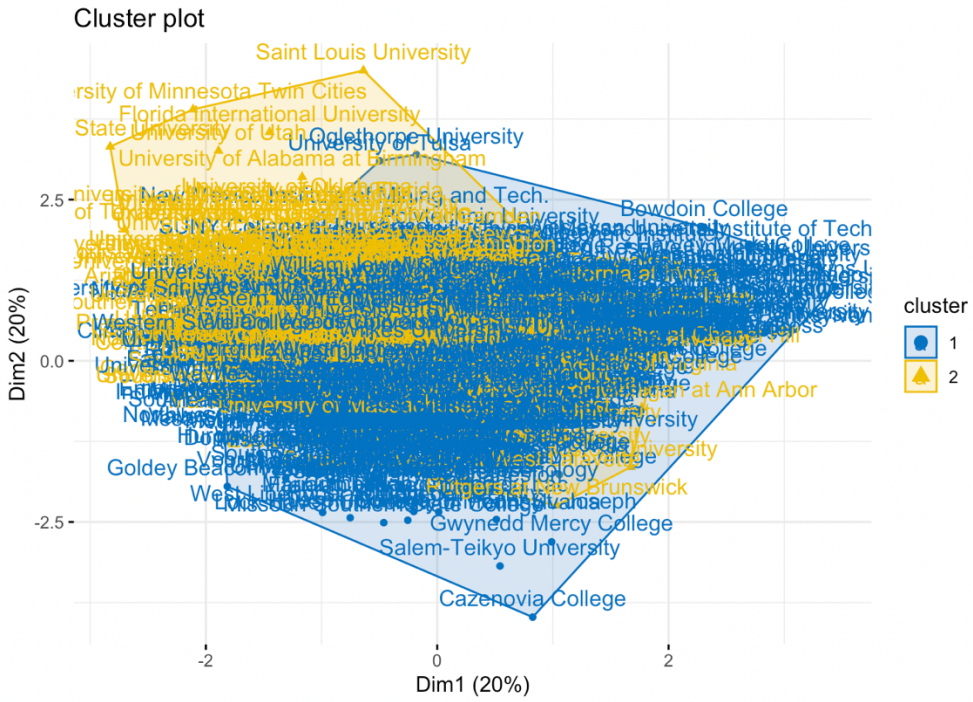
- RC4 sums up the educational qualification of the faculty at the colleges.
- It defines if the faculty hold a PhD or a terminal degree.

RC3 - Student Expenses

- RC3 gathers the details of the expenses incurred for books and personal expenses of the students.
- It summarises the personal expenses the student spends on their development.

Clustering

- Moving ahead, we use the factors obtained for EFA for clustering.



- As there are 777 colleges, the cluster plot looks very messy.
- We can try to generate a confusion matrix for the above analysis and see the results more clearly.

```
##           Actual
## Clustered Private Public
##   Private     550     71
##   Public       15    141
```

- We see that the clustering has classified the colleges with 88.9% accuracy.
- We can conclude that we can classify the type of college based on the data provided.

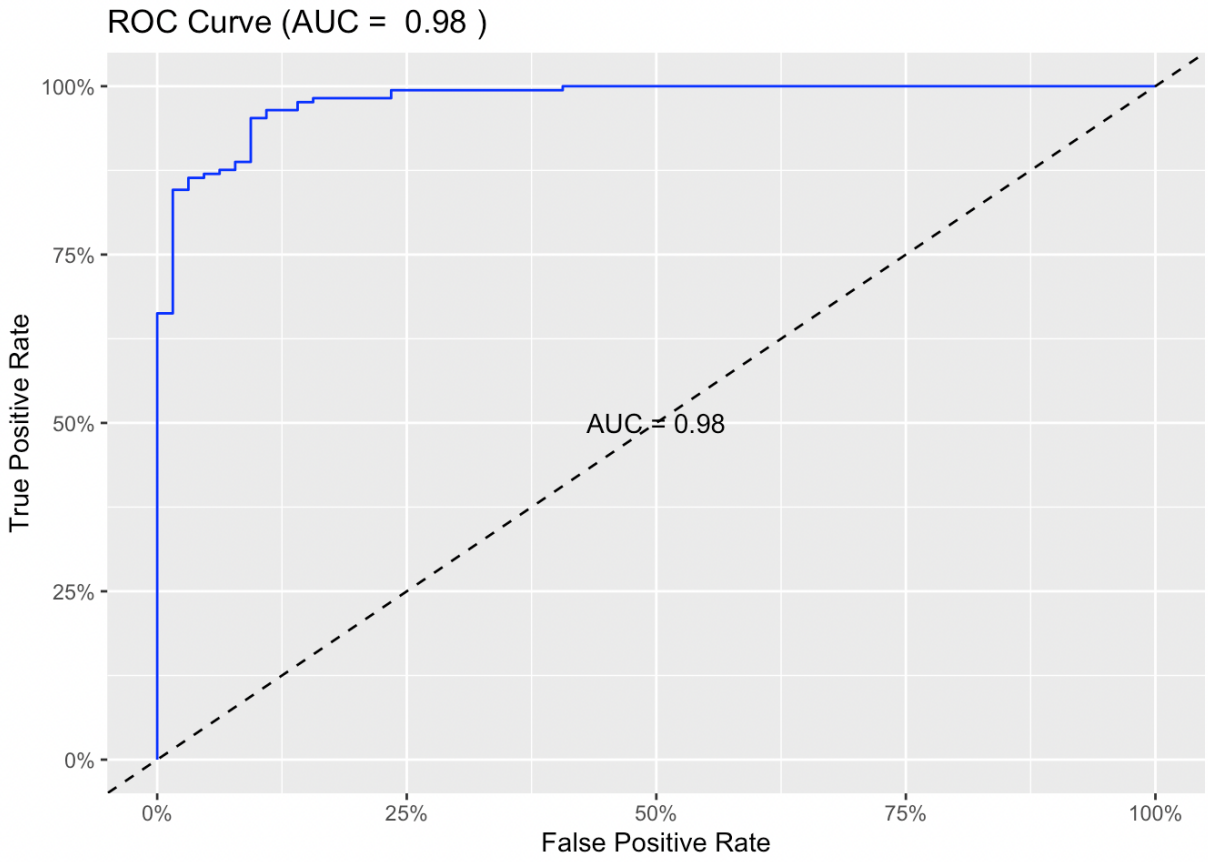
Q2 SOLUTION:

- We define training and testing sets to perform logistic regression as the output variable is qualitative and has two factors (Public/Private)

```
##
## Call:
## glm(formula = Ytrain ~ ., family = "binomial", data = x)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7422  -0.0422   0.0462   0.1673   3.1599
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.5676725  2.1562290  -0.263   0.7923
## Apps        -0.0007502  0.0003174  -2.363   0.0181 *
## Accept       0.0002637  0.0005422   0.486   0.6267
## Enroll       0.0012106  0.0008669   1.396   0.1626
## Top10perc    0.0164093  0.0317300   0.517   0.6050
## Top25perc    0.0092338  0.0213646   0.432   0.6656
## F.Undergrad -0.0002595  0.0001415  -1.835   0.0666 .
## P.Undergrad -0.0001175  0.0001842  -0.638   0.5237
## Outstate     0.0005294  0.0001204   4.399 1.09e-05 ***
## Room.Board   0.0005943  0.0003199   1.858   0.0632 .
## Books        0.0028475  0.0016885   1.686   0.0917 .
## Personal    -0.0003295  0.0003204  -1.028   0.3038
## PhD         -0.0742546  0.0332388  -2.234   0.0255 *
## Terminal    -0.0347455  0.0295821  -1.175   0.2402
## S.F.Ratio   -0.1001612  0.0646275  -1.550   0.1212
## perc.alumni  0.0601375  0.0251824   2.388   0.0169 *
## Expend       0.0002451  0.0001394   1.759   0.0786 .
## Grad.Rate    0.0192999  0.0129682   1.488   0.1367
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 636.8  on 543  degrees of freedom
## Residual deviance: 175.3  on 526  degrees of freedom
## AIC: 211.3
##
## Number of Fisher Scoring iterations: 8
```

```
##           actual
## predicted  No Yes
##       No   55   5
##       Yes   9 164
```

- The regression gives a confusion matrix with an accuracy of 94% and a precision of 97%.



- The AUC is obtained to be 97.98% which is excellent and tells us that our prediction works well.

CONCLUSION:

- **Based on the given data, we could classify the type of college (Public or Private)**
    - The accuracy of the classification came out to be 88.9%
    - We have used Exploratory Factor Analysis and Clustering for this classification.
- **We could also predict the type of college (Public or Private) based on the variables provided.**
    - Using logistic regression, we predicted the type of college with 94% accuracy and 97% precision.
    - An AUC of 98% for the ROC curve shows good prediction.