Sree Bhanu Nori
Anastasie Mvula Nsimba
Grace Shin
Anusha Eregowda
STAT 515
Due: 05/13/2019

Final Project: Graduate School Admissions Data

Introduction:

The dataset is from Kaggle.com and was originally for students from India who are applying to Masters degree programs in the United States in order to predict admissions chances using several different variables. The dataset should give an insight of the factors of admissions that are important in graduate school admissions. The data collected has the following decision variables:

- GRE Score: range from 290 to 340
- TOEFL Score: range from 92 to 120
- University Rating: ranks from 1 to 5
-Statement of Purpose (SOP): ranks from 1 to 5
- Letter of Recommendation Strength (LOR): ranks from 1 to 5
- Undergraduate GPA (CGPA): ranks from 6.8 to 9.92
- Research Experience: Binary data, either 0 or 1
- Chance of Admit: value ranking from 0.34 to 0.97

Before applying different models to the dataset, there were initial steps that were taken to understand the data. These steps were to plot the data and to get the data summary, shown in Appendix A.Figure 1 shows the scatter plots for the AdmitChance versus different variables such as CGPA, GRE, TOEFL, and LOR. Figures 2 and 3 are the histograms of different variables to find the number of times the variables occurs.

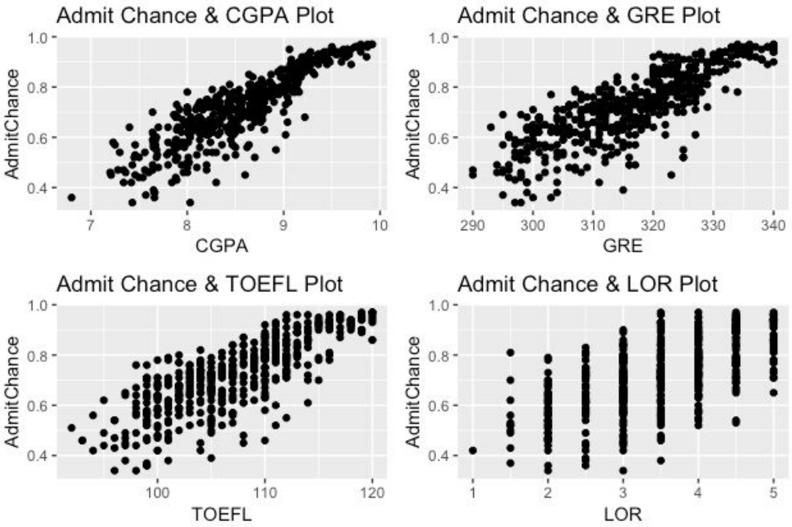


Figure 1: Scatterplot of the CGPA, GRE, TOEFL, and LOR versus AdmitChance

The scatter plots show a linear relationship between CGPA, GRE, and TOEFL. LOR, or letter of recommendation, has the points stacked. The majority of the points are between 3 and 4 for LOR.

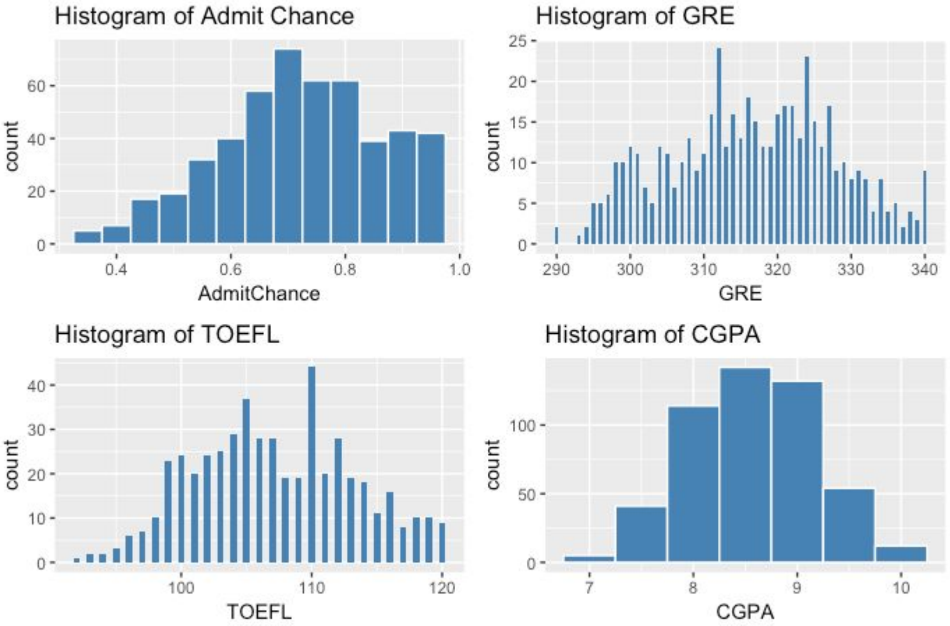


Figure 2: Histogram of AdmitChance, GRE, TOEFL, and CGPA

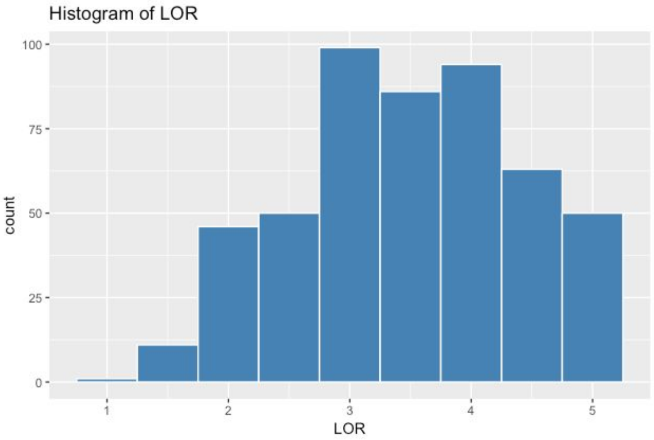


Figure 3: Histogram of LOR

AdmitChance has the most counts between the values of 0.6 to 0.8, GRE has the highest counts at approximately 310 and between 320 and 330, TOEFL has the highest count at 110, CGPA between 8 and 9, and LOR, letter of recommendation, between 3 and 4.

Hypothesis:
What are the factors that can affect a student's admission?

In order to answer the hypothesis, we will be using the Cross-Validation method to evaluate the metrics. The dataset will be split into Train and Test datasets.

Afterwards, the following modeling algorithms will be applied to the dataset and a Training dataset will be created:
- Correlation Plots
- Linear Regression
- Random Forest
- Principal Component Analysis
- K means clustering

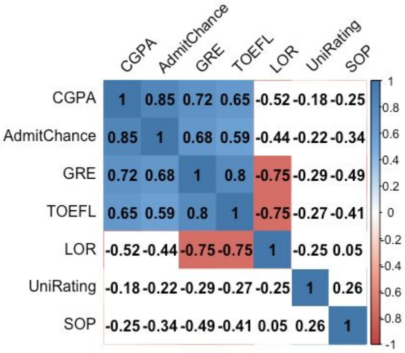
Correlation Matrix:

The goal of the matrix is to determine the dependency between multiple variables at the same time. In this case we chose the correlation matrix to help us answer the following questions.

1. Is there a relationship between CGPA and admission to university.
2.Which aspects enhances the chance of admission?
Do all variables in the dataset contribute to a positive decision on admission, or do just one or two of the variables contribute?
To answer this question, we will visualize the data first and then we will need to separate out the individual effects of each variables.

Here we decided to plot a correlogram, which appears to be more suitable for our questions. The Correlogram gives a visual representation of the relationship of one variable vector and all the data in the dataset, shown below in Figure 4.

Admissions Criteria: According strength of correlation



	CGPA	AdmitChance	GRE	TOEFL	LOR	UniRating	SOP
CGPA	1	0.85	0.72	0.65	-0.52	-0.18	-0.25
AdmitChance	0.85	1	0.68	0.59	-0.44	-0.22	-0.34
GRE	0.72	0.68	1	0.8	-0.75	-0.29	-0.49
TOEFL	0.65	0.59	0.8	1	-0.75	-0.27	-0.41
LOR	-0.52	-0.44	-0.75	-0.75	1	-0.25	0.05
UniRating	-0.18	-0.22	-0.29	-0.27	-0.25	1	0.26
SOP	-0.25	-0.34	-0.49	-0.41	0.05	0.26	1

Figure 4: Admissions Criteria Correlation Strength

The variables CGPA, GRE, and TOEFL indicate a strong positive correlation. It has a weaker correlation with University Rating, SOP and LOR. The Research has the least correlation. which means that y increases with x. Since correlation does not mean causation, we decided to further investigate the findings.

Fit of the Model: Model selection approach

Once we fit the model, we identified the significant variables in the decision process of admission. GRE, LOR, CGPA, Research experience and TOEFL were statistically significant at a full model, since the p value showed the number below 0.05.

We chose to use the subset regression over stepwise, because it displays all possible models based on the adjusted r squared.

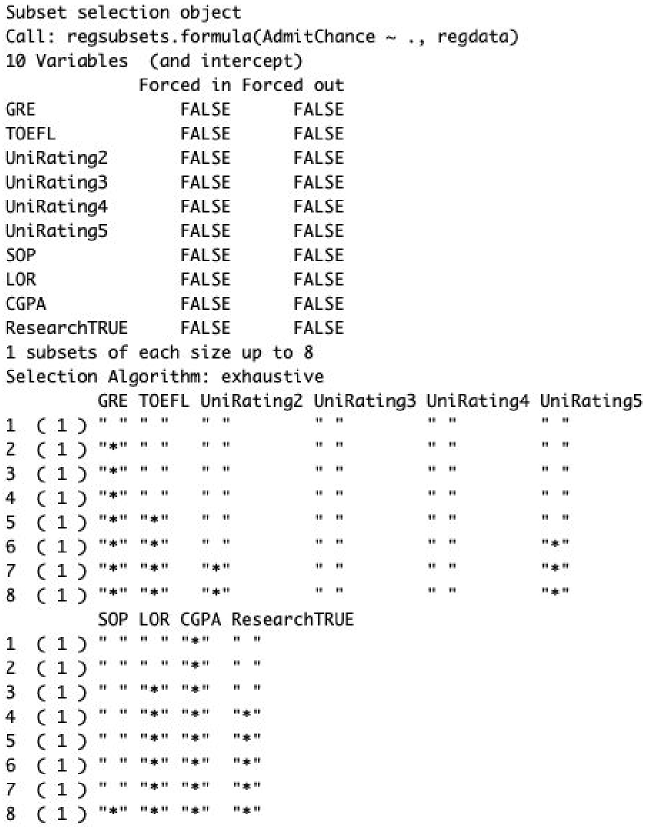
We removed SOP in the model, since it had a p value of 0.73 and did not appear as a reliable predictor. Hence, we chose to fit the model using all predictors except SOP.

We remove the ones with the smallest coefficient, since they would not have much of an impact on the dependent variable.

The Model selection approach consists of testing all possible combination of the predictor variables, and then it enables us with the help of statistical criteria to select the best model.

In our scenario, we have 8 predictor variables in the data.

The best model contains a number of predictors, which will be best to use the RSS on, shown in Figure 5



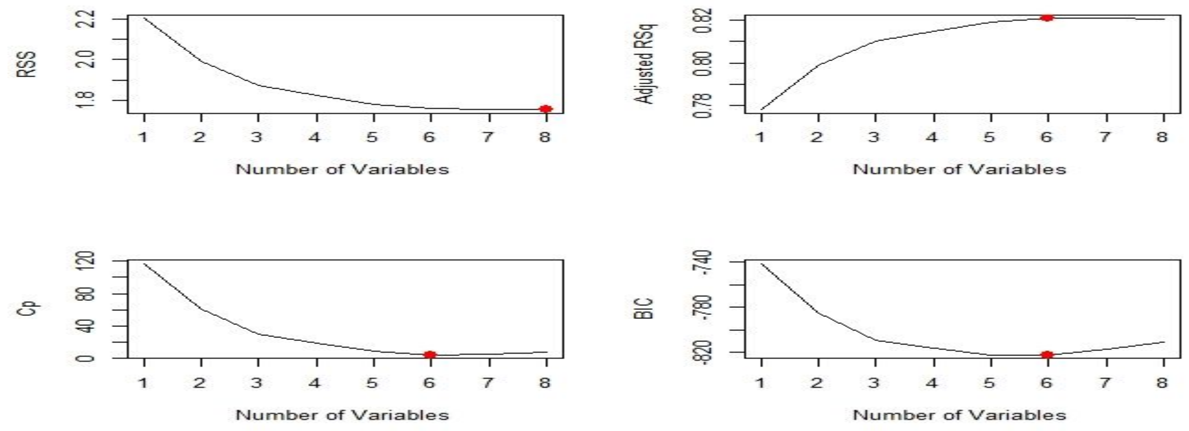
```
Subset selection object
Call: regsubsets.formula(AdmitChance ~ ., regdata)
10 Variables  (and intercept)
             Forced in Forced out
GRE              FALSE      FALSE
TOEFL            FALSE      FALSE
UniRating2       FALSE      FALSE
UniRating3       FALSE      FALSE
UniRating4       FALSE      FALSE
UniRating5       FALSE      FALSE
SOP              FALSE      FALSE
LOR              FALSE      FALSE
CGPA             FALSE      FALSE
ResearchTRUE     FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         GRE TOEFL UniRating2 UniRating3 UniRating4 UniRating5
1  ( 1 ) " " " "   " "        " "        " "        " "       
2  ( 1 ) "*" " "   " "        " "        " "        " "       
3  ( 1 ) "*" " "   " "        " "        " "        " "       
4  ( 1 ) "*" " "   " "        " "        " "        " "       
5  ( 1 ) "*" "*"   " "        " "        " "        " "       
6  ( 1 ) "*" "*"   " "        " "        " "        "*"       
7  ( 1 ) "*" "*"   "*"        " "        " "        "*"       
8  ( 1 ) "*" "*"   "*"        " "        " "        "*"       
         SOP LOR CGPA ResearchTRUE
1  ( 1 ) " " " " "*"  " "         
2  ( 1 ) " " " " "*"  " "         
3  ( 1 ) " " "*" "*"  " "         
4  ( 1 ) " " "*" "*"  "*"         
5  ( 1 ) " " "*" "*"  "*"         
6  ( 1 ) " " "*" "*"  "*"         
7  ( 1 ) " " "*" "*"  "*"         
8  ( 1 ) "*" "*" "*"  "*"         
```

Figure 5: RSS for Best Model for Predictors

The output shows that GRE and CGPA are the best two variable models followed by LOR and Research. In general, the asterisk indicates that one variable is included in the corresponding model.

Since, we will not know which model to pick, we applied some statistical criteria. The goal is to estimate the prediction error of each model and select the one with the lower prediction error.

In order to pick the best model, we decided to plot Residual Sum of Squares, adjusted R2, Cp, and BIC.
This visualization will give us the best tool to compare the different models.
However, these metrics could be affected by overfitting on the predictive analytics.
And the mean squared error is 0.0036, which can be considered as a base case.



Figure#6: Variable errors

So, we will need a further step by using the K Fold cross validation**.**

K Fold cross validation:

Here we need to divide the data into k subsets, where the small part is the test data set and the bigger portion of the data, the remaining subset for the training data. We chose to divide the data into 7 subsets. We can observe after the validation that 4 to 7 several variable scenarios are stable at 0.0035.

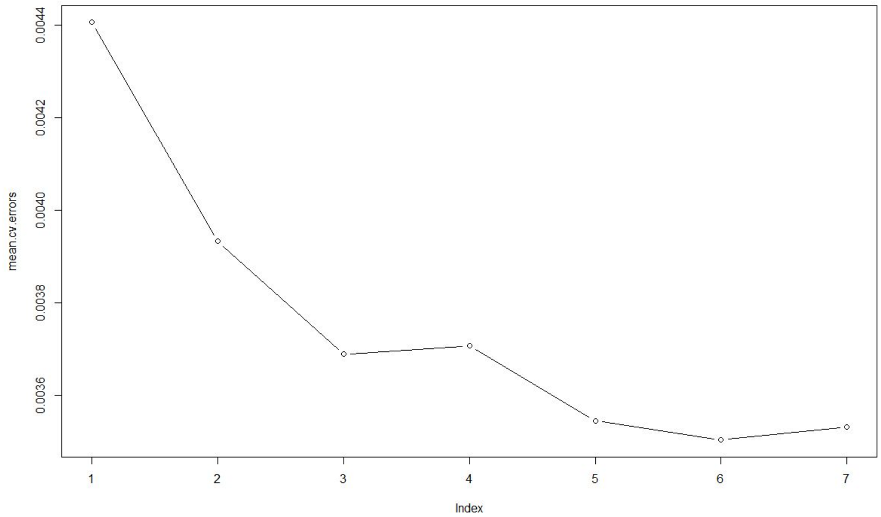


Figure 7: Cross Validation Errors

Choosing a Model: Validation Set

The validation set approach generates accurate estimates of the test error. Therefore, it is imperative to only use the training observations to select the variables and to fit the model.
We implemented the validation test approach by splitting the observations into a training- and test set. In order to obtain the same value every time we split the training- and test set, we set a seed.
We calculate the validation set error for each number of variables.
Then we run it in a loop for each number of variables in the model. So, we can obtain the coefficients of the training set and then we compute the MSE.
The validation errors are stable at 0.0035 for 5 out 7. Now we can use the best subset model on the full data set.

Choosing a Model: Cross Validation Approach

In order to make predictions for each model we need to calculate the test errors on the subset and then store it in the matrix.
In this scenario, we need the loop to perform the cross validation.
The Folds element that equal j are in the test set and the remaining are in the training set.
Then we find the mean across the fold MSE for each model. The best case is 0.0035 for 5 variable models, not too different from the base case i.e, 0.0036.

Random Forest:

We considered the training data and perform random forest analysis on it with importance plot.

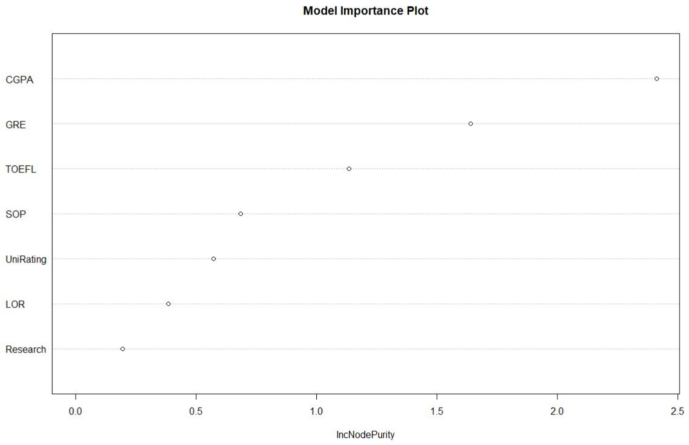


Figure 8: Importance plot

According to the plot we can observe that CGPA followed by the GRE score seem to be important variables in order to get the admit when compared to rest of the variables.
The error we obtained with 500 number of trees is,

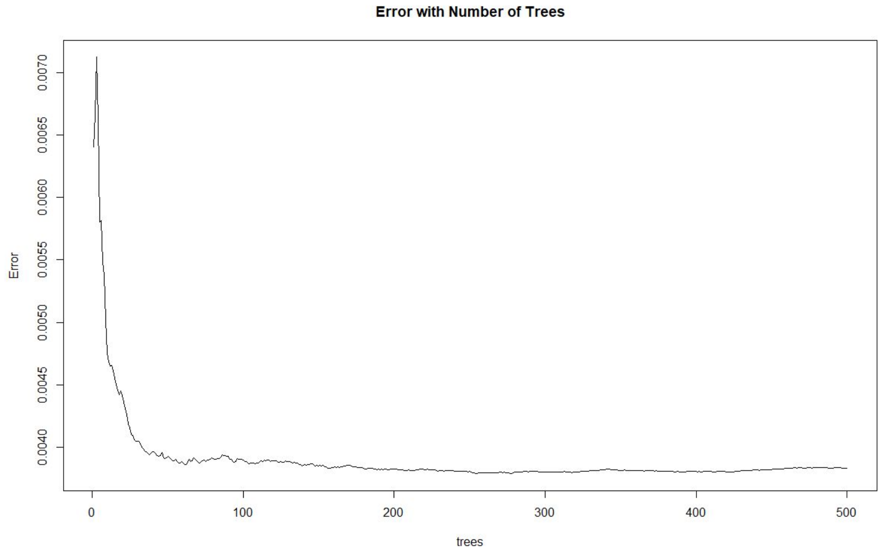


Figure 9:Errors with number of trees.

Next, we can calculate the RMSE.

Rmse with all variables is 0.06331481.

Rmse with a few variables is 0.07294912.

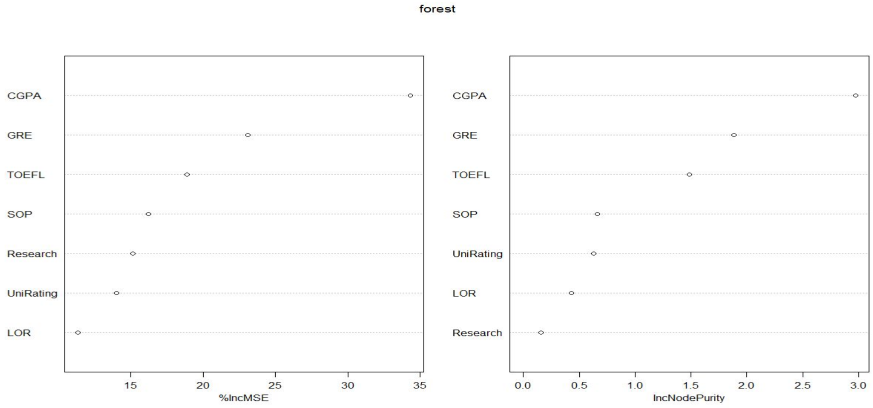
Rmse with all the variables except for research variable is 0.06458465.

Hence, we can know that based on the RMSE value, the model with all the predictors excluding research is performed best.

Random Forest with CV:

In random forest with CV, we used the 10 fold method. By grabbing all the rows with the id 'n', aggregate them into a test set. All the other rows (the other 9 parts) go in the training set. We run the model on the test set, save the prediction in a dataframe, the test set in another.

Then we compared them to get our performance measure. When we compare, we almost see the same results.



Principal Component Analysis:
Principal component Analysis is one of the useful technique for analysis of data with large attribute set. It pictures a better visualization for the variation present in the dataset. It is specifically helpful for wide datasets, where each row of data has many variables . It is also called as Dimension Reduction tool or Compressor.

No of variables represent the number of dimensions in which the data is projected. By applying the principal component Analysis algorithm on the wide dataset, it compresses the original variables into a smaller number of principal components.

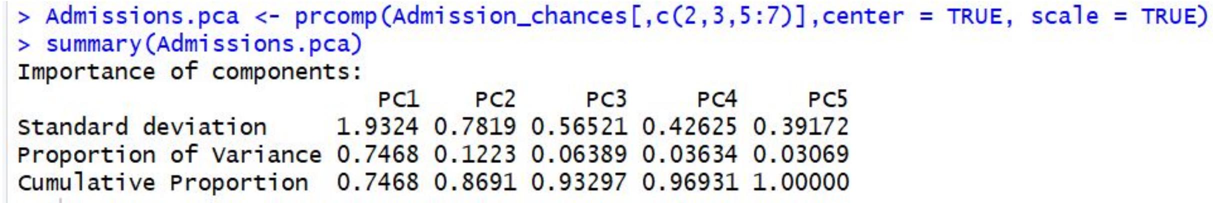
This may bring down to 2 or 3 dimensional graph that explains the maximum variance in data. The first Principal component (PC1) showcases the maximum variance in the data followed by the second Principal component (PC2). This means that the PC1 at the direction where there is maximum spread out of data. The other Principal components does not explain much variance.

Each principal component explains certain percentages of the total variance in the dataset.

(i) Computing the Principal Components
We have 8 variables in total.

Since PCA works best with numerical data, we will exclude the binary data(Research) , University rating(not much aid) , and Serial No(represents the row no). So now we have 5 attributes with 500 rows of data.

We pass this dataset to procomp() function , which is a principal component functionality and assign to a object. We can view this PCA object with summary().



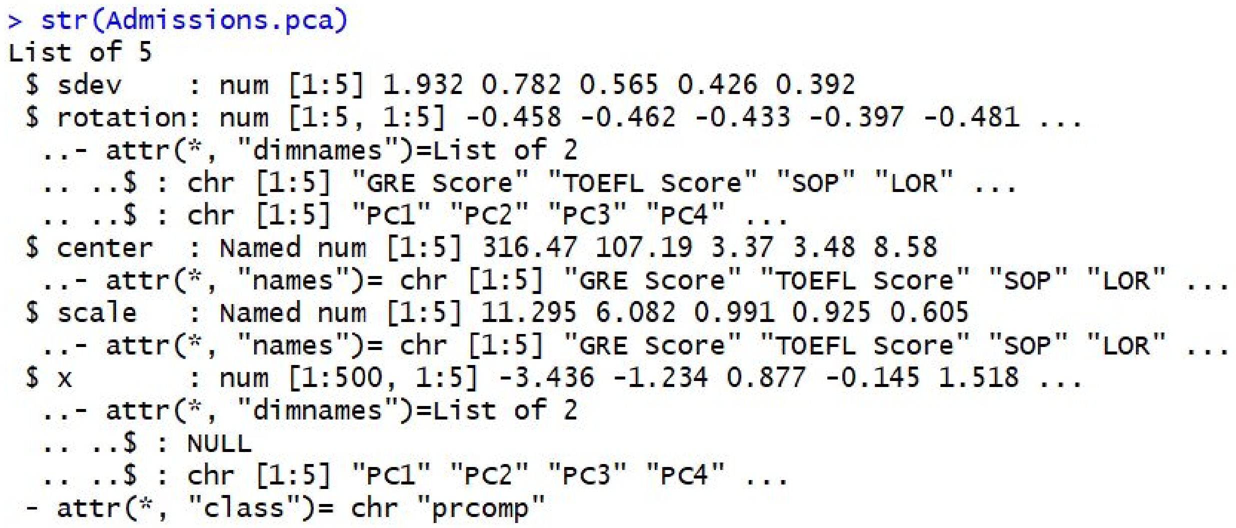
```
> Admissions.pca <- prcomp(Admission_chances[,c(2,3,5:7)],center = TRUE, scale = TRUE)
> summary(Admissions.pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     1.9324 0.7819 0.56521 0.42625 0.39172
Proportion of Variance 0.7468 0.1223 0.06389 0.03634 0.03069
Cumulative Proportion  0.7468 0.8691 0.93297 0.96931 1.00000
```

Figure 10: PCA

We get 5 principal components from the result which are called as PC1, PC2, PC3, PC4, PC5. The summary tells us that the PC1 explains 74.6% of the total variance in data which is the maximum, PC2 explains 12.2% of the variance, PC3 explains 6.3% of the variance , PC4 explains 3.6% of the variance and PC5 explains 3% of the variance in data.

So, PC1 and PC2 explains 84% of the variance in data.

We can see the structure of the PCA object by calling the str()

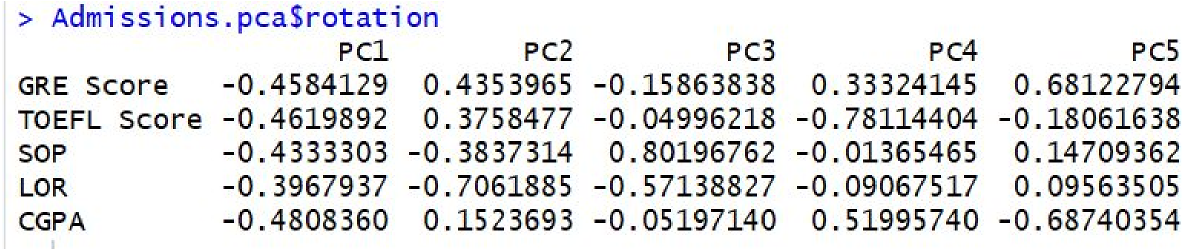


```
> str(Admissions.pca)
List of 5
 $ sdev    : num [1:5] 1.932 0.782 0.565 0.426 0.392
 $ rotation: num [1:5, 1:5] -0.458 -0.462 -0.433 -0.397 -0.481 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:5] "GRE Score" "TOEFL Score" "SOP" "LOR" ...
  .. ..$ : chr [1:5] "PC1" "PC2" "PC3" "PC4" ...
 $ center  : Named num [1:5] 316.47 107.19 3.37 3.48 8.58
  ..- attr(*, "names")= chr [1:5] "GRE Score" "TOEFL Score" "SOP" "LOR" ...
 $ scale   : Named num [1:5] 11.295 6.082 0.991 0.925 0.605
  ..- attr(*, "names")= chr [1:5] "GRE Score" "TOEFL Score" "SOP" "LOR" ...
 $ x       : num [1:500, 1:5] -3.436 -1.234 0.877 -0.145 1.518 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : NULL
  .. ..$ : chr [1:5] "PC1" "PC2" "PC3" "PC4" ...
 - attr(*, "class")= chr "prcomp"
```

Figure 11: PCA

The PCA object has 5 variables associated with it:

- **sdev:** standard deviation of each principal components
- **# rotation:** the contribution of each variable to each principal component
- **# center:** sample mean for the original variable
- **# scale :** standard deviation of the original variable
- **# x :** value of the PCA for every sample points
-



```
> Admissions.pca$rotation
                   PC1        PC2         PC3         PC4         PC5
GRE Score   -0.4584129  0.4353965 -0.15863838  0.33324145  0.68122794
TOEFL Score -0.4619892  0.3758477 -0.04996218 -0.78114404 -0.18061638
SOP         -0.4333303 -0.3837314  0.80196762 -0.01365465  0.14709362
LOR         -0.3967937 -0.7061885 -0.57138827 -0.09067517  0.09563505
CGPA        -0.4808360  0.1523693 -0.05197140  0.51995740 -0.68740354
```

Figure 12: PCA

PC1 has almost same coefficients, they being the same weightage.
PC2 has GRE and TOEFL with positive coefficients and others with negative coefficients.

(ii) Plotting PCA

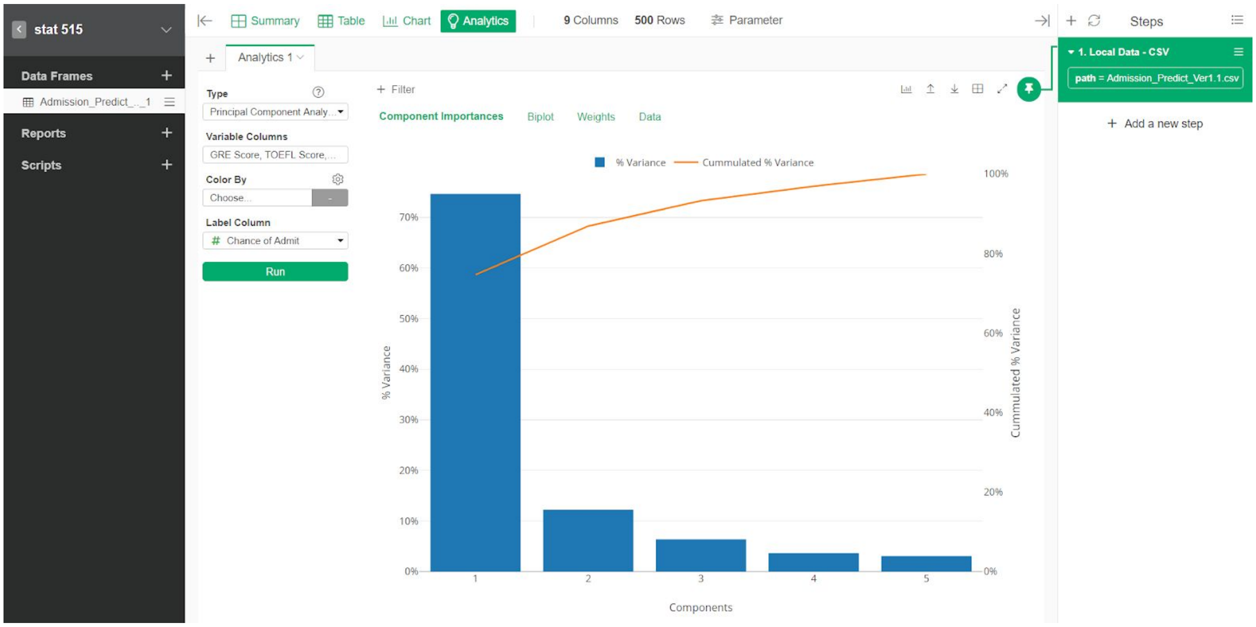
We can make a biplot to visualize the Principal component Analysis. Biplot allows us to visualize the data and how they are similar to one another or different from the other in the PCA. It also showcases the variables that contribute to each principal component.

We have implemented the biplot using **Exploratory** Tool.

Using Exploratory Tool

Please refer to Appendix C for the complete explanation of how the tool was used.

Exploratory is a simple and user friendly UI tool that is used to analyze deep insights in to machine learning algorithms or say Data Science.
The exploratory tool on providing the dataset gives me the ‘Component Importance’, ‘Biplot’ and the ‘Weights’.



Type: Principal Component Analy...
Variable Columns: GRE Score, TOEFL Score, ...
Color By: Choose...
Label Column: # Chance of Admit
1. Local Data - CSV: path = Admission_Predict_Ver1.1.csv

Figure 13: PCA

The component importance shows the scree plot and cumulative scree plot in one graph.

The PC1 shows the maximum variance in the data i.e 74.7% of variance explained.
The PC2 shows the second maximum variance in the data i.e 12.2% variance explained.
Scree plot is a line plot that showcases the Principal components in the decreasing order of contribution to variance in data. Below is the Scree plot achieved.
Each principal component represents the fraction of total variance in the data .

Cumulative scree Plot: This plot shows the cumulative variance explained by all Principal components. SInce its cumulative, it’s a increasing linear functionality.

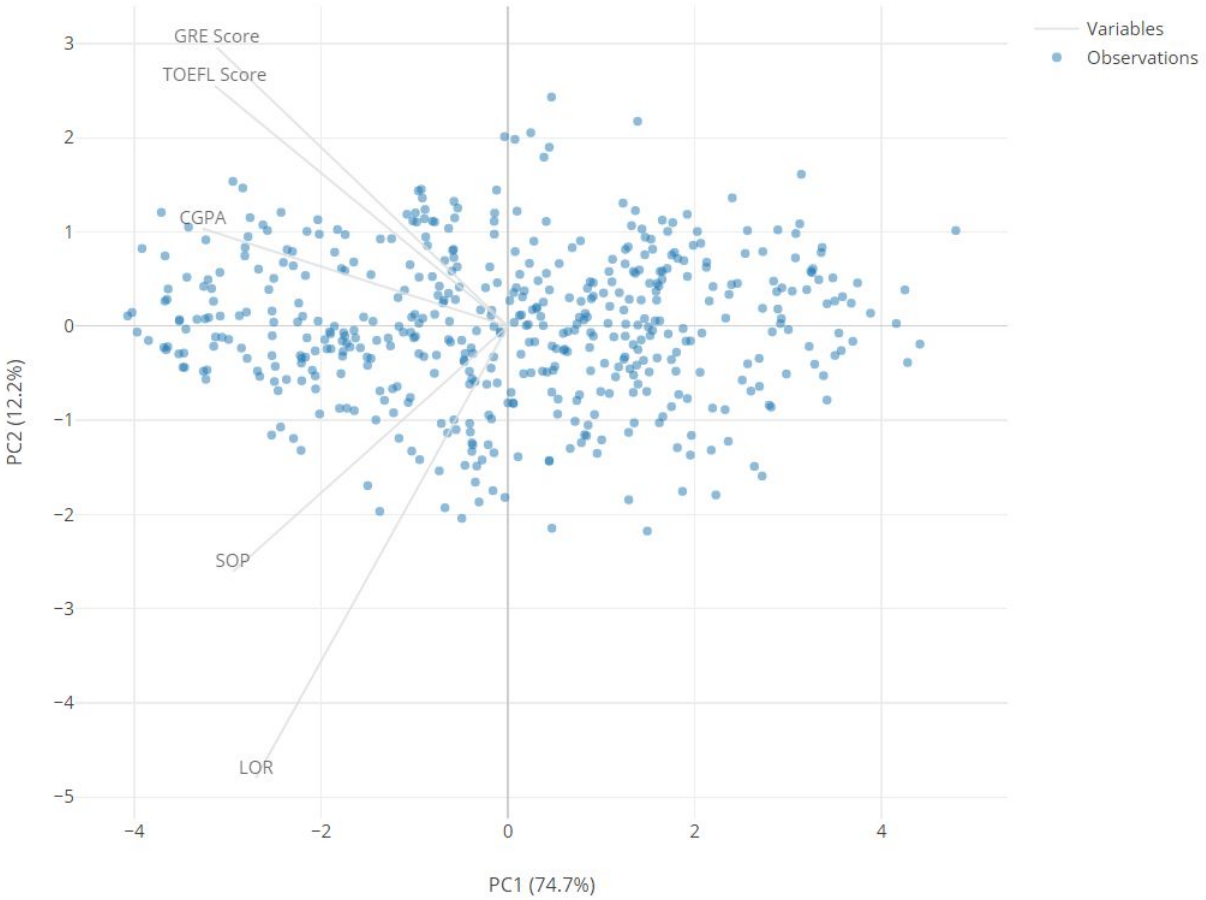


Figure 14: PCA

The biplot is produced by the Exploratory tool by running Principal component Analysis tool.

For the PC1 component to be high, all variables should have a good score which is high. Left hand side of the PC1 has the good marks and right hand side has the bad marks. All the variables have the same weight and hence pushes the PC1 to the left hand side.

Coming to PC2, for the PC2 score to be high, the student should have scored high GRE and high TOEFL score and low LOR and SOP. This makes the PC2 have a high score and having high SOP and LOR and low GRE and TOEFL score would push PC2 down.

Conclusions and Future Work

According to the models, the factors that have the most significant influence on graduate school admissions are:
GRE and CGPA
This was calculated by looking at the p-value and rejecting the hypothesis when the number is smaller than the statistically significant.

Future work of the project can be done in various ways by comparing the data to other graduate school admissions data for countries other than India. Another way to continue to work on the project would be to explore the K means clustering modeling method shown below. Figures 15 and 16 below is for various k values from 2 to 5 and the number of clusters ideal for the dataset.

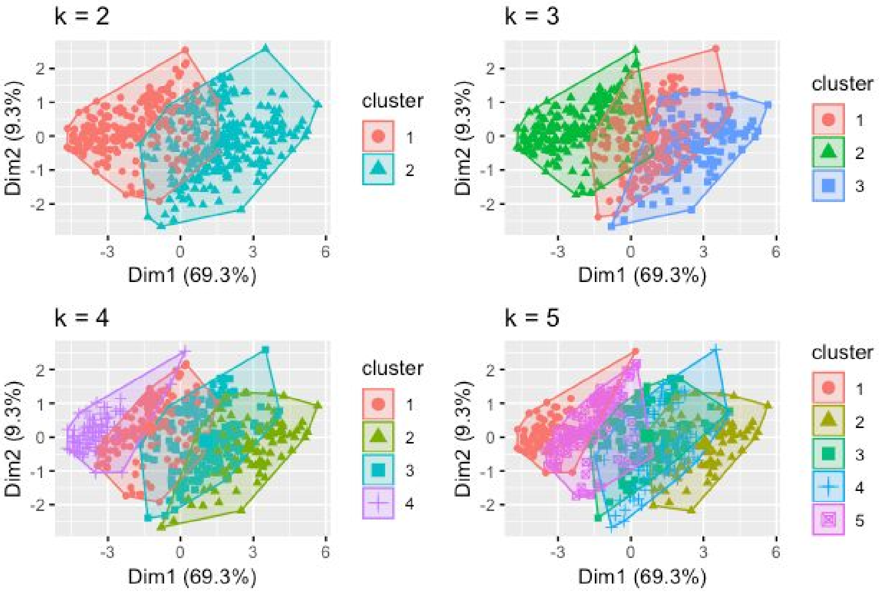


Figure 15: k means clustering for k = 2, 3, 4, and 5 clusters

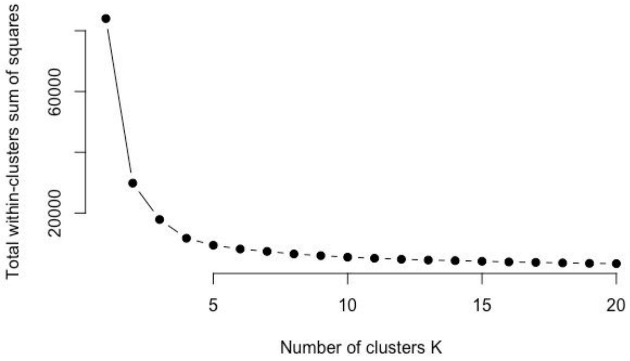


Figure 16: Plots of number of k clusters

The plot in Figure 6 graphs the total within clusters sum of squares with the number of clusters. The optimal number of clusters is approximately 5 clusters. Other pairwise scatter plots and k means square clustering plots that were created with the data are in Appendix B. Future work for the dataset can be to determine the factors that affect the number of clusters and what variables are grouped in each of the clusters by looking at the cluster data and principal components. Limitations of the data is that the admission data was just the rate and not whether or not the applicant got into the college and which college the student applied to.

References:

Acharya, M. S., Armaan, A., & Antony, A. S. (2019). Graduate Admissions. Retrieved from https://www.kaggle.com/mohansacharya/graduate-admissions

Dorbala, R. (2018, August 23). Correlogram in R Studio. Retrieved from https://www.youtube.com/watch?v=2jeOeYSvozQ

K-means Cluster Analysis. (n.d.). Retrieved from https://uc-r.github.io/kmeans_clustering

Scatter Plot in R using ggplot2 (with Example). (n.d.). Retrieved from https://www.guru99.com/r-scatter-plot-ggplot2.html

Soni, N. (2017, August 30). Correlogram in R. Retrieved from https://www.youtube.com/watch?v=DFVjUFWGjHw

Visualize correlation matrix using correlogram. (n.d.). Retrieved from http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram

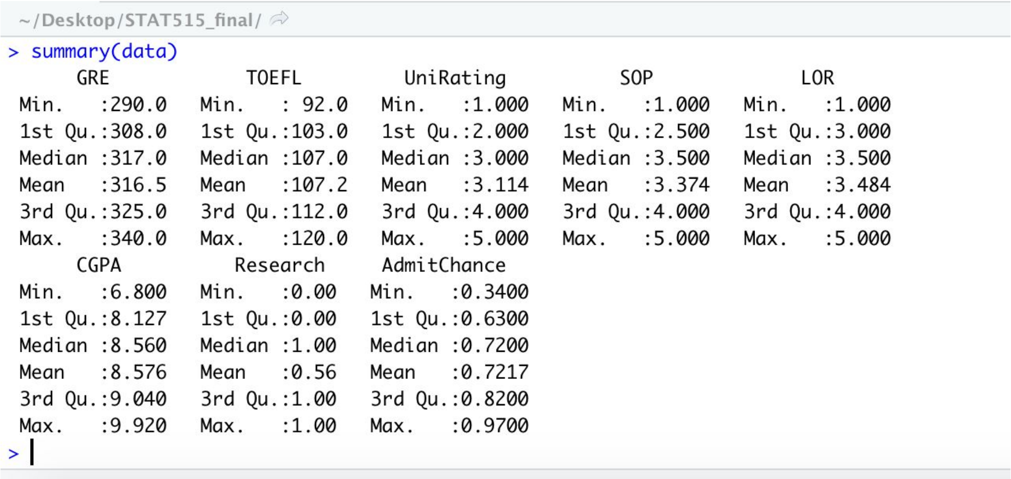
Wright, K. (2018, July 09). Examples for the corrgram package. Retrieved from https://cran.r-project.org/web/packages/corrgram/vignettes/corrgram_examples.html

Principal Component Analysis in R

https://www.datacamp.com/community/tutorials/pca-analysis-r

Appendix

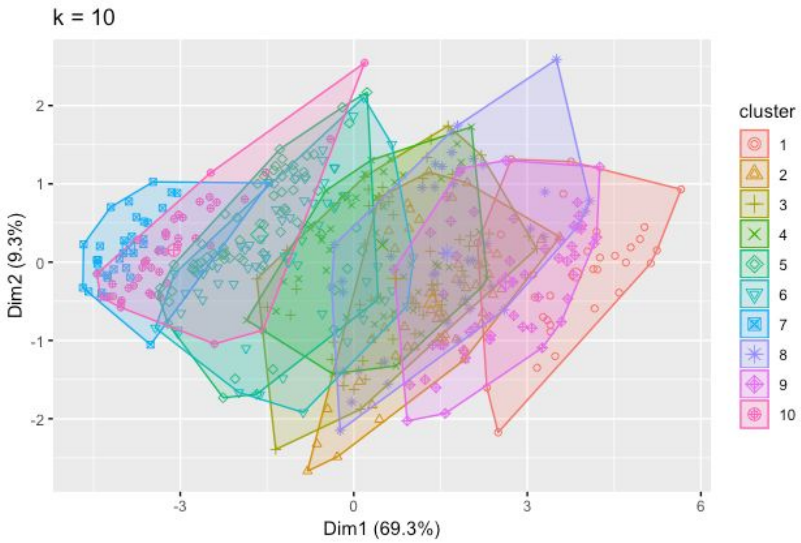
Appendix A: Summary of the data (mean, median, mode, max, min, and quartiles)

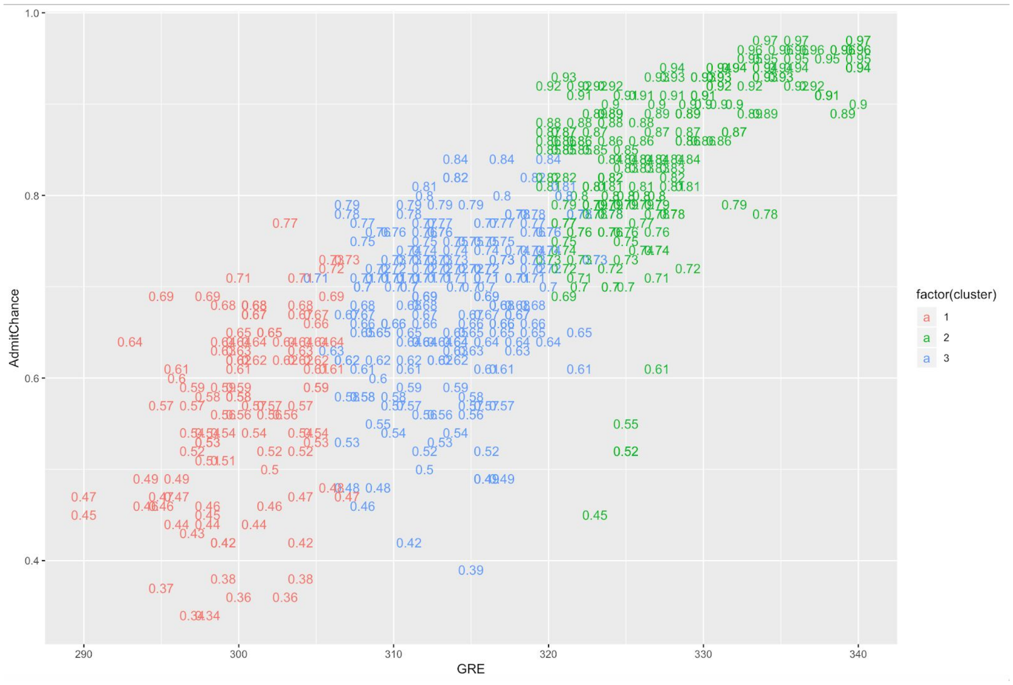


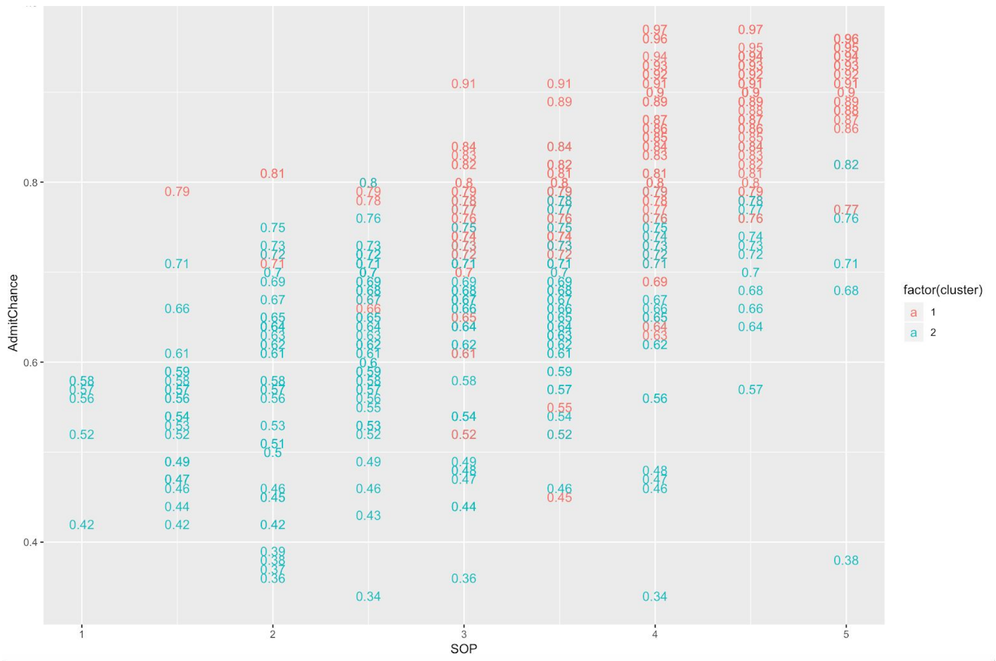
```
> summary(data)
      GRE            TOEFL          UniRating          SOP             LOR       
 Min.   :290.0   Min.   : 92.0   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:308.0   1st Qu.:103.0   1st Qu.:2.000   1st Qu.:2.500   1st Qu.:3.000  
 Median :317.0   Median :107.0   Median :3.000   Median :3.500   Median :3.500  
 Mean   :316.5   Mean   :107.2   Mean   :3.114   Mean   :3.374   Mean   :3.484  
 3rd Qu.:325.0   3rd Qu.:112.0   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :340.0   Max.   :120.0   Max.   :5.000   Max.   :5.000   Max.   :5.000  
      CGPA          Research     AdmitChance    
 Min.   :6.800   Min.   :0.00   Min.   :0.3400  
 1st Qu.:8.127   1st Qu.:0.00   1st Qu.:0.6300  
 Median :8.560   Median :1.00   Median :0.7200  
 Mean   :8.576   Mean   :0.56   Mean   :0.7217  
 3rd Qu.:9.040   3rd Qu.:1.00   3rd Qu.:0.8200  
 Max.   :9.920   Max.   :1.00   Max.   :0.9700  
> 
```

Appendix B: k means clustering cluster plots

Below shows 10 clusters and the pairwise scatter plots to compare the clusters.




AdmitChance
1.0
0.8
0.6
0.4
290
300
310
320
330
340
GRE
factor(cluster)
a 1
a 2
a 3

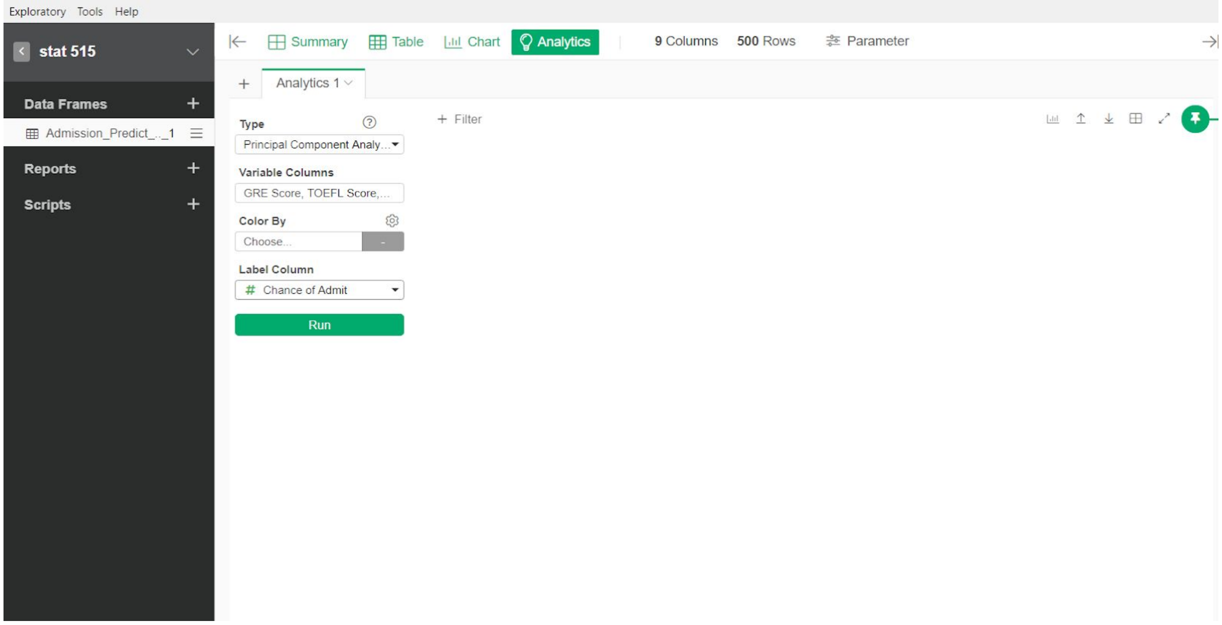

AdmitChance
0.8
0.6
0.4
1
2
3
4
5
SOP
factor(cluster)
a 1
a 2

Appendix C: Step by Step instruction to work with Exploratory tool

How to start of with the Exploratory tool ?

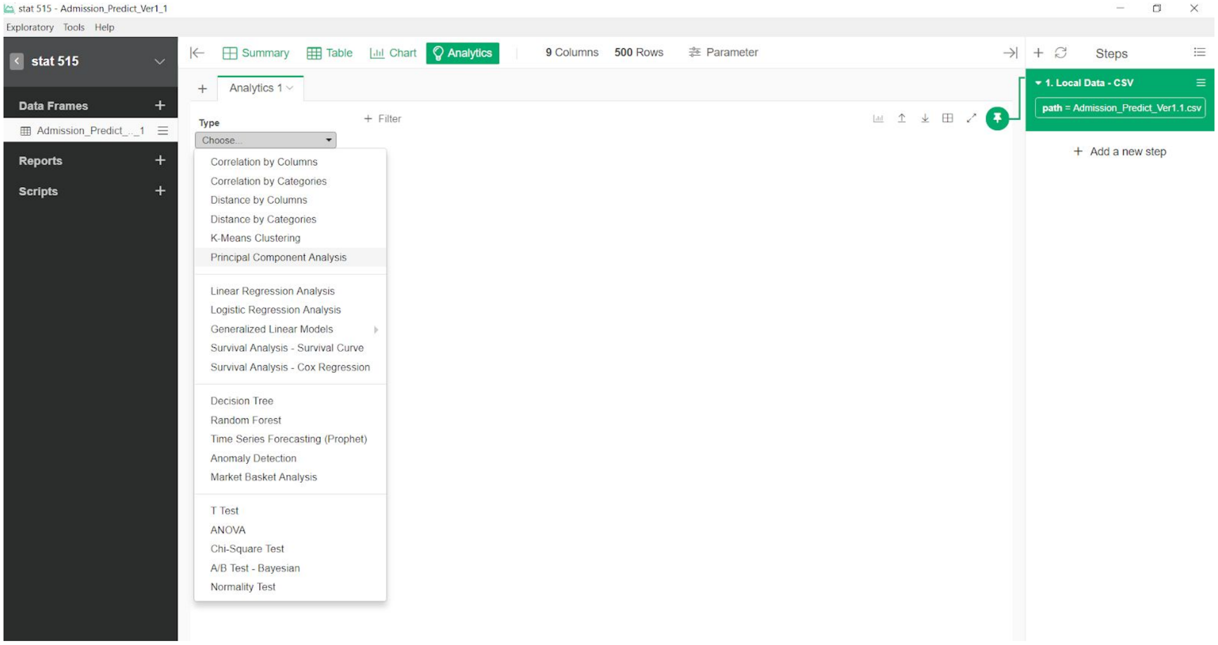
We have to create a project after the tool boots.
Entering into the project, We browse and upload the dataset into the tool under Data Frames section.
Next choose **Analytics** tab in the top bar.
There will be 4 sections available on the left side of the page.

- Type - the type of algorithm you want to apply on the dataset
- Variable columns - Represents the predictors you want to choose
- Color by - Gives a distinct color to the group/clusters
- Label column - Choose the response variable



Type: Principal Component Analy...
Variable Columns: GRE Score, TOEFL Score,...
Color By: Choose...
Label Column: Chance of Admit

The below picture shows the ample type of algorithm that you can choose for your dataset



path = Admission_Predict_Ver1.1.csv

On clicking ‘Run’, it gives me the ‘Component Importance’, ‘Biplot’ and the ‘Weights’.