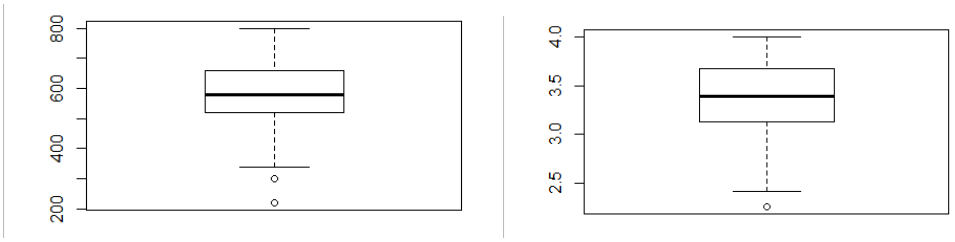
Project Name: College Admission Model

- Find the missing values. (if any, perform missing value treatment)

Ans: No missing value

- Find outliers (if any, then perform outlier treatment)

Ans: Checked using boxplot. Outliers are removed by using the equation: (Q1-1.5*IQR) to (Q2+1.5*IQR)



GRE GPA

- Find the structure of the data set and if required, transform the numeric data type to factor and vice-versa.

Ans: Before

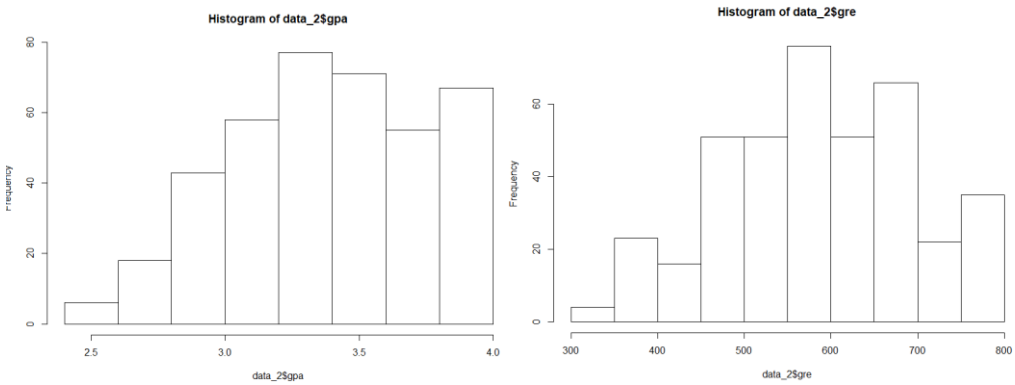
```
$ admit      : int  0 1 1 1 0 1 1 0 1 0 ...
$ gre        : int  380 660 800 640 520 760 560 400 540 700 ...
$ gpa        : num  3.61 3.67 4 3.19 2.93 3 2.98 3.08 3.39 3.92 ...
$ ses        : int  1 2 2 1 3 2 2 2 1 1 ...
$ Gender_Male: int  0 0 0 1 1 1 1 0 1 0 ...
$ Race       : int  3 2 2 2 2 1 2 2 1 2 ...
$ rank       : int  3 3 1 4 4 2 1 2 3 2 ...
```

After factor transformation

```
$ admit      : Factor w/ 2 levels "0","1": 1 2 2 2 1 2 2 1 2 1 ...
$ gre        : int  380 660 800 640 520 760 560 400 540 700 ...
$ gpa        : num  3.61 3.67 4 3.19 2.93 3 2.98 3.08 3.39 3.92 ...
$ ses        : Factor w/ 3 levels "1","2","3": 1 2 2 1 3 2 2 2 1 1 ...
$ Gender_Male: Factor w/ 2 levels "0","1": 1 1 1 2 2 2 2 1 2 1 ...
$ Race       : Factor w/ 3 levels "1","2","3": 3 2 2 2 2 1 2 2 1 2 ...
$ rank       : Ord.factor w/ 4 levels "1"<"2"<"3"<"4": 3 3 1 4 4 2 1 2 3 2 ...
```

- Find whether the data is normally distributed or not. Use the plot to determine the same.

Ans: GRE data is not normally distributed : mean (591.2) > median (580.0), so right skewness, GPA also not normally distributed : median (3.4) > mean (3.398), so left skewness



- Normalize the data if not normally distributed.

Ans: Normalized using Scale function

- Use variable reduction techniques to identify significant variables.

Ans: Used logistic regression to see the important variables based on the P-value. And found gre, gpa, and rank as important variables.

- Run logistic model to determine the factors that influence the admission process of a student (Drop insignificant variables)

Ans: Dropped all the variables except gre, gpa, and rank

- Calculate the accuracy of the model and run validation techniques.

Ans: model is tested on the 30% test set and got accuracy of 68.38%

- Try other modelling techniques like decision tree and SVM and select a champion model. Determine the accuracy rates for each kind of model . Select the most accurate model. Identify other Machine learning or statistical techniques

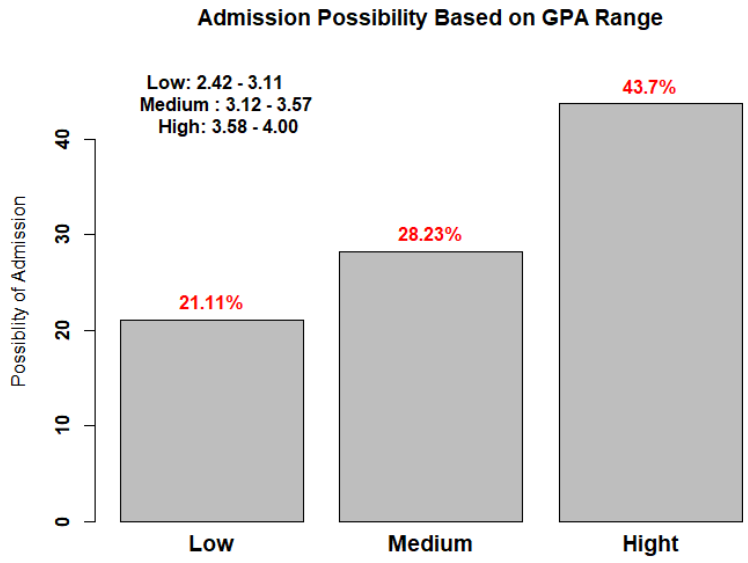
Ans: Applied Decision Tree, SVM, Random Forest, and Naïve Bayes. From all of them Random Forest shows the best result (Accuracy 70.09%). The accuracies of other models are given below:

Algorithm	Accuracy%
Logistic Regression	68.38
SVM	69.23
Decision Tree	68.38
Random Forest	70.09%
Naïve Bayes	70.09%

- Categorize the average of grade point into High, Medium, and Low (with admission probability percentages) and plot it on a point chart. rates for each kind of model . Select the most accurate model. Identify other Machine learning or statistical techniques

Ans: GPA data are categorized using K-means clustering using K=3. The clusters and admission probability of them are summarized below:

	GPA range	Admission Probability
Cluster 1	2.42 - 3.11	21.11%
Cluster 2	3.12 - 3.57	28.33%
Cluster 3	3.58 - 4.00	43.7%



Point cloud plot: