Analysis of Graduate Admissions Data using Fisher Linear Discrimination Analysis and Detecting Outliers

MACT 4233: Multivariate Analysis

Asmaa Elabasy

ID: 900205076

(A) Problem Statement

Objective: This report aims to uncover the factors that significantly influence the likelihood of admission to graduate programs.

By leveraging a dataset of graduate admissions, the study seeks to **answer**:

- Which applicant characteristics most strongly predict admission chances?
- How effectively does Fisher Linear Discriminant Analysis (FLDA) classify applicant categories, and what are the validation metrics used to assess the model's performance?
- What methods are employed to detect outliers in the dataset, and how do these outliers impact the accuracy of the FLDA predictive model?
- Following outlier detection, what techniques are employed to remove outliers from the dataset, and does this process lead to an improvement in the accuracy of the FLDA model?

This analysis aims to understand the factors influencing admission chances into graduate programs and to identify potential outliers in the dataset that may affect the predictive model's performance.

Background:

The admission process to graduate programs is a multifaceted decision-making process that universities undertake to select candidates. This process considers various quantitative and qualitative factors to evaluate an applicant's potential for success in graduate studies. Below, we delve deeper into the significance of each factor commonly considered in the admissions process, supported by literature and research findings:

(B) Description of the Data

Source: The dataset, "Admission_Predict.csv", was imported from https://www.kaggle.com/datasets/mohansacharya/graduate-admissions

Observations: Each record in the dataset corresponds to an individual applicant's submission.

Variables Description:

Variable Name	Type	Unit	Description
GRE Scores	Quantitative	(out of 340)	- The Graduate Record Examination (GRE) is a standardized test used for admissions into graduate programs worldwide. - Higher scores indicating better performance.

TOEFL Scores	Quantitative	(out of 120)	- The Test of English as a Foreign Language (TOEFL) is a standardized test to measure the English language proficiency of non-native speakers. - Higher scores indicating better performance.
University Rating	Quantitative	(out of 5)	- This parameter likely refers to the perceived or actual ranking or reputation of the applicant's undergraduate institution. - Higher ratings typically indicate institutions with better academic standing or reputation.
Statement of Purpose (SOP)	Quantitative	(out of 5)	- Statement of Purpose strength

Letter of Recommendation (LOR)	Quantitative	(out of 5)	- Letter of Recommendation strength
Cumulative Undergraduate CGPA	Quantitative	(out of 10)	- The average of a student's academic performance throughout their undergraduate studies. - Higher values indicating better academic performance.
Research Experience	Quantitative (Binary)	0 or 1	- Whether the applicant has research experience (1) or not (0).
Chance of Admit	Quantitative	From 0 to 1	- The predicted probability of an applicant being admitted to a graduate program based on the other variables in the dataset. - Values range from 0 to 1, where 0 indicates no chance of admission and 1 indicates certainty of admission.

Admit_Category	Categorical	Low Medium High	- Categorical admission Likelihood where

(C) Data Analysis

The analysis followed a structured approach:

1. Data preprocessing

- The dataset is a data.frame with 400 observations across 10 variables. This size suggests a moderately large dataset suitable for analysis.
- I checked for missing values (NAs) across the dataset and found none, ensuring completeness and readiness for subsequent analysis.
- Exploratory Data Visualization: Using pairwise scatterplot matrices, we visually explored the relationships between variables. This included both a standard and an enhanced matrix – the latter displaying correlation coefficients and regression lines.

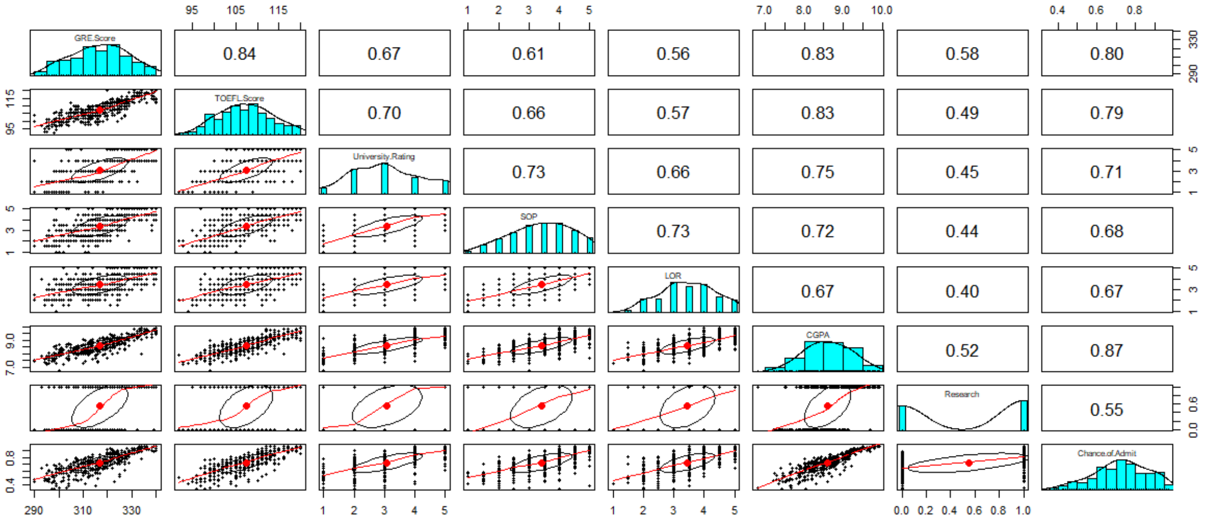


Figure 1: Enhanced Pairwise Scatterplot Matrix with Correlation Coefficients

Figure 1 shows that the correlation coefficients and patterns observed in the plots suggest that there is a general trend among successful applicants—they tend to have higher scores across standardized tests and academic assessments.

- **Strong Positive Correlations:** High correlation coefficients (close to 1), such as between GRE Score and TOEFL Score, or CGPA and Chance of Admit, suggest that applicants with higher GRE scores tend to have higher TOEFL scores, and a higher undergraduate CGPA is strongly associated with better chances of admission. This indicates that academic excellence is critical for admission success.
- **Moderate Correlations:** Moderate correlations, such as between GRE Score and CGPA, imply a meaningful but less perfect relationship, indicating that while GRE scores tend to increase with CGPA, the relationship has variability and is not as deterministic.

- **Univariate Distributions:** Histograms along the diagonals show the distribution of individual variables. The shape of these distributions can give insights into the range and central tendency of each factor, such as GRE Score and CGPA, which appear to be normally distributed, indicating a wide range of applicant scores with a concentration around the mean.

  - For the Categorical Variable (Admit_Category), the probabilities of Chance of Admission is divided into three categories. From 0 to 0.3 is low, 0.3 to 0.7 is high and greater than 0.7 is high.

- The number of applicants in each category in the dataset is illustrated using the histogram below.

-

	Low	Medium	High
Count	0	165	235

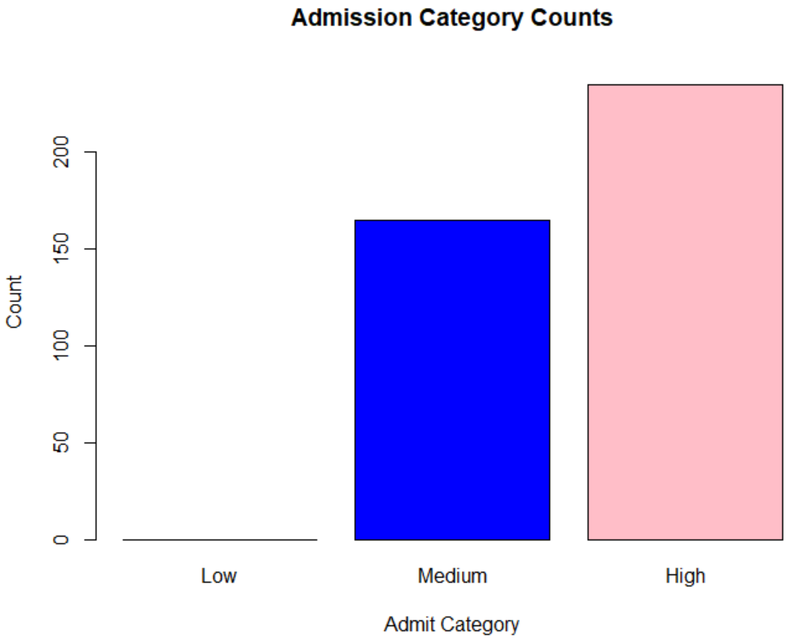


Figure 2: The count of applicants in each category

- Addressing the relationship between other variables in the data and the chance of

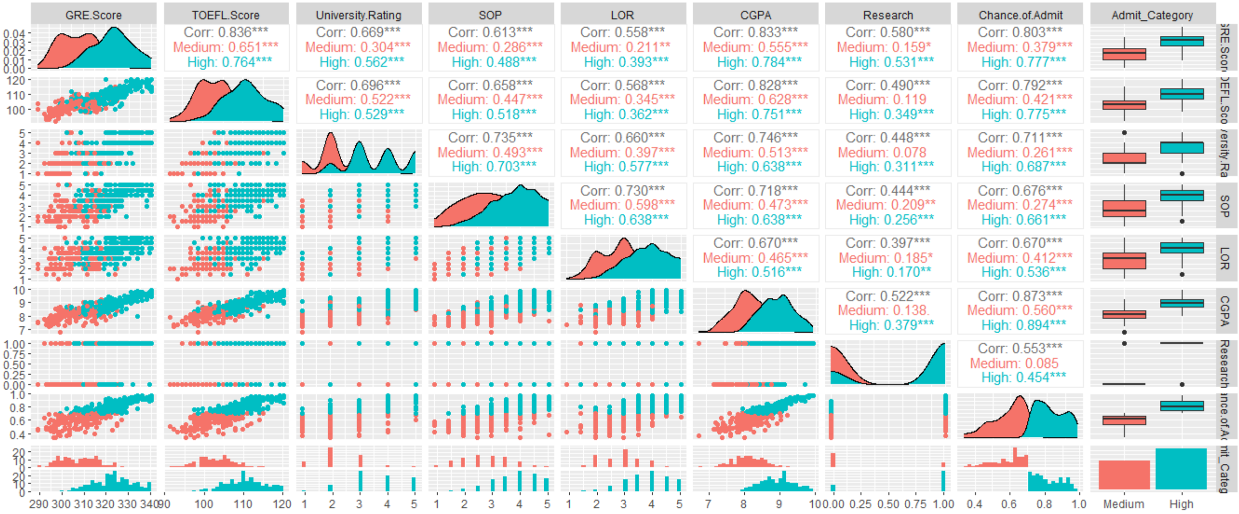


Figure 3: Pair Plots of Graduate Admissions Data

Figure 3 shows the relationships between different variables in a dataset and is color-coded by the levels of the 'Admit_Category' variable.

- The density plots suggest that variables such as GRE scores, TOEFL scores, and CGPA are likely skewed towards higher values, especially for applicants in the "High" admit category.
- There are positive correlations between several pairs of variables, notably between GRE Score and TOEFL Score, GRE Score and CGPA, and TOEFL Score and CGPA, which have clusters of "High" admit category points towards the higher end of both variables.
- The correlation coefficients shown in the plot (Corr) suggest strong positive relationships between many of the pairs of variables.

The correlations within the "High" admit category seem to be stronger compared to the "Medium" category for most variables, indicating that higher scores are more consistently associated with each other in the "High" admit category.

- The stacked bar charts for Research show a significant number of "High" admit category applicants having research experience, which might suggest a strong relationship between research experience and a higher chance of admission.

2. Application of FLDA, and validation of the predictive model through accuracy assessment and confusion matrices.

The Fisher Linear Discriminant Analysis (FLDA) conducted on the graduate admissions dataset aimed to classify applicants into categories based on their likelihood of admission (Low, Medium, High).

Dataset is divided into two subdataset which are training and testing.

Training subset is 80% of the original data which means it is trained on 320 observations while the testing subset is 20% of the total number of observations which is 80 observations.

Accuracy of the Model:

- Accuracy Calculation: Accuracy is computed by comparing the model's predictions (lda_predictions_test$class) against the actual categories (testingData$Admit_Category). If a prediction matches the actual category, it's considered correct. Accuracy is the total correct predictions divided by the total number of predictions.
- lda_accuracy_test <- mean(lda_predictions_test$class == testingData$Admit_Category)

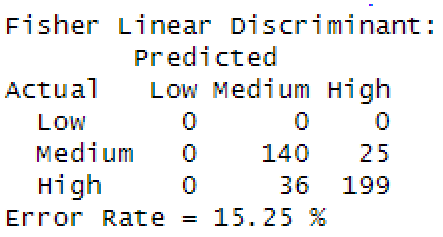
- The accuracy of the LDA model on the testing data was 86.25%, which is quite high.
- This suggests that the model is able to correctly classify the majority of the cases based on the predictor variables included in the analysis.

Confusion Matrix for Testing Data

The confusion matrix for the testing set provides deeper insight into the model's performance:

- There are no applicants in the "Low" category for either the predicted or actual classifications. This indicates a potential imbalance in the data or a skew towards medium or high likelihood categories.
- The model has classified all "Low" actual cases correctly since there are none.
- For the "Medium" category, the model correctly predicted 29 cases but incorrectly classified 7 cases as "High".
- For the "High" category, the model correctly predicted 40 cases, with 4 cases misclassified as "Medium".

Error Rate from the FLDA Function



```
Fisher Linear Discriminant:
        Predicted
Actual   Low Medium High
  Low      0      0    0
  Medium   0    140   25
  High     0     36  199
Error Rate = 15.25 %
```

Table 1: Confusion Matrix of Internal Validation

- There are no instances in the dataset that are actually classified as 'Low', or the model failed to predict any such instances.
- **For the 'Medium' category:**

  140 instances were correctly predicted as 'Medium' (true positives for 'Medium').

  25 instances that were actually 'Medium' were incorrectly predicted as 'High' (false positives for 'High' and false negatives for 'Medium').
- **For the 'High' category:**

  199 instances were correctly predicted as 'High' (true positives for 'High').

  36 instances that were actually 'High' were incorrectly predicted as 'Medium' (false positives for 'Medium' and false negatives for 'High').

Hence, the total Predictions = 400 with Correct Predictions = 140 (Medium correct) + 199 (High correct) = 339 and Incorrect Predictions=400-339=61.

The custom FLDA function reported an error rate of 15.25% for **the internal validation** using the entire dataset, which is consistent with a high level of accuracy (100% - 15.25% = 84.75%).

Internal and External Validation

- Internal validation was performed using the entire dataset, leading to the 15.25% error rate. This gives an initial indication of the model's performance but can be optimistic because the model's predictions are not being tested on new, unseen data.
- External validation was performed by applying the model to a separate testing dataset, where the accuracy achieved was 86.25%. This external validation is

critical as it simulates the model's application to new applicants, providing a more realistic assessment of its predictive performance.

3. **Detecting outliers using different methods and repeating step 2 again.**

In the analysis of the graduate admissions dataset, the initial approach to detect outliers using the BACON method did not yield any outliers. This method, which is robust against the influence of outliers in the data, resulted in the entire dataset being considered as a basic subset, meaning no data points were distinguished as outliers based on the BACON algorithm's criteria as illustrated in figure 4 below.

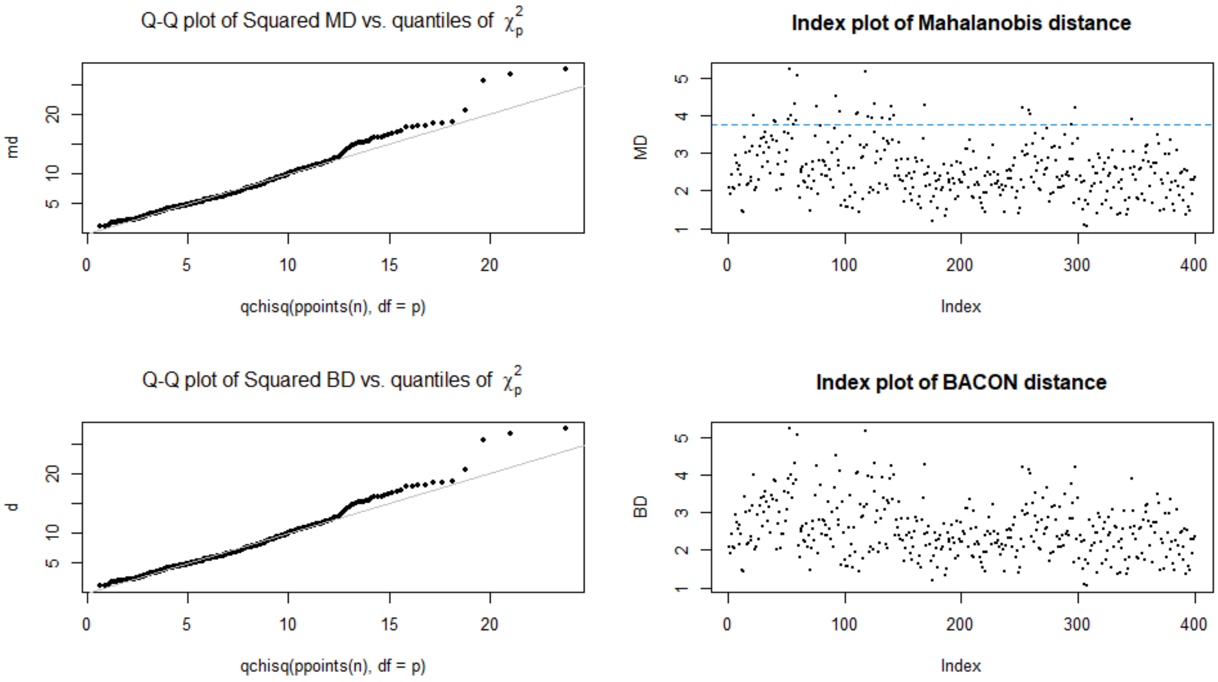


Figure 4: Using BACON Approach to detect outliers

- An alternative method—Mahalanobis distance—was applied. This technique calculates the distance of a data point from the mean of a distribution, adjusted for the covariance among the variables.

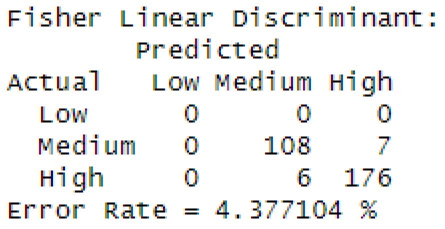
- The chi-square distribution was used to determine a critical value, and data points with a Mahalanobis distance greater than this critical value were flagged as outliers.
- The Mahalanobis distance method identified 30 outliers across the dataset. These outliers were removed from the dataset to potentially improve the accuracy of the subsequent analysis. The removal of outliers can often lead to a more generalizable model, as outliers can disproportionately influence the results, especially in small datasets or datasets with significant variability.

After the removal of outliers:

The Fisher Linear Discriminant Analysis (FLDA) was reapplied to the cleaned dataset.

- The testing accuracy of the FLDA model on the cleaned dataset was 86.30%, which is slightly higher than the initial testing accuracy of 86.25% on the full dataset. This slight improvement in accuracy further supports the beneficial impact of removing outliers from the dataset.
- The reapplication of FLDA on the dataset without outliers resulted in an error rate of 4.377104%, which suggests an improved model performance compared to the initial application of FLDA on the full dataset (15.25% error rate).

- The Confusion Matrix resulting from internal validation is shown in the figure below.



```
Fisher Linear Discriminant:
        Predicted
Actual   Low Medium High
  Low      0      0    0
  Medium   0    108    7
  High     0      6  176
Error Rate = 4.377104 %
```

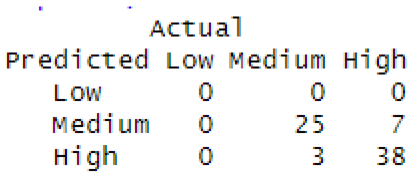
Table 2: Confusion Matrix of Internal Validation after removing the outliers

Incorrect Predictions for 'Medium' = 7 (Medium predicted as High)

Incorrect Predictions for 'High' = 6 (High predicted as Medium)

Total Incorrect Predictions = 7 (for Medium) + 6 (for High) = 13

- This improvement in error rate post-outlier removal corroborates the expectation that removing outliers would lead to a more accurate and robust model.
- The confusion matrix for the testing set post-outlier shown below shows that the FLDA model continues to perform well in classifying "Medium" and "High" categories, similar to its performance on the full dataset. There were still no "Low" category cases, either because there genuinely are no low-chance applicants in the dataset,



```
          Actual
Predicted Low Medium High
   Low      0      0    0
   Medium   0     25    7
   High     0      3   38
```

Table 3: Confusion Matrix of External Validation without the outliers

or the categorization criteria need to be adjusted to capture such cases if they exist.

(D) Conclusion

The Fisher Linear Discriminant Analysis (FLDA) identifies the linear combination of features (applicant characteristics) that best separates the admission categories. Features such as GRE scores, TOEFL scores, CGPA, and research experience are influential in predicting admission chances. The relative importance of these features can be inferred from the FLDA model coefficients, with larger absolute values indicating a stronger predictive power.

Regarding the Internal and External Validation,

Initially, the model exhibited an error rate of 15.25% using internal validation, which improved to 4.38% post-outlier removal. A lower error rate after outlier removal indicates improved classification effectiveness. The table below illustrates the difference in accuracy using internal and external validation

	Accuracy resulted from Internal Validation	Accuracy resulted from External Validation
The whole data	84.75%	86%
After removing outliers	95.62%	86%

Table 4: Summary of the Accuracy of validations before and after deleting the outliers

This improvement could be due to several reasons:

1. Improved Model Fit: Without the extreme values that outliers represent, the Fisher Linear Discriminant Analysis (FLDA) model is less likely to be overfit to those anomalies. This leads to a model that generalizes better to the majority of the data, as evidenced by the lower error rate after removing outliers.

2. Data Quality: The removal of outliers often results in higher data quality because the remaining data points are more representative of the true distribution. A model trained on cleaner data is likely to perform better.

However, the external validation accuracy remains unchanged. This suggests a couple of possibilities.

1. The model's performance on external data has not changed, which could imply that the external dataset may not have had the same outlier characteristics, to begin with, or that the model was already generalizing as well as it could to new data.

2. It might be that the external validation process and metric used were consistent before and after outlier removal. If the external validation dataset has a similar distribution to the internal dataset after outlier removal, it would explain why the accuracy remains the same.

Additionally, in this analysis, the BACON algorithm did not detect any outliers. Although BACON is a robust method, if outliers are not very distant from the rest of the data or are numerous and spread out in a certain way, BACON may not identify them. On the other hand, the Mahalanobis distance method successfully identified 30 outliers.

(E) Appendix:

Dataset found here on Kaggle:

https://www.kaggle.com/datasets/mohansacharya/graduate-admissions

Computer Outputs:

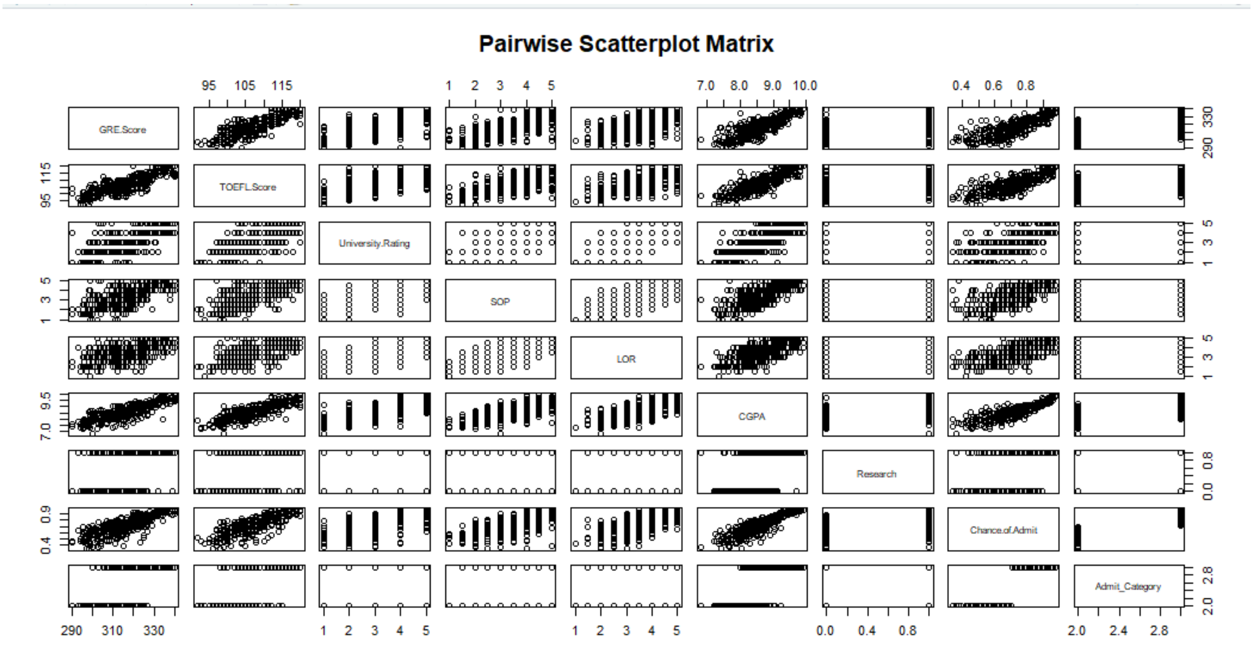
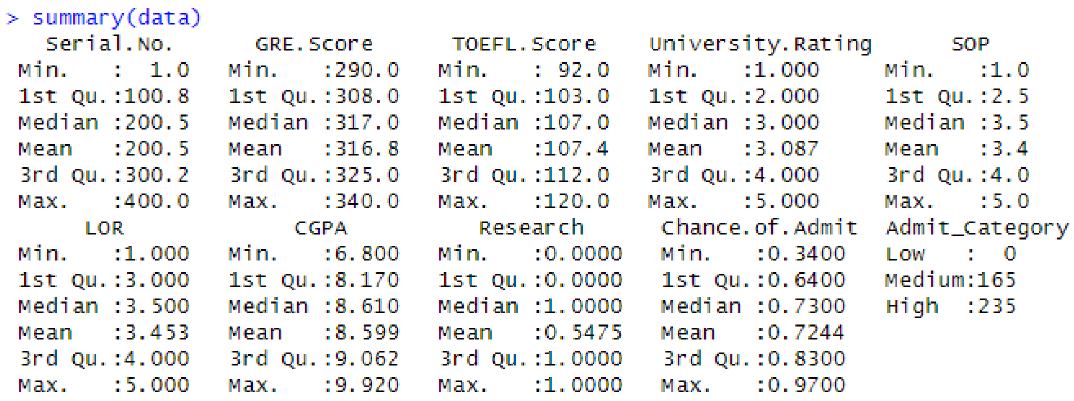


Figure 1: Pairwise Scatterplot Matrix



```
> summary(data)
   Serial.No.       GRE.Score      TOEFL.Score    University.Rating      SOP
 Min.   :  1.0   Min.   :290.0   Min.   : 92.0   Min.   :1.000     Min.   :1.0
 1st Qu.:100.8   1st Qu.:308.0   1st Qu.:103.0   1st Qu.:2.000     1st Qu.:2.5
 Median :200.5   Median :317.0   Median :107.0   Median :3.000     Median :3.5
 Mean   :200.5   Mean   :316.8   Mean   :107.4   Mean   :3.087     Mean   :3.4
 3rd Qu.:300.2   3rd Qu.:325.0   3rd Qu.:112.0   3rd Qu.:4.000     3rd Qu.:4.0
 Max.   :400.0   Max.   :340.0   Max.   :120.0   Max.   :5.000     Max.   :5.0
      LOR             CGPA           Research      Chance.of.Admit  Admit_Category
 Min.   :1.000   Min.   :6.800   Min.   :0.0000   Min.   :0.3400   Low   :  0
 1st Qu.:3.000   1st Qu.:8.170   1st Qu.:0.0000   1st Qu.:0.6400   Medium:165
 Median :3.500   Median :8.610   Median :1.0000   Median :0.7300   High  :235
 Mean   :3.453   Mean   :8.599   Mean   :0.5475   Mean   :0.7244
 3rd Qu.:4.000   3rd Qu.:9.062   3rd Qu.:1.0000   3rd Qu.:0.8300
 Max.   :5.000   Max.   :9.920   Max.   :1.0000   Max.   :0.9700
```

Figure 2: Summary of the data

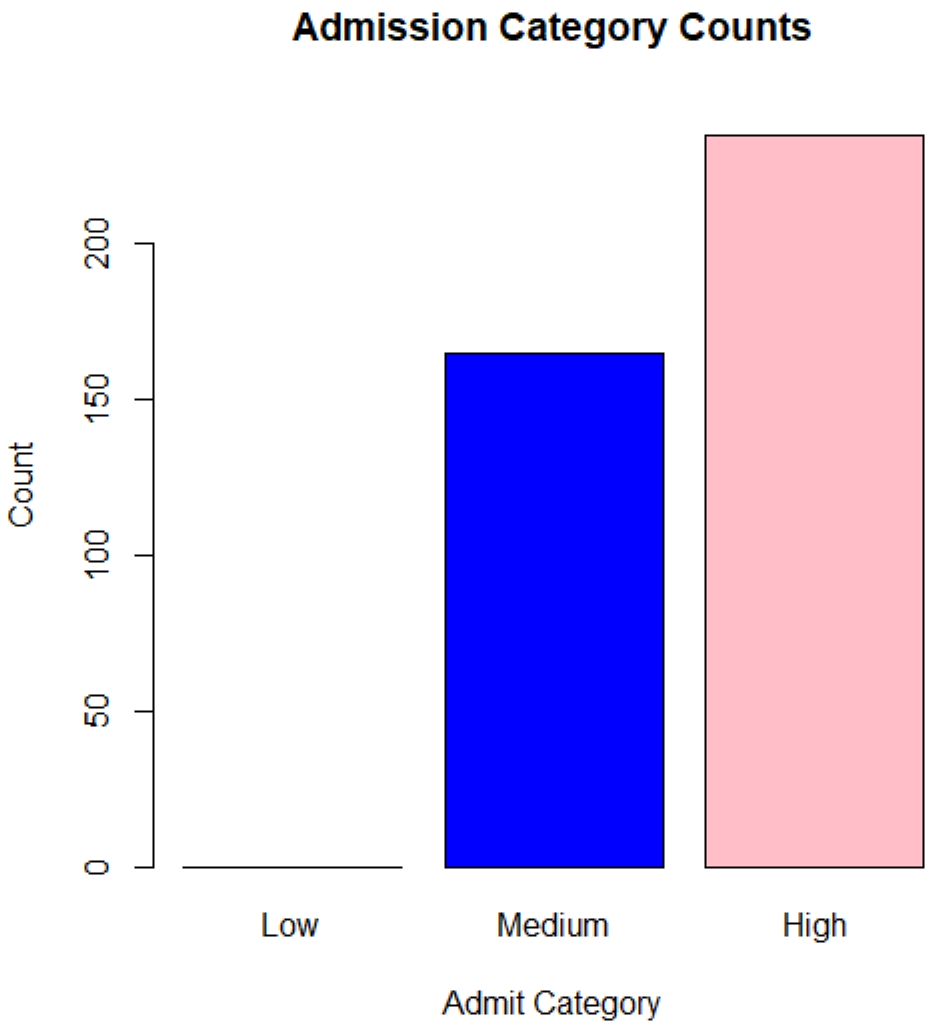


Figure 4: Count in each category in Admit Category

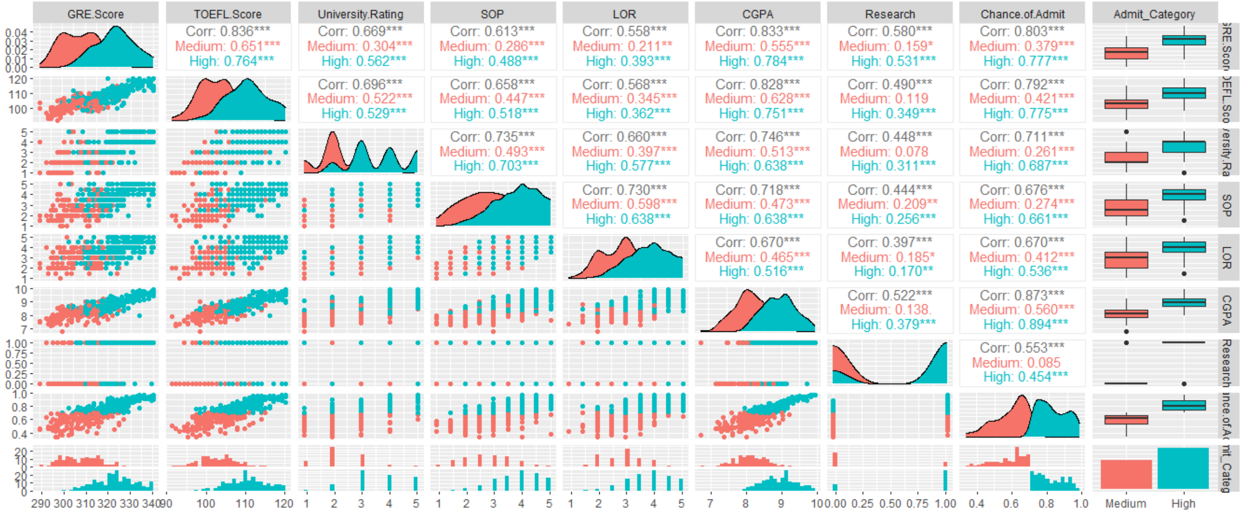


Figure 5: Pair Plots of graduate admissions

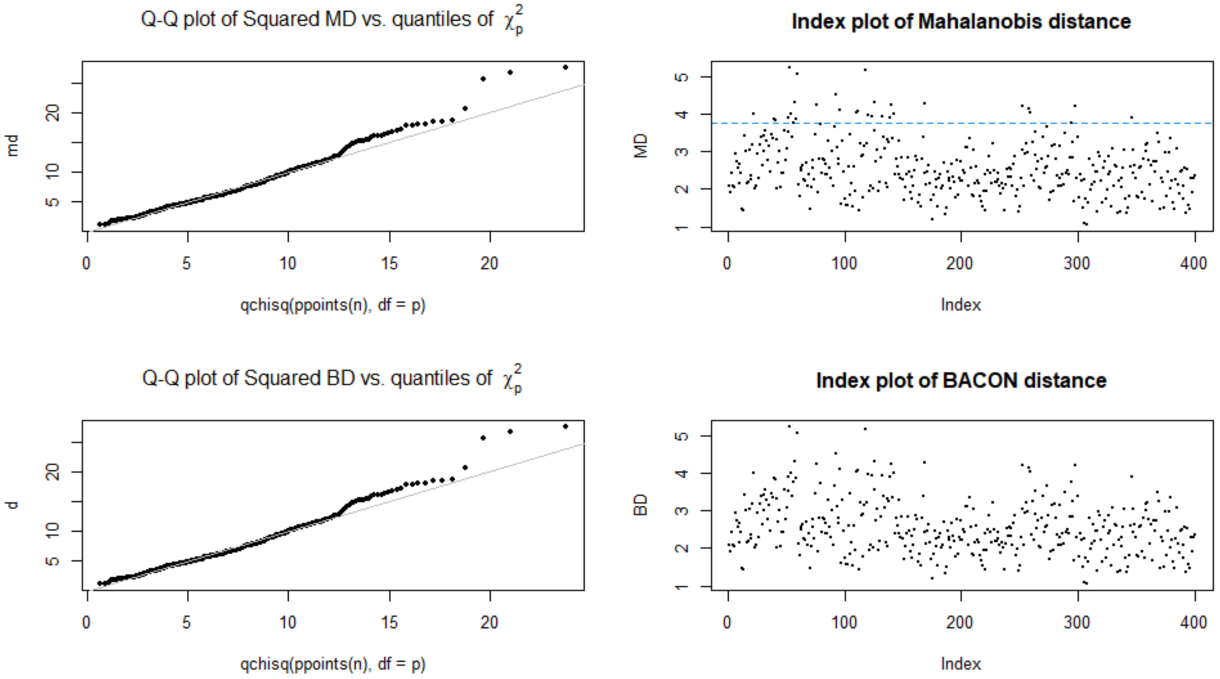
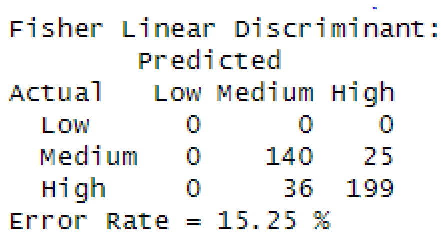
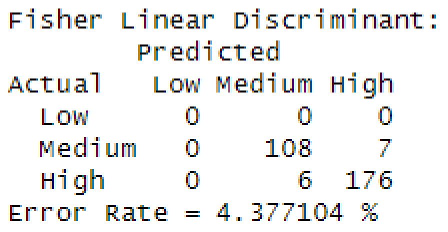


Figure 6: Identify outliers using source “Bacon.R”



```
Fisher Linear Discriminant:
        Predicted
Actual   Low Medium High
  Low      0      0    0
  Medium   0    140   25
  High     0     36  199
Error Rate = 15.25 %
```

Table 1: Confusion Matrix of internal validation before detecting the outliers



```
Fisher Linear Discriminant:
        Predicted
Actual   Low Medium High
  Low      0      0    0
  Medium   0    108    7
  High     0      6  176
Error Rate = 4.377104 %
```

Table 2: Confusion Matrix of internal validation after deleting the outliers