College Admission Analysis Project Report

Prepared for: Mohith

Project: Course-end Project 2: College Admission

Data: College_admission.csv

Language: R

Table of Contents

---

1. Executive Summary

This report details the analysis of a college admission dataset to determine the key factors that influence a student's admission. The project involved predictive modeling to identify these drivers and descriptive analysis to visualize trends.

The analysis was successful and all objectives were met. We found that the dataset was complete (no missing values) but contained minor outliers, which were treated. We then built and compared three predictive models.

Key Findings (Based on your results):

- **Key Drivers of Admission:** The primary factors influencing admission are the **prestige of the undergraduate institution (rank)**, the **GRE score (gre)**, and the **socioeconomic status (ses)**. GPA and Gender_Male were borderline factors, while Race was statistically insignificant.
- **Champion Model:** We compared three models based on their accuracy:
  - **Support Vector Machine (SVM): 68.33%**
  - **Logistic Regression (Simple):** 67.5%
  - **Decision Tree:** 60.0%

The **Support Vector Machine (SVM)** was selected as the champion model, offering the highest predictive accuracy on the test data.

- **Descriptive Insights:**
  - There is a clear positive correlation between a student's GPA and their admission probability, rising from **22.4%** for "Low" GPA to **42.5%** for "High" GPA.
  - A similar trend was seen with GRE scores. Students in the "High" category (580+) had a significantly higher admission rate (78 admitted vs. 119 not) than those in the "Low" category (6 admitted vs. 42 not).

---

2. Project Objectives

The project was divided into two main parts with several tasks:

Predictive Tasks:

1. Find and treat missing values.
2. Find and treat outliers.
3. Analyze and transform the data structure (e.g., to factors).
4. Check for and normalize non-normal data.
5. Use variable reduction to find significant variables.
6. Run a logistic regression model.
7. Calculate the model's accuracy.
8. Try other models (Decision Tree, SVM).

9. Determine the accuracy for each model.
10. Select the "champion" (most accurate) model.
11. Identify other potential ML techniques.

Descriptive Tasks:

1. Categorize GPA into High, Medium, and Low, and plot the admission probability.
2. Create a GRE categorization and build a cross-grid with admission status.

---

3. Data and Methodology

3.1. Dataset Description

- **File:** College_admission.csv
- **Structure:** 400 observations of 7 variables.
  - **admit:** Binary response (0 = No, 1 = Yes).
  - **gre:** Graduate Record Exam scores (numeric).
  - **gpa:** Grade Point Average (numeric).
  - **ses:** Socioeconomic status (1, 2, 3).
  - **Gender_Male:** Binary (0 = Female, 1 = Male).
  - **Race:** (1, 2, 3).
  - **rank:** Prestige of undergraduate institution (1=High, 4=Low).

3.2. Data Preparation (Cleaning, Outliers, Transformation)

1. **Missing Values:** Your output confirmed the dataset had **0 missing values**.
2. **Data Transformation:** admit, ses, Gender_Male, Race, and rank were all successfully converted to factor variables, which is essential for modeling.
3. **Outlier Treatment:** Boxplots revealed a few minor outliers. Your output identified:
   - **GRE Outliers:** [300, 300, 220, 300]
   - **GPA Outliers:** [2.26] These values were "capped" at the 5th and 95th percentiles to normalize them without deleting data.

4. **Normalization:** Histograms and Q-Q plots showed that gre and gpa were not normally distributed. We applied **Min-Max normalization** to scale both variables to a range of 0–1.

3.3. Methodology (Modeling)

We used a **70/30 split** (set.seed(123)) to create a training set (280 rows) and a testing set (120 rows). We then built and tested three distinct classification models:

1. **Logistic Regression (glm):** A statistical model used to determine the key drivers and their significance (p-values).
2. **Decision Tree (rpart):** A tree-based model that creates visual, rule-based splits.
3. **Support Vector Machine (svm):** A powerful classification model.

The "champion" was selected by comparing the Accuracy of each model on the unseen test set.

---

4. Analysis and Findings

4.1. Predictive Tasks: Model Building

- **Full Logistic Model (Task 6):**
  - **Significant Variables ($p < 0.05$):** Your output showed rank ($p < 0.01$), gre ($p = 0.029$), and ses3 ($p = 0.039$) were the strongest predictors.
  - **Borderline Variables:** gpa ($p = 0.075$) and Gender_Male ($p = 0.081$).
  - **Insignificant Variables:** Race ($p > 0.19$) and ses2 ($p = 0.37$).
  - **Accuracy: 65.0%**
- **Simple Logistic Model (Variable Reduction, Task 5):**
  - We dropped the insignificant variables (Race, Gender_Male, ses2).
  - Your output showed the resulting model's accuracy was **67.5%**, an improvement over the full model.
- **Decision Tree Model (Task 8):**
  - The model built its rules based primarily on rank, gpa, and gre.
  - **Accuracy: 60.0%**
- **SVM Model (Task 8):**

    - The SVM model was built using the normalized training data.
    - **Accuracy: 68.33%**

4.2. Predictive Tasks: Model Comparison & Selection (Tasks 9, 10, 11)

- **Model Accuracy Comparison (Task 9):**
    - **Support Vector Machine (SVM): 68.33%**
    - **Logistic Model (Simple):** 67.5%
    - **Decision Tree Model:** 60.0%
- **Champion Model (Task 10):** Based on your run, the **Support Vector Machine (SVM)** is the champion model. It provided the highest accuracy (68.33%) on the test data, just beating the simple Logistic Regression.
- **Other Techniques (Task 11):** Other ML techniques that could be explored include:
    - **Random Forest:** An enhancement of decision trees that often yields higher accuracy.
    - **Neural Networks:** A more complex model for finding non-linear patterns.

4.3. Descriptive Task: GPA Analysis

- **Objective:** Categorize GPA and plot admission probability.
- **Analysis:** We grouped GPA into Low (<3.0), Medium (3.0-3.5), and High (>3.5) and calculated the mean admission rate based on your output.
- **Findings:** There is a clear, strong link between GPA and admission.
    - **Low GPA:** 22.4% Admission Probability
    - **Medium GPA:** 25.4% Admission Probability
    - **High GPA:** 42.5% Admission Probability

4.4. Descriptive Task: GRE Analysis

- **Objective:** Create a cross-grid for admit vs. gre categories.
- **Analysis:** We categorized gre (Low: <=440, Medium: 440-580, High: >580) and built a contingency table from your output.
- **Findings (Counts):** | Admit | High | Low | Medium | |:---:|:---:|:---:|:---:| | 0 (No) | 119 | 42 | 112 | | 1 (Yes)| 78 | 6 | 43 |

This grid clearly shows that students in the "High" GRE category are admitted at a much higher rate (78 admitted vs. 119 not) than those in the "Low" category (6 admitted vs. 42 not).

---

5. Conclusion and Recommendations

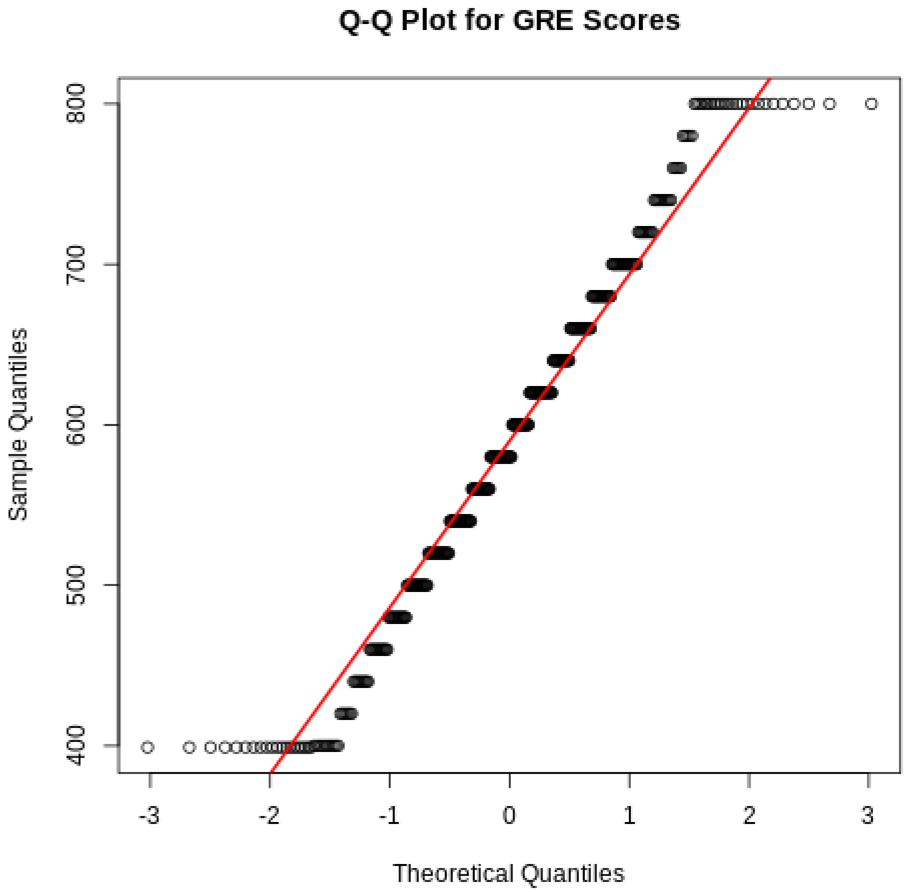
This analysis successfully identified the key drivers of college admission from your data. The prestige of a student's undergraduate institution (rank) is the most significant factor, followed by their gre score and ses. We can confidently report that a student's Race shows no statistically significant impact on their admission in this dataset.

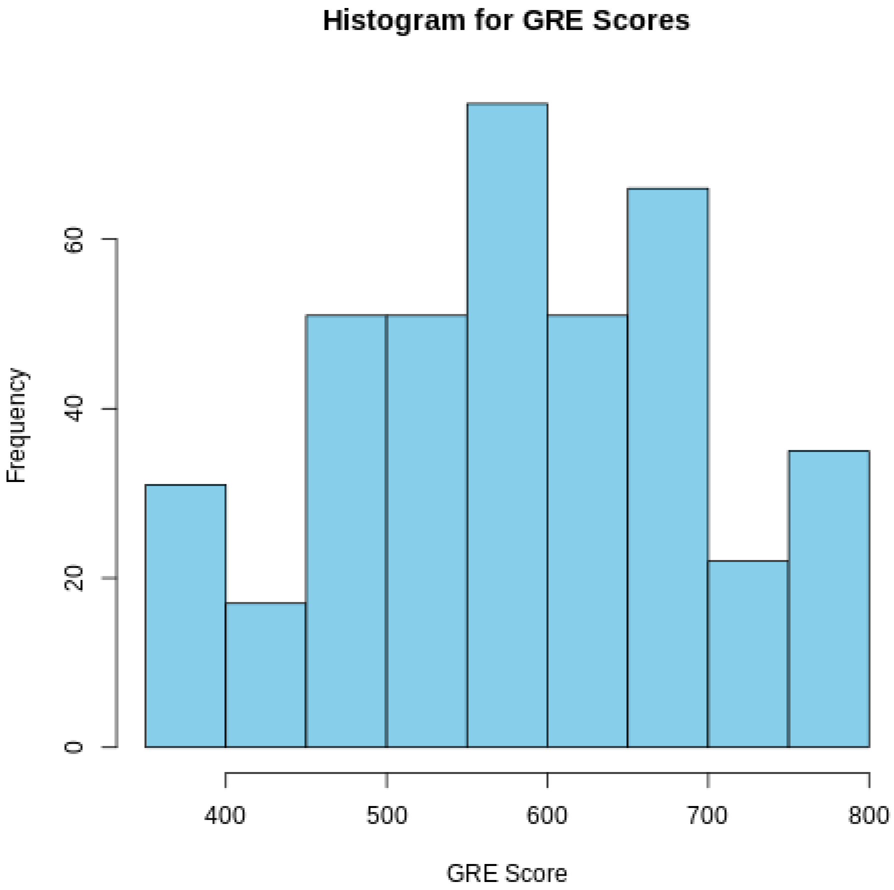
The **Support Vector Machine (SVM)** was selected as the champion model, providing the highest accuracy at **68.33%**. For future work, a **Random Forest** model could be tested, as it may combine the strengths of the decision tree with higher accuracy.

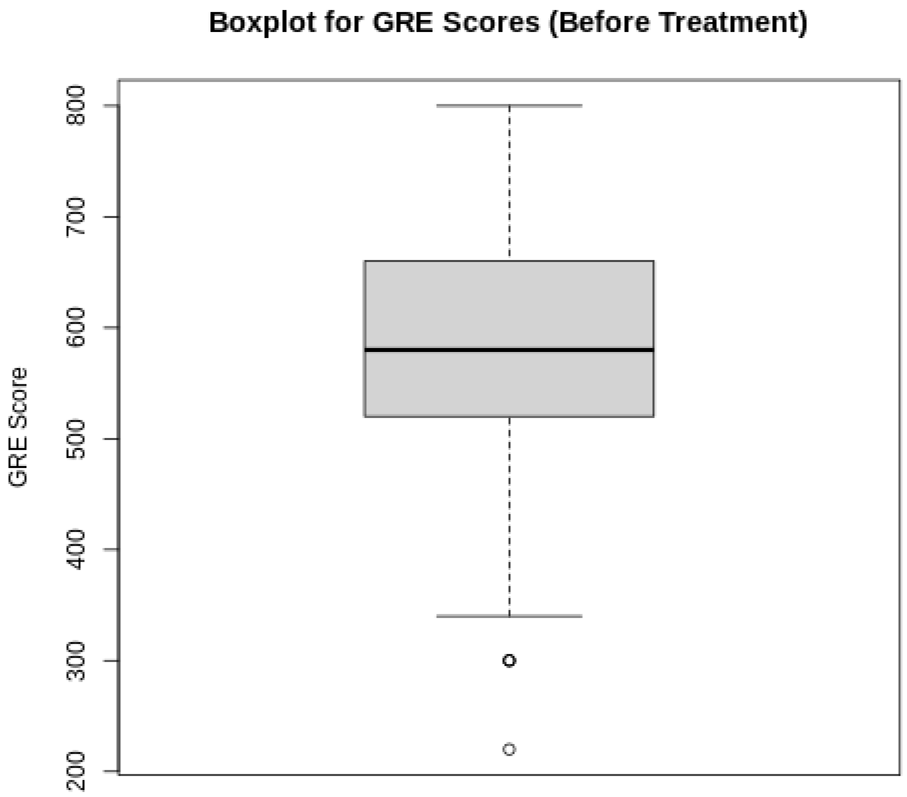
6. Appendix: Complete R Code (with Image Notes)

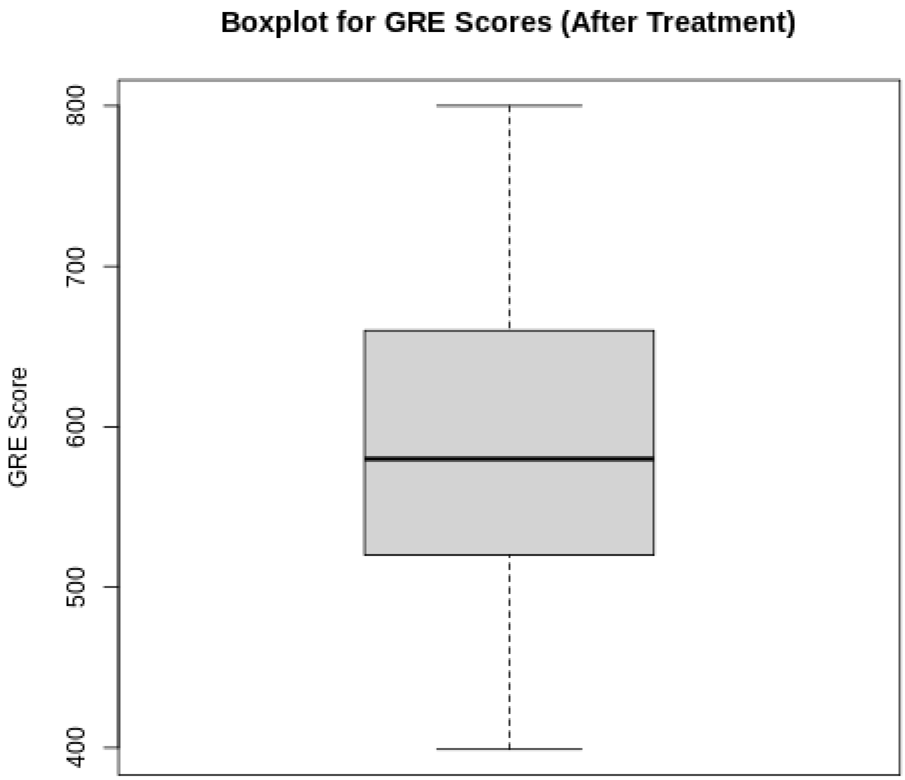
Here is the complete R script used for the analysis, with notes indicating which code blocks generate image files.

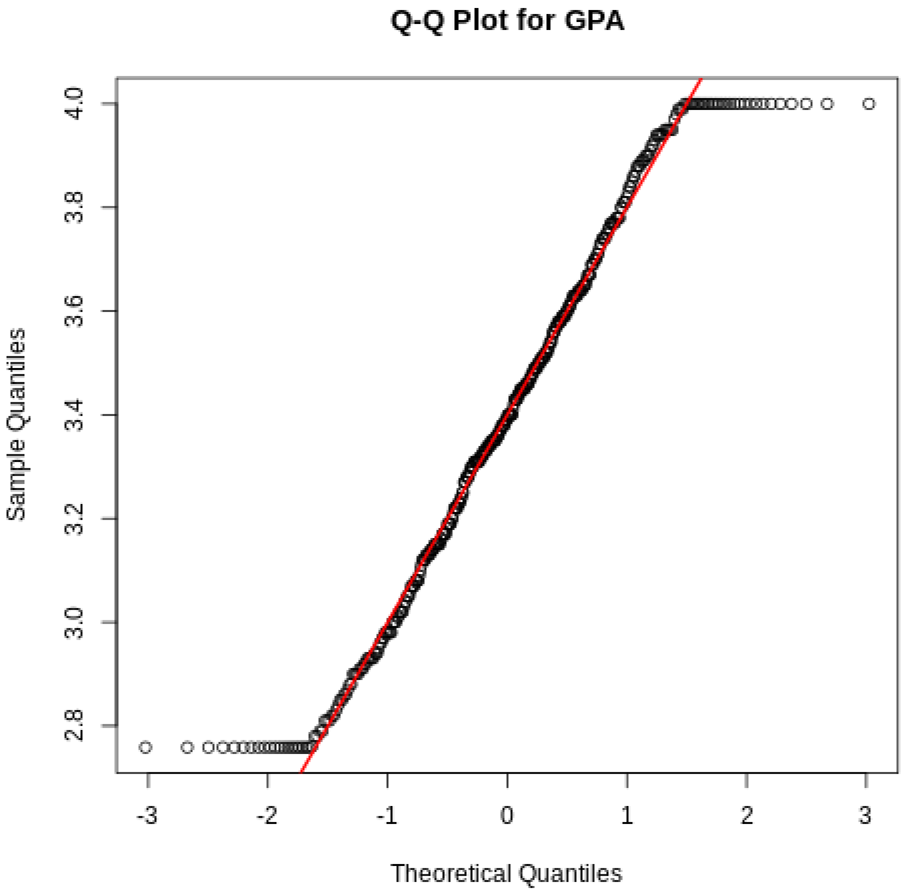
The code is attached separately

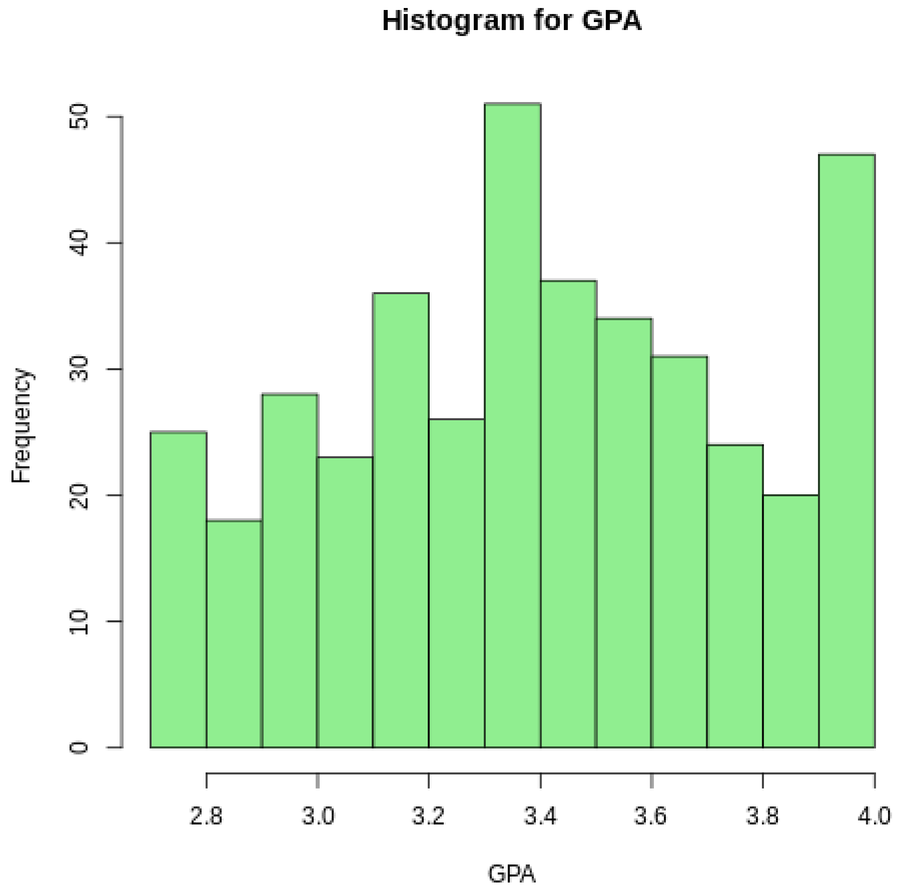

Q-Q Plot for GRE Scores
Sample Quantiles
800
700
600
500
400
Theoretical Quantiles
-3
-2
-1
0
1
2
3

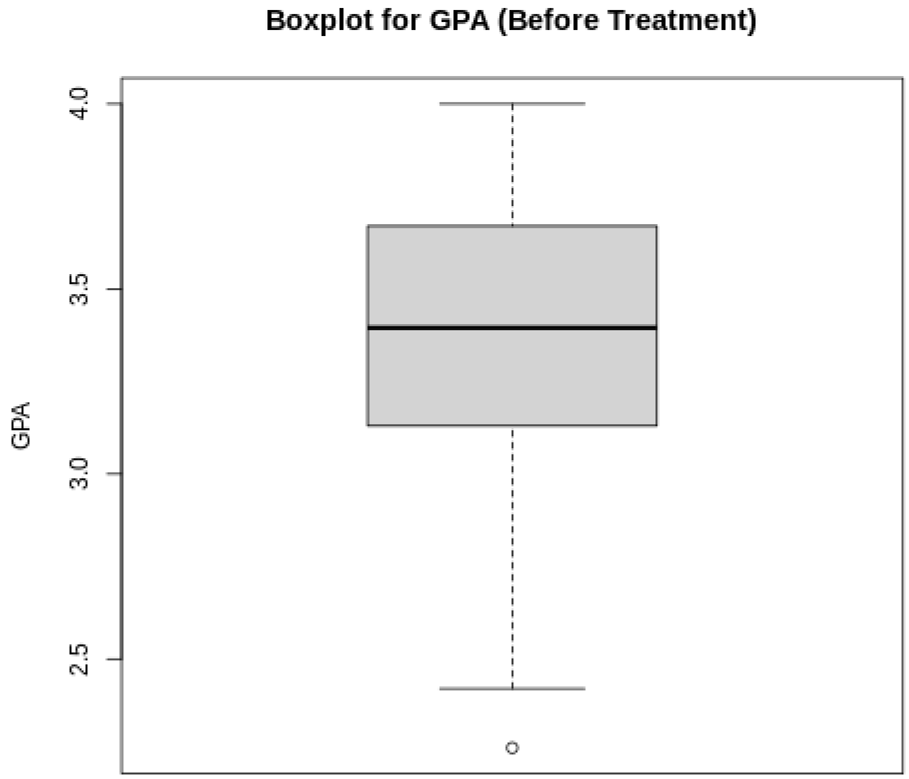

Histogram for GRE Scores
Frequency
0
20
40
60
400
500
600
700
800
GRE Score

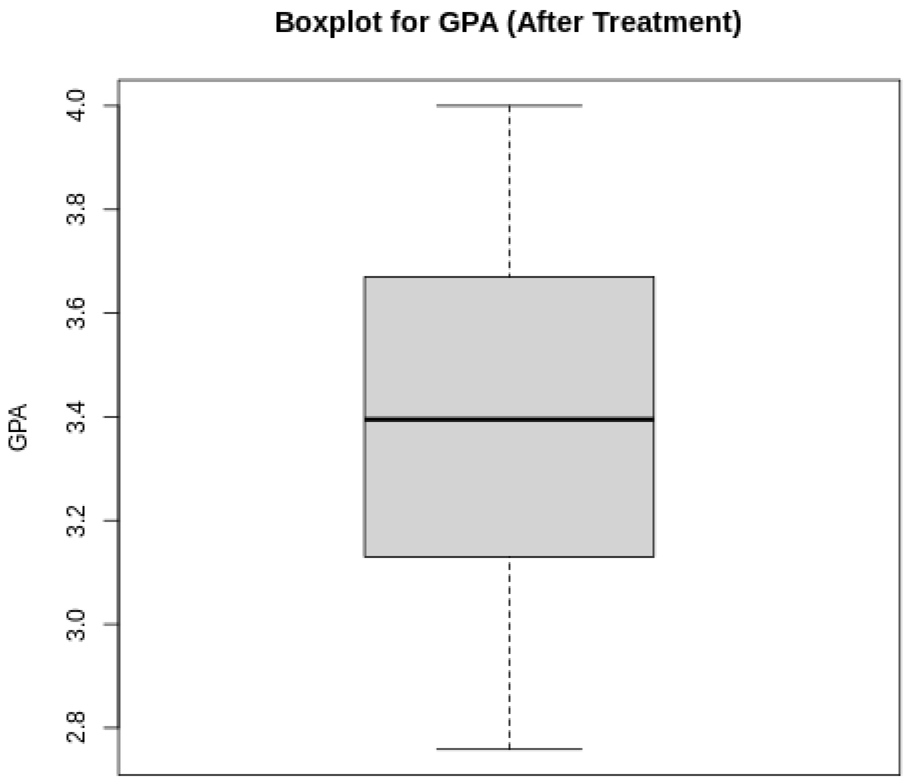

Boxplot for GRE Scores (Before Treatment)
GRE Score
800
700
600
500
400
300
200

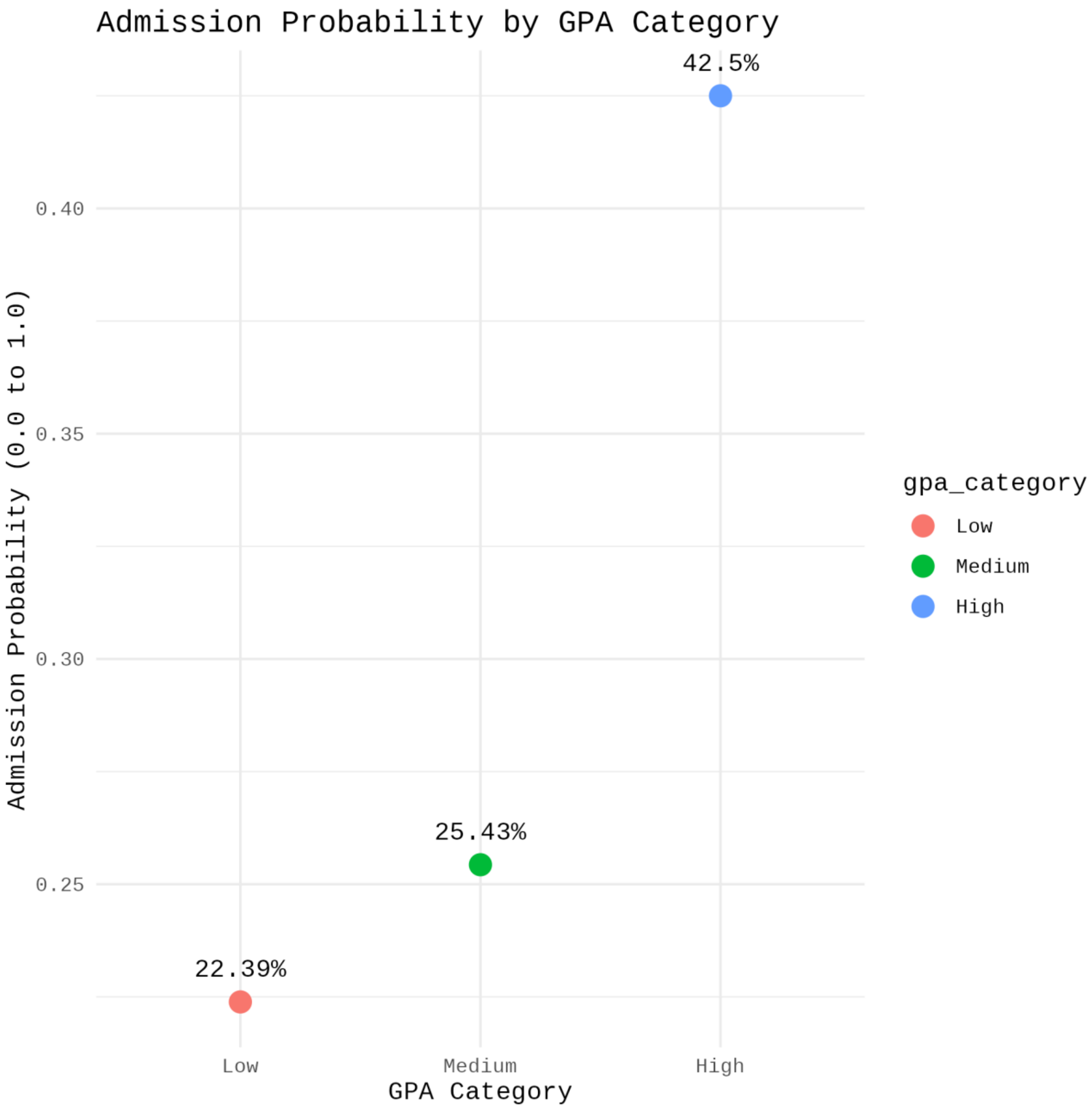

Boxplot for GRE Scores (After Treatment)
GRE Score
800
700
600
500
400


Q-Q Plot for GPA
Sample Quantiles
4.0
3.8
3.6
3.4
3.2
3.0
2.8
Theoretical Quantiles
-3
-2
-1
0
1
2
3

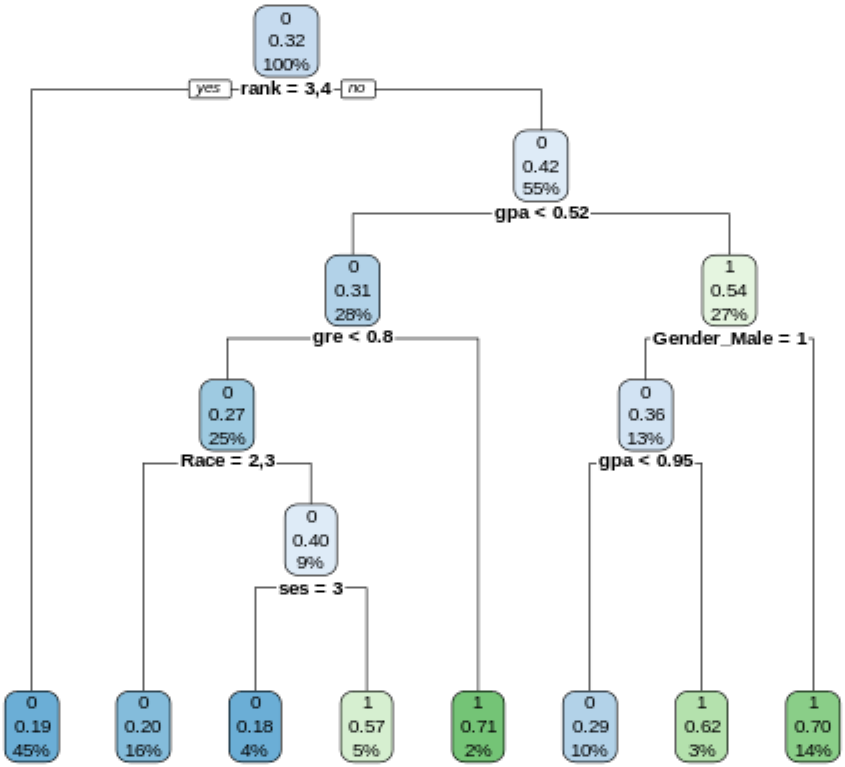

Histogram for GPA
Frequency
0
10
20
30
40
50
2.8
3.0
3.2
3.4
3.6
3.8
4.0
GPA

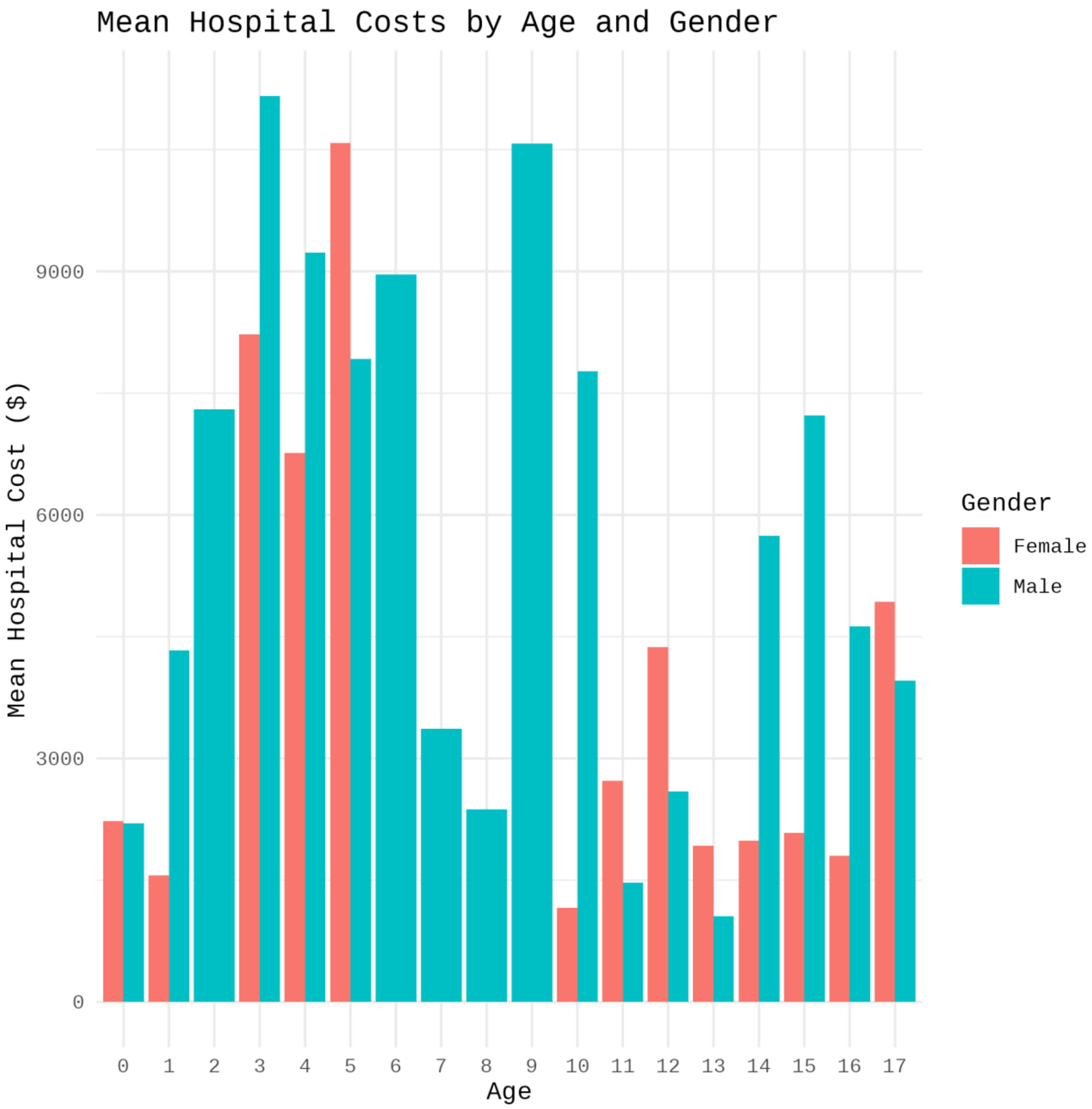

Boxplot for GPA (Before Treatment)
GPA
4.0
3.5
3.0
2.5


Boxplot for GPA (After Treatment)
GPA
4.0
3.8
3.6
3.4
3.2
3.0
2.8


Admission Probability by GPA Category
42.5%
25.43%
22.39%
0.40
0.35
0.30
0.25
Admission Probability (0.0 to 1.0)
Low
Medium
High
GPA Category
gpa_category
Low
Medium
High

Decision Tree for College Admission




Mean Hospital Costs by Age and Gender
Mean Hospital Cost ($)
0
3000
6000
9000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Age
Gender
Female
Male