**Final project**

**The final project involves a case study that demonstrates the methods learned in this course. Each project report must be completed by groups of two to three students. Larger groups are expected to produce more and higher-quality work.**

**Students who do not form a group by November 1st will have to report to the teaching staff and will be randomly assigned. To find a group, use Piazza to let other students know you are looking for a team. Describe what your interests are so that other students with similar interests can join you. Contact the teaching assistants regarding any issues with the group.**

**The grade for the project report will be based on individual contributions, overall presentation, and the significance of the study. The contribution of each group member must be clearly indicated on the first page of the report. The report should be 8-10 pages including writing and visualizations, but excluding code.**

**The project must contain a regression analysis (linear, logistic or more advanced) and a method covered in the third part of the lecture (ROC, CV, PCA, k-NN, k-Means).**

**The rubric is as follows:**

* **Reporting (30%): Are there clear research questions that you asked, and did you address these in an orderly fashion? Did you make well justified conclusions? Is your project sensible and easy to read?**
* **Data Processing (10%): How much work was necessary to get your data and bring it in a format that is useful for further analysis.**
* **Visualization and Methodology (40%): Do your visualizations follow best practices? Do they support the hypothesis? Is your methodology appropriate? Does this give insight to your project? Are the methods tailored to your specific topic and data (not generic or off-the-shelf)?**
* **Code (10%): Is your code well-organized and easy to read? Is your code reproducible? Is your code documented? Is your code reasonably efficient? Did you use appropriate data structures and algorithms?**

**Each category will be graded points according to the table below and weighted appropriately.**

| **Grade** | **Points** |
| --- | --- |
| **Good** | **10** |
| **Satisfactory** | **8** |
| **Poor** | **6** |
| **Partial Work** | **4** |
| **No Work** | **0** |

**Final Project Outline**

**1. Introduction**

• **Research Questions**: Clearly define the research questions or hypotheses being addressed.

• **Objective**: State the purpose and significance of the study.

• **Overview of Methods**: Briefly outline the regression analysis and the advanced method from the course (e.g., ROC, CV, PCA, k-NN, k-Means).

**2. Data Processing**

• **Data Acquisition**: Describe the dataset, including its source, relevance, and scope.

• **Data Cleaning**: Detail the steps taken to clean and preprocess the data (handling missing values, outliers, etc.).

• **Data Formatting**: Explain any transformations or formatting required to prepare the data for analysis.

**3. Methodology**

• **Regression Analysis**:

• Specify the type of regression used (linear, logistic, etc.).

• Justify its suitability for the research questions.

• Explain the implementation and key variables analyzed.

• **Advanced Method**:

• Detail the chosen advanced method (e.g., PCA, k-NN, etc.).

• Justify its relevance to the problem.

• Describe how the method was applied and interpreted.

**4. Results**

• **Visualizations**:

• Present graphs, plots, or charts to illustrate key findings.

• Ensure visualizations follow best practices and support the hypotheses.

• **Statistical Insights**:

• Report and interpret key results from the regression and advanced method analyses.

• Highlight any significant patterns or relationships.

**5. Discussion**

• **Conclusion**:

• Summarize the findings in relation to the research questions.

• Discuss the implications of the results.

• **Limitations**:

• Address any limitations in the study design or data.

• Suggest areas for further research.

**6. Contributions**

• Clearly state the contributions of each group member on the first page:

• Data processing

• Analysis

• Visualization

• Report writing

• Code implementation

**7. Appendix**

• **Code**: Include well-organized and documented code.

• **Additional Figures**: Present supplementary visualizations if needed.

• **Data Sources**: Provide links or citations for datasets used.

**Grading Focus**

• **Reporting (30%)**: Clear research questions, orderly analysis, and well-justified conclusions.

• **Data Processing (10%)**: Detailed and efficient handling of data.

• **Visualization & Methodology (40%)**: Best-practice visuals, appropriate and insightful methods tailored to the topic.

• **Code (10%)**: Readable, reproducible, efficient, and well-documented code.

Ensure the report is 8-10 pages (excluding code), easy to read, and demonstrates a cohesive and high-quality team effort.

1. **Introduction**

Research Questions:

Can historical admission data be used to predict the probability of acceptance for future applicants? What are the key features that contribute most significantly to predicting graduate admission chances?

Objective:

The goal of this study is to develop a predictive model for graduate admissions using historical data. By analyzing past admissions, we aim to understand the influence of different factors such as GRE Score, TOEFL Score, CGPA, and other application elements, and to use this information to predict future admission outcomes.

Overview of Methods:

We will perform regression analysis to create a model that predicts the "Chance of Admit" based on key features in the dataset. Additionally, advanced techniques such as Principal Component Analysis (PCA) will be used for dimensionality reduction, and k-Nearest Neighbors (k-NN) will be employed to explore the relationships between applicants with similar profiles.

**2. Data Processing**

**Data Acquisition**:

The dataset used in this study is sourced from Kaggle (https://www.kaggle.com/datasets/mohansacharya/graduate-admissions). It contains information about graduate admissions from a variety of applicants, with features including GRE Score, TOEFL Score, University Rating, Statement of Purpose (SOP) strength, Letter of Recommendation (LOR) strength, CGPA, research experience, and the corresponding "Chance of Admit". The dataset consists of 500 entries, each representing an applicant, and provides a comprehensive overview of factors influencing admission decisions.

**Data Cleaning**:

**2. Data Processing and Analysis**

The dataset was loaded into R and transformed to a long format for easier analysis. Missing values were removed, and outliers in CGPA were capped between the 5th and 95th percentiles to minimize their impact. Categorical variables such as University Rating and Research were converted into factors, and duplicate rows were removed to ensure data quality. These steps helped prepare the dataset for effective statistical modeling and exploratory analysis.

To ensure data quality, the following data cleaning steps were performed:

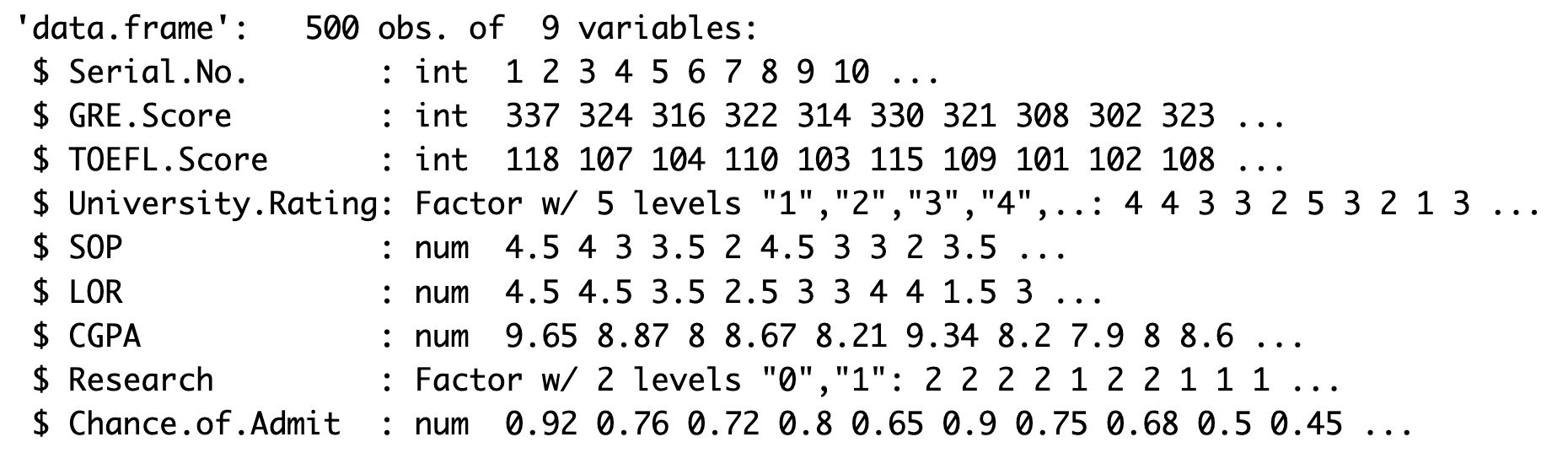
1. **Missing Values**:  
   The dataset was checked for missing values using sum(is.na(data)) and colSums(is.na(data)). Any rows containing missing values were removed using the na.omit() function, resulting in a cleaned dataset (data\_clean).
2. **Outlier Detection and Treatment**:  
   A boxplot was used to visualize potential outliers in the CGPA variable (boxplot(data$CGPA, main = "Boxplot for CGPA")). Outliers were defined as values below the 5th percentile or above the 95th percentile. These outliers were capped to the 5th and 95th percentile values, respectively, to reduce their impact on the analysis (data$CGPA[data$CGPA < qnt[1]] <- qnt[1] and data$CGPA[data$CGPA > qnt[2]] <- qnt[2]).
3. **Duplicate Rows**:  
   Duplicate rows were identified using sum(duplicated(data)) and removed to create a final clean dataset without redundancy (data\_clean <- data[!duplicated(data), ]).

**Data Formatting**:  
The data was transformed into a long format using the pivot\_longer() function from the tidyrpackage. This transformation makes it easier to perform exploratory analysis and visualize trends for each variable.The long format dataset allows for detailed exploration of the relationships between each feature and the "Chance of Admit". This transformation is particularly useful when performing visualizations or comparing distributions across different groups, such as university ratings or research experience.

* 1. **Categorical Variables**:  
     The University Rating and Research columns were converted into factors, as these variables represent categorical data. By converting these variables into factors, we ensured that statistical modeling and visualization would appropriately treat them as categorical rather than continuous.

**Structure of data:**

The structure of the transformed dataset was shown using the str() function. This verification ensures that each variable is properly classified, either as numeric or categorical, which is critical for correct analysis and modeling.



**Categorical Data Analysis**

A close-up of a white paper

Description automatically generatedA comparison of a bar graph

Description automatically generated with medium confidence

Fig1. Frequency Distribution

The frequency distribution of the variable University Rating was calculated, the output indicates the count of applicants in each of the five university rating categories (1 to 5). We can see that most students are from undergraduate universities that in rank 2-4.

The frequency distribution for the variable Research is showing the count of applicants who had research experience (Research = 1) versus those who did not (Research = 0). We can see that, applicants with or without research experiment is quiet even.

Fig2. Bar Plot

The left bar plot visually represents the relationship between the University Rating and the Chance of Admit. From the plot, it is evident that higher university ratings tend to correspond to higher chances of admission, implying that applicants from higher-rated universities are more likely to be admitted.

The right boxplot of Research versus Chance of Admit shows that applicants with research experience generally have a higher probability of admission compared to those without research experience. This suggests that having research experience is a valuable asset when applying for admissions.

**Numerical Data Analysis**

**Scatter Plot**

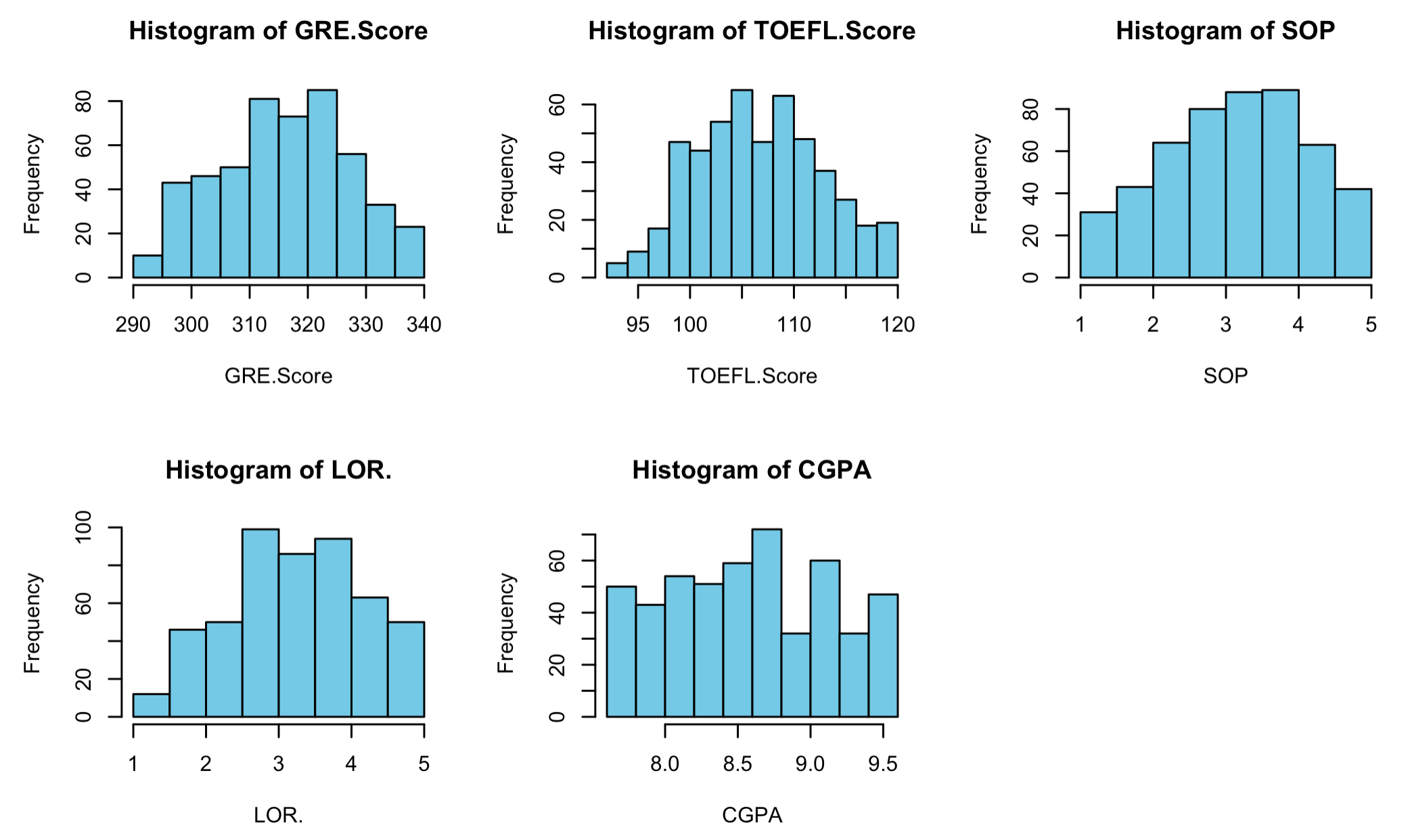
A group of graphs showing different colored points

Description automatically generated with medium confidence

The plots for numerical variables vs. Chance of Admit were generated to visualize the relationships between key variables and admission chances.

The scatter plots indicate that GRE Score, TOEFL Score, SOP, LOR, and CGPA are all positively correlated with Chance of Admit, with CGPA showing the strongest effect. Higher scores in each of these variables generally lead to an increased likelihood of admission, highlighting the importance of strong academic performance, compelling statements of purpose, and good letters of recommendation in the admissions process.

**Histogram**



The distributions of “GRE Score” and “TOEFL Score” are roughly normal, which suggests that these standardized test scores are common metrics that applicants perform similarly on. Both “SOP” and “LOR” are concentrated towards higher ratings, reflecting that most applicants are strong in terms of these subjective measures. “CGPA” has a fairly uniform distribution, but still indicates a high level of academic achievement among applicants.

A graph with red squares

Description automatically generated

The correlations suggest that strong academic credentials (measured by GPA, GRE, and TOEFL scores) tend to be associated with strong application elements like SOP and LOR, contributing positively to the applicant’s overall profile. However, moderate correlations for SOP and LOR also indicate variability, suggesting that these qualitative aspects may depend on factors beyond academic metrics alone.

1. **Model Selection**

From analyzing on each independent variables, we conclude that each independent variables does shown an positive correlation with the admission rate, which is a strong indication for linear regression to see relation between admission rate and these factors.

Linear Regression:

When doing linear regression, we used training and testing datasets in the regression serves to evaluate how well the model generalizes to unseen data, ensuring that it doesn’t overfit or underfit.

A screenshot of a computer

Description automatically generated

This full model was used to predict admission chances using GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, and Research experience. From the output, we figured out that Significant predictors include GRE Score (p = 0.00144), TOEFL Score (p = 0.00883), LOR (p = 0.00100), CGPA (p < 2e-16), and Research experience (p = 0.00271), indicating positive relationships with admission chances. CGPA was the most influential factor. University Rating and SOP were not statistically significant.

In addition, the model's test R-squared value is 0.900757, meaning approximately 90.08% of the variance in the Chance of Admission is explained by the model, indicating a good fit. The Mean Squared Error (MSE) is 0.0023, suggesting that the model's predictions are reasonably accurate.

Thus, we decided to implement the reduced linear regression model that excludes the insignificant variables (University Rating and SOP) and do linear regression again.

A screenshot of a computer

Description automatically generated

The reduced linear regression model was used to predict admission chances based on GRE Score, TOEFL Score, LOR, CGPA, and Research experience. All predictors were found to be statistically significant, with p-values less than 0.01, indicating strong relationships with admission chances. GRE Score (p = 0.00111), TOEFL Score (p = 0.00611), LOR (p < 0.0001), CGPA (p < 2e-16), and Research experience (p = 0.00269) all had positive coefficients, implying that higher values in these factors are associated with increased chances of admission. The model's test R-squared value is 0.90056, suggesting that 90.06% of the variance in admission chances is explained by these variables. In addition, comparing with full model, the R^2 only reduce slightly amount, saying the reduced model still explain the admission rate well. The Mean Squared Error (MSE) is 0.0023, indicating reasonably accurate predictions. Overall, the reduced model is effective in predicting admission chances, with CGPA being the most influential factor.

Based on these results, we are quite satisfied with the reduced model and have decided to use it as the final model for predicting admission chances, as it balances simplicity and predictive performance effectively. The final model is expressed as follows:

Chance of Admit = -1.3925 + 0.001944 \* GRE\_Score + 0.002383 \* TOEFL\_Score + 0.017759 \* LOR + 0.130684 \* CGPA + 0.023283 \* Research