# STAT403 Final Report: Analysis of Graduate School Admission

Gefei Shen, Zhikai Li, Tao Zhang

June 2024

## Introduction

Nowadays, it is common that gaining admission to a graduate program is highly competitive, thus understanding the key factors that influence acceptance of graduate program from different university has become a very popular focus among students. Our project aims to delve into this critical issue by analyzing a dataset that highly related to predict graduate admissions from undergraduate students perspective.

Our dataset is sourced from Kaggle, and it comprises information on 401 applicants from Indian Perspective, including eight different variables that may influence the likelihood of admission. These variables include GRE Score, TOEFL Score, University Rating, Statement of Purpose (SOP), Letter of Recommendation Strength (LOR), Undergraduate GPA (CGPA), Research, and the Chance of Admit. For the GRE Score, TOFEL Score, CGPA and Chance of Admit, they are numerical variables. For University Rating, Statement of Purpose (SOP), Letter of Recommendation Strength (LOR), and Research are catergorical variables. For categorical variables like the Letter of Recommendation (LOR) and Statement of Purpose (SOP), which are rated on a scale from 0 to 5 as higher means the more recommendation letters and the better the university ranking. Thus we want to explore these variables to find their relationship and impact on the Chance of Admit.

The first focus of our analysis is to determine which factors most significantly affect the likelihood of being admitted to a graduate program. Our first research question investigates the relative importance of various academic and personal factors, such as GRE scores, university ratings, and LORs, in the admissions process. By identifying these key factors, we aim to provide a clear understanding of the criteria that admissions committees consider critical in their evaluations. This insight can be important for prospective students in enhancing their profiles and improving their chances of acceptance.

Our second research question delves into the role of CGPA in influencing admission chances across universities with different ratings. Since the GPA are usually known as the most important factors for academic evaluations, it is crucial for us to understand how it interacts with the prestige and quality of different universities. We will explore whether a high CGPA consistently boosts admission probabilities regardless of the university rating, or if its impact varies significantly based on the institution's ranking. We think this analysis aims can help undergraduate students have a better understanding of importance of their CGPA for graduate admission.

# Methods

## Question 1

To analysis which factors would be most significantly affect the chance of be admitted and build our prediction model. We first give a overview about the relationship between chance of being admitted and all other predictors using pair() function. Then, we decompose the factors into the numerical variables (e.g TOEFL score, GRE score, etc.) and categorical variables (e.g letter of recommendation, research, etc.). Firstly, we make the correlation plot to find the correlation between numerical variables and chance of being admitted. Then, we use anova test to find the correlation between categorical variable and chance of being admitted. After that we can have know all features that have significantly relationship between them and the Change of Admit, we can use them as parameters for our prediction model. At this point, we can already finish our model, but we also want to use cross-validation as our resampling-based approach to see that whether our model can be improved. Since it can provide a more accurate estimate of model performance by reducing overfitting especially when our sample size is not extremely large.

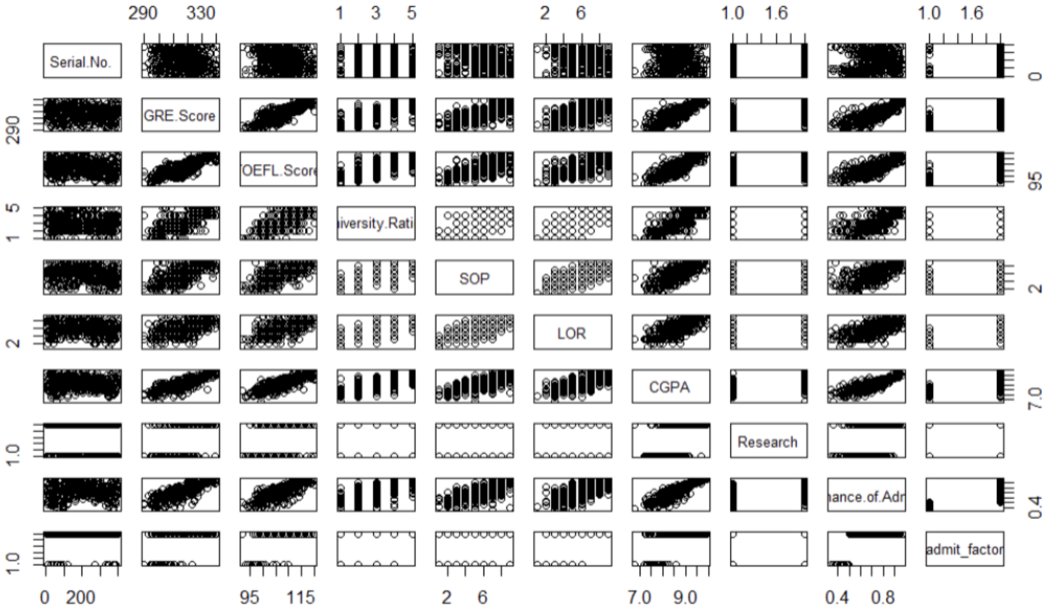
## Question 2

Given a specific ranking of the university, we choose logistic regression model to compare the coefficient of the CGPA. From the coefficient, we could find the weights of CGPA across the universities with different ranking. The larger coefficient means CGPA weight more. Then, we use the Kernel Density Estimator to give a distribution of CGPA over each university rankings. Lastly, we would draw a plot about the interaction term between university ranking and CGPA versus the probability of admission, and we will be able to provide conclusion about how does CGPA influence admission chances across universities with varying ratings.

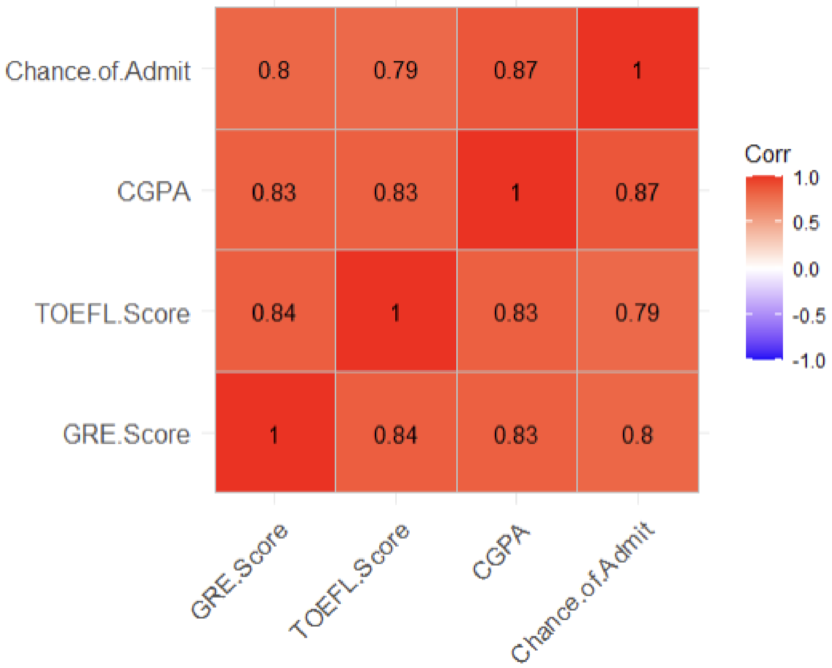
# Results

## Question 1

Here are the results of our survey using correlation for all variables.

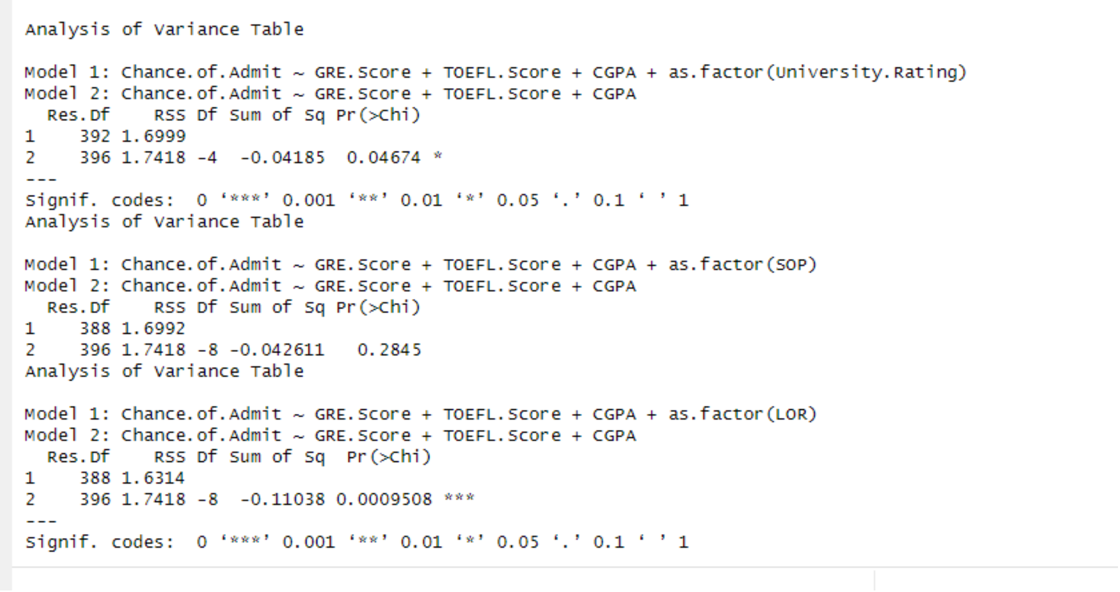


Correlation is used to determine whether there is a relationship between numerical variables.
In this case, the predictor variable: Chance of Admit is a numerical variable.
After that, we convert the Chance of Admit into categorical variables. When it's greater than 0.5, it's 1,0 otherwise.
In the comparison between numerical variables and numerical variables, we find that they are all related. This means that the predictors of all the numerical variables we selected can be used.

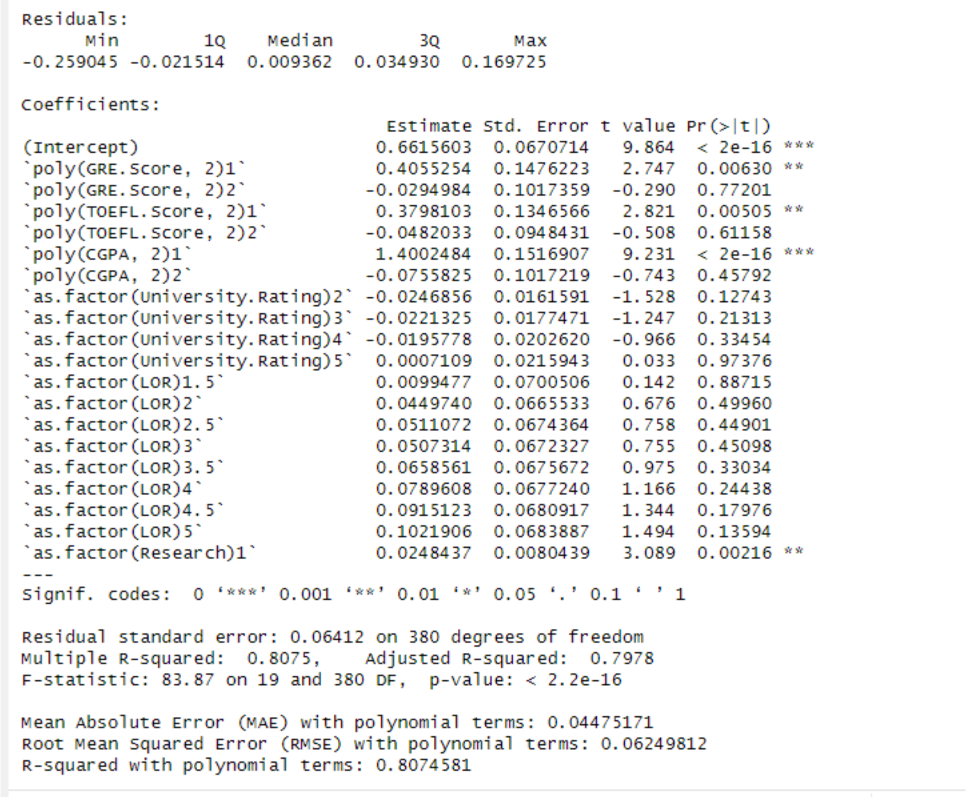


We use all available numerical variables as predictors, and then add a categorical variable in turn to form a new categorical variable.

We use the anova test to determine whether this categorical variable can be added to the prediction model. In the end, we found that all variables other than SOR could be used, which means all of these categorical features except SOR contribute significant influence to the Change of Admit.



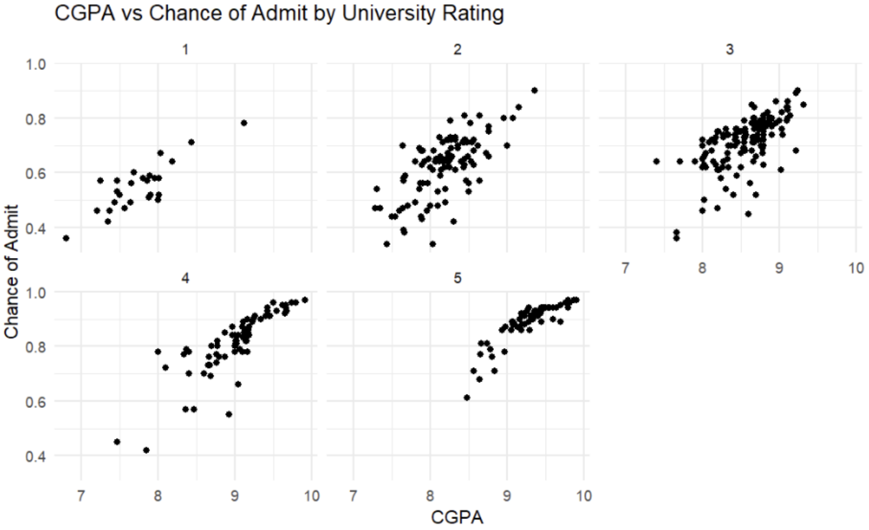
```
Analysis of Variance Table

Model 1: Chance.of.Admit ~ GRE.Score + TOEFL.Score + CGPA + as.factor(University.Rating)
Model 2: Chance.of.Admit ~ GRE.Score + TOEFL.Score + CGPA
  Res.Df    RSS Df Sum of Sq Pr(>Chi)
1    392 1.6999
2    396 1.7418 -4  -0.04185  0.04674 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Analysis of Variance Table

Model 1: Chance.of.Admit ~ GRE.Score + TOEFL.Score + CGPA + as.factor(SOP)
Model 2: Chance.of.Admit ~ GRE.Score + TOEFL.Score + CGPA
  Res.Df    RSS Df Sum of Sq Pr(>Chi)
1    388 1.6992
2    396 1.7418 -8 -0.042611   0.2845
Analysis of Variance Table

Model 1: Chance.of.Admit ~ GRE.Score + TOEFL.Score + CGPA + as.factor(LOR)
Model 2: Chance.of.Admit ~ GRE.Score + TOEFL.Score + CGPA
  Res.Df    RSS Df Sum of Sq  Pr(>Chi)
1    388 1.6314
2    396 1.7418 -8  -0.11038 0.0009508 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



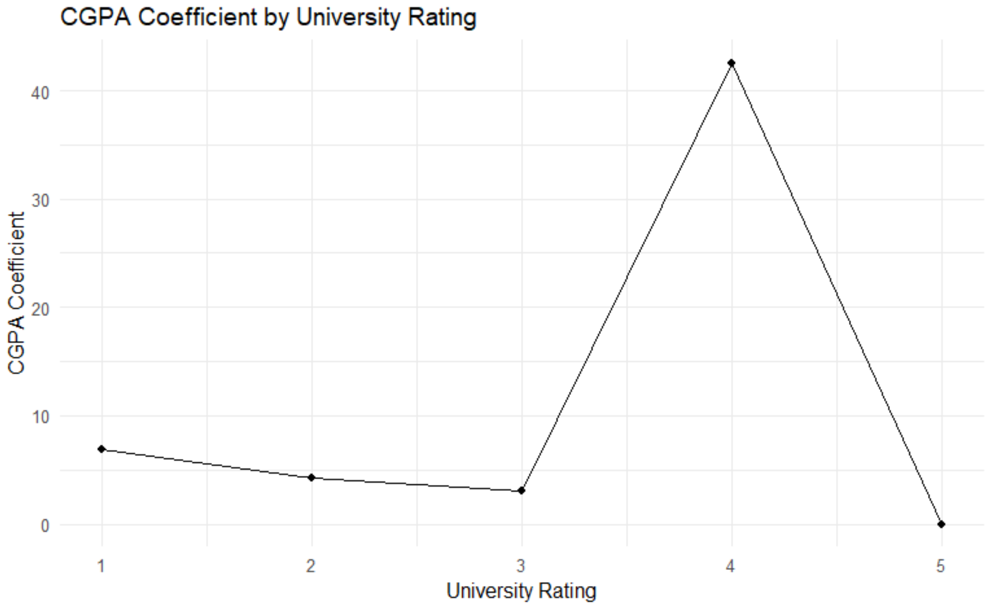
```
Residuals:
      Min        1Q    Median        3Q       Max
-0.259045 -0.021514  0.009362  0.034930  0.169725

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      0.6615603  0.0670714   9.864  < 2e-16 ***
`poly(GRE.Score, 2)1`            0.4055254  0.1476223   2.747  0.00630 **
`poly(GRE.Score, 2)2`           -0.0294984  0.1017359  -0.290  0.77201
`poly(TOEFL.Score, 2)1`          0.3798103  0.1346566   2.821  0.00505 **
`poly(TOEFL.Score, 2)2`         -0.0482033  0.0948431  -0.508  0.61158
`poly(CGPA, 2)1`                 1.4002484  0.1516907   9.231  < 2e-16 ***
`poly(CGPA, 2)2`                -0.0755825  0.1017219  -0.743  0.45792
`as.factor(University.Rating)2` -0.0246856  0.0161591  -1.528  0.12743
`as.factor(University.Rating)3` -0.0221325  0.0177471  -1.247  0.21313
`as.factor(University.Rating)4` -0.0195778  0.0202620  -0.966  0.33454
`as.factor(University.Rating)5`  0.0007109  0.0215943   0.033  0.97376
`as.factor(LOR)1.5`              0.0099477  0.0700506   0.142  0.88715
`as.factor(LOR)2`                0.0449740  0.0665533   0.676  0.49960
`as.factor(LOR)2.5`              0.0511072  0.0674364   0.758  0.44901
`as.factor(LOR)3`                0.0507314  0.0672327   0.755  0.45098
`as.factor(LOR)3.5`              0.0658561  0.0675672   0.975  0.33034
`as.factor(LOR)4`                0.0789608  0.0677240   1.166  0.24438
`as.factor(LOR)4.5`              0.0915123  0.0680917   1.344  0.17976
`as.factor(LOR)5`                0.1021906  0.0683887   1.494  0.13594
`as.factor(Research)1`           0.0248437  0.0080439   3.089  0.00216 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06412 on 380 degrees of freedom
Multiple R-squared:  0.8075,    Adjusted R-squared:  0.7978
F-statistic: 83.87 on 19 and 380 DF,  p-value: < 2.2e-16

Mean Absolute Error (MAE) with polynomial terms: 0.04475171
Root Mean Squared Error (RMSE) with polynomial terms: 0.06249812
R-squared with polynomial terms: 0.8074581
```

After we know all important factors, We used cross validation to divide the sample into ten parts for training. The summary output of a linear regression model shows significant predictors above: GRE Score, TOEFL Score, and CGPA (all polynomial terms), and Research experience. Key metrics include a Residual Standard Error of 0.06412, an R-squared of 0.8075, and an Adjusted R-squared of 0.7978.
We can also see that the Mean Absolute Error (MAE) is 0.04475171, and the Root Mean Squared Error (RMSE) is 0.06249812, indicating a reasonably good fit with polynomial terms included. Finally, we used the model we built to predict the value. With a tolerance of 0.05, the accuracy rate is about 68%. With a tolerance of 0.1, the accuracy rate is about 90%.

## Question 2

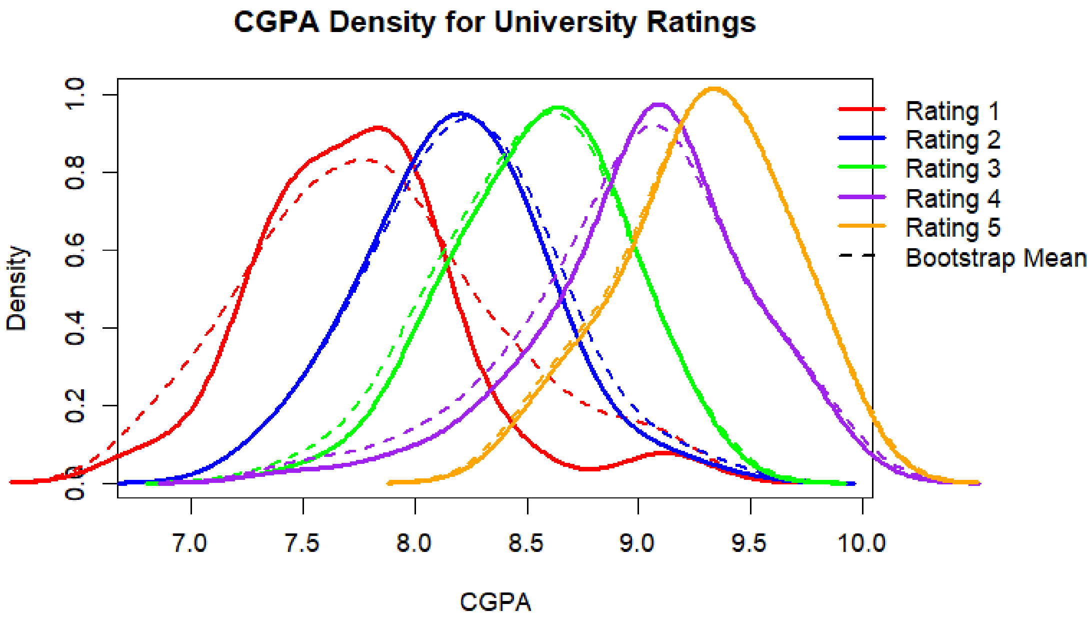


We first investigate whether CGPA is still associated with Chance of Admit at different college levels. We can see from our result that for all universities with different University Rating, the CGPA are positively related to Chance of Admit.

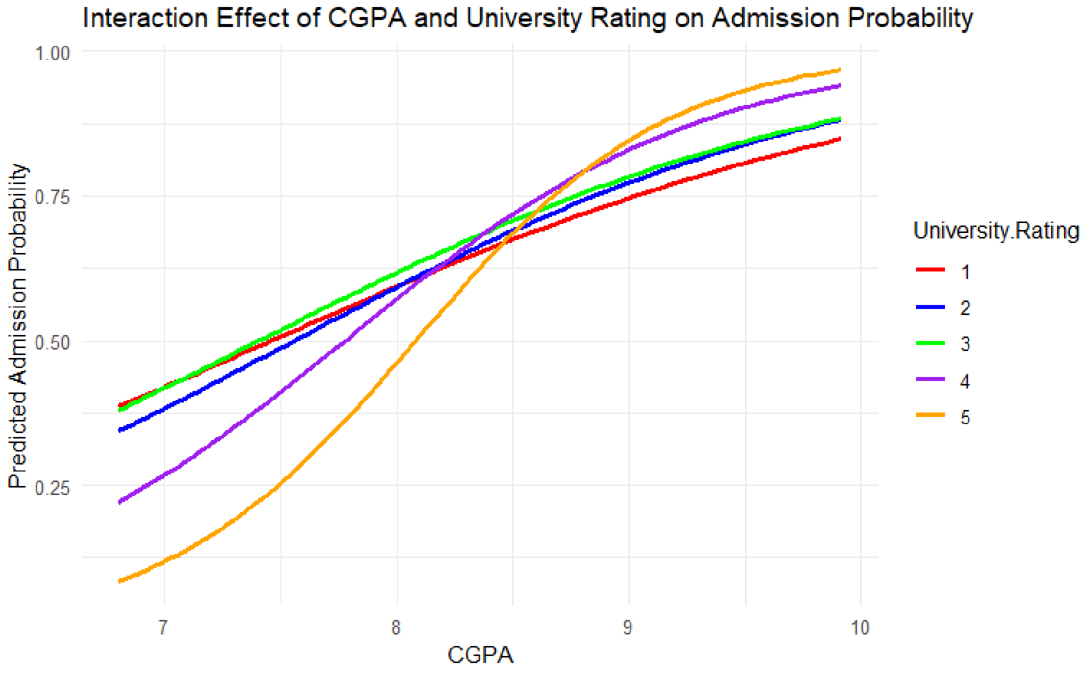


Then we utilizing logistic regression, we first extract the coefficients of CGPA. We can see that from the University Rating 1 to 3, the CGPA coefficient tend to decrease, and when the University Rating is at 4, the CGPA coefficient become the highest. After the University Rating increase to 5, the CGPA coefficient decrease a lot and it become the lowest coefficient compares to all previous coefficients. From these results, we can conclude that with the increasing of university rankings, the impact of changes in CGPA on admission is gradually decreasing, except for the universities that have University Rating of 4. For universities at Rating 4, the level of CGPA for different applicants plays a crucial role in whether they can be admitted.

We have summarized this phenomenon into a logic: for universities with Ratings 1, 2, and 3, the CGPA used by students when applying can have a significant impact, but for universities with Ratings 4, changes in their CGPA levels can have a significant impact. This may be because the CGPA used by students applying to universities within this range is diverse, and they also tend to prefer using CGPA to determine whether a student is suitable. For the top tier universities, changes in CGPA have little impact on admissions because these types of universities often care more about content beyond CGPA than other ranked universities. However, it is worth noting that this does not mean that top universities do not care about their students' GPA. The occurrence of this phenomenon is more likely due to the low diversity and high CGPA requirements of students applying to such universities, and these universities still have strict CGPA requirements for students, which is reflected in our analysis below.



Then we use Gaussian kernel to generate the plot above. We could find it makes sense that as the rating of university increase, the expectation or the average of the CGPA increases. To be more specific, when students apply university with rating 1, the average CGPA is roughly 7.8 and when students apply university with rating 5, their average CGPA is roughly 9.5. This indicates that the higher the ranking score of a university, the higher the CGPA requirements for admitted students.



From our plot of interaction effect between CGPA and Admission Probability, we can see that for universities with lower and medium rankings, such as rating of 1, 2, and 3, an increase in CGPA for applicants can lead to a stable increase in admission probability, while the difference in admission probability between high and low CGPAs is not huge. However, for higher ranked universities such as rating 4 and 5, the increase in CGPA does not result in a linear and stable increase in admission probability. We can see that a low CGPA is very difficult for students to be admited by these universities, and the increase in admission probability from a relatively low CGPA is obvious. After reaching a certain level of CGPA, the improvement starts to become very limited. These phenomena also conform to our previous analysis, and then we can have our conclusion: high ranked or rating universities place great emphasis on students' CPGA, while the CGPA requirements for students in medium and lower rating's universities are relatively less stringent, but for higher rating universities, after reaching a certain level of CGPA, they start to care more about other factors of students, and the increase in admission rate brought about by the simple improvement of CGPA becomes very small. This means that undergraduate students who aim to apply to better universities need to focus on their competitiveness in other aspects, such as TOFEL, GRE scores, or recommendation letters, while maintaining their CGPA.

# Discussion

For problem 2, we use logsitic regression to investgate the relationship between CGPA and university with differnt rating. However, logistic regression assumes that all observations are independent. In our dataset, students within the same school are likely to be more similar to each other than to students from different schools due to shared resources, teaching methods, and environments. Ignoring this dependency can lead to biased estimates and underestimated standard errors. Therefore, I think we could use the Hierarchical Linear model (HLM). HLM is designed to handle nested data structures, where students are nested within universities. It allows for the inclusion of both fixed effects (overall average effects) and random effects (university-specific deviations).

$$\begin{aligned}\text{CGPA}_{ij} &= \beta_0 + \beta_1 \text{University\_Rating}_{ij} + u_j + \epsilon_{ij} \\ u_j &\sim N(0, \sigma_u^2) \\ \epsilon_{ij} &\sim N(0, \sigma^2)\end{aligned}$$

where $\text{CGPA}_{ij}$ is the CGPA of the $i$-th student in the $j$-th university, $\beta_0$ is the intercept term, $\beta_1$ is the fixed effect coefficient of university rating, $\text{University\_Rating}_{ij}$ is the university rating of the $i$-th student in the $j$-th university, $u_j$ is the random effect for the $j$-th university, representing the specific influence of each university, assumed to follow a normal distribution with mean 0 and variance $\sigma_u^2$, and $\epsilon_{ij}$ is the residual error term, representing individual-level differences, assumed to follow a normal distribution with mean 0 and variance $\sigma^2$.

```
data <- data.frame(School_Rating = factor(school_ratings),
    CGPA = cgpa_simulations)
# Fit mixed-effects model
model <- lmer(CGPA ~ School_Rating + (1 | School_Rating), data = data)
summary(model)
```