How to Get into Graduate Program Successfully?

Kaiwen Fu

Department of Statistics

University of Connecticut

May 1, 2023

Abstract

Getting into a graduate program is a significant milestone for students aspiring to pursue higher education. It requires thorough research, planning, and preparation to increase the chances of being accepted into a graduate program successfully. This paper presents a comprehensive statistical analysis on how to get into a graduate program for international students, covering essential factors such as GPA, GRE, TOEFL, SOP, etc. We analyze the relation among the chance of admission and other factors to get the result and provide suggestions. By following the advice presented in this article, prospective students may increase their chances of gaining admission to their desired graduate program and achieve their academic and professional goals.

Keywords: education, graduate program, statistical analysis, chance of admission.

1 Introduction

Why is the graduate program so important for graduates? Historically, graduate education in the United States has played a critical role in the success of the U.S. workforce and

economy, and the link between graduation education and American prosperity has become tighter today (Wendler et al., 2012). Most of graduates are going to apply to a graduate program before they get employed, because there is a positive relation for graduate students between their work ability and solving graduate coursework (Mason et al., 2009). Not only could graduates find a good job position, graduate study is also a training for their emotion capabilities. Jaeger (2003) found that emotional abilities could be developed further in the conventional graduate classroom and there is a significant correlation between emotional intelligence and academic achievement. All of these factors could increase their competitiveness in the job market.

In recent decades, international students enrolled in American institutions of higher education have been increasing and Powers (1990) claimed that the number of non-U.S. graduate students is growing and will continue. There is no doubt that international students play an important role in American society. Chellaraj et al. (2008) shows that if the number of foreign graduate students were to increase by 10%, there would be a corresponding rise of 4.5% in patent applications, 6.8% in university patent grants, and 5.0% in non-university patent grants.

In order to determine the chance of admission for international students, Gupta et al. (2016) considered standardized test scores and GPA as well as university reputations as factors to predict the chance of admission. We are going to involve more factors which were not covered before to analyze the chance of admission, and find the correlation between each variable.

The rest of the paper is organized as follows. The data will be presented in Section 2. Section 3 describes the methods. The results are reported in Section 4. A discussion concludes in Section 5.

Table 1: Summary of variables

	GRE.Score	TOEFL.Score	University.Rating	SOP	LOR	CGPA	Research
Minimum	290.0	92.0	1.0	1.0	1.0	6.800	0
1st quantile	308.0	2.0	2.0	2.5	3.0	8.127	0
Median	317.0	107.0	3.0	3.5	3.5	8.560	1
Mean	316.5	107.2	3.1	3.4	1.0	8.576	0.56
3rd quantile	325.0	112.0	4.0	4.0	1.0	9.040	1
Maximum	340.0	120.0	5.0	5.0	5.0	9.920	1

2 Data Description

This data was collected by Mohan S Acharya (2019) in order to help international students in shortlisting universities with their profiles. In this dataset, there are 500 samples and each sample has 7 independent variables, GRE Scores (out of 340), TOEFL Scores (out of 120), University Rating (out of 5), Statement of Purpose and Letter of Recommendation Strength (out of 5), Undergraduate/College GPA (out of 10), Research Experience (either 0 or 1), and a dependent variable, the chance of admission (out of 1). First, this data does not contain any empty value. Table 1 summarizes some features of the independent variables in our dataset

First of all, GRE Score, TOEFL Scores, Statement of Purpose, Letter of Recommendation strength, College GPA, and Chance of Admission are all continuous variables, but University Rating is categorical variable. Figure 1 shows the distribution of the GRE score, where we can say most students' GRE scores are at range 320 to 325 and only few students get GRE score at 290 to 295. Figure 2 shows the distribution of the TOEFL Score. Students who got 105 are the most and only few students got 90 below at TOEFL test. Figure 3 shows the distribution of the SOP score. Most of students who have SOP score at 2 to 4 and only few students have the score 1 and 5. Figure 4 shows the distribution of the LOR Score. Most of Students' letter of recommendation strength lies in the range 2.5 to 4, and only few students get weak recommendation letter. Figure 5 shows the frequency of university rating for the ordinal variable. We can say that most students are at university rating 3, and few students

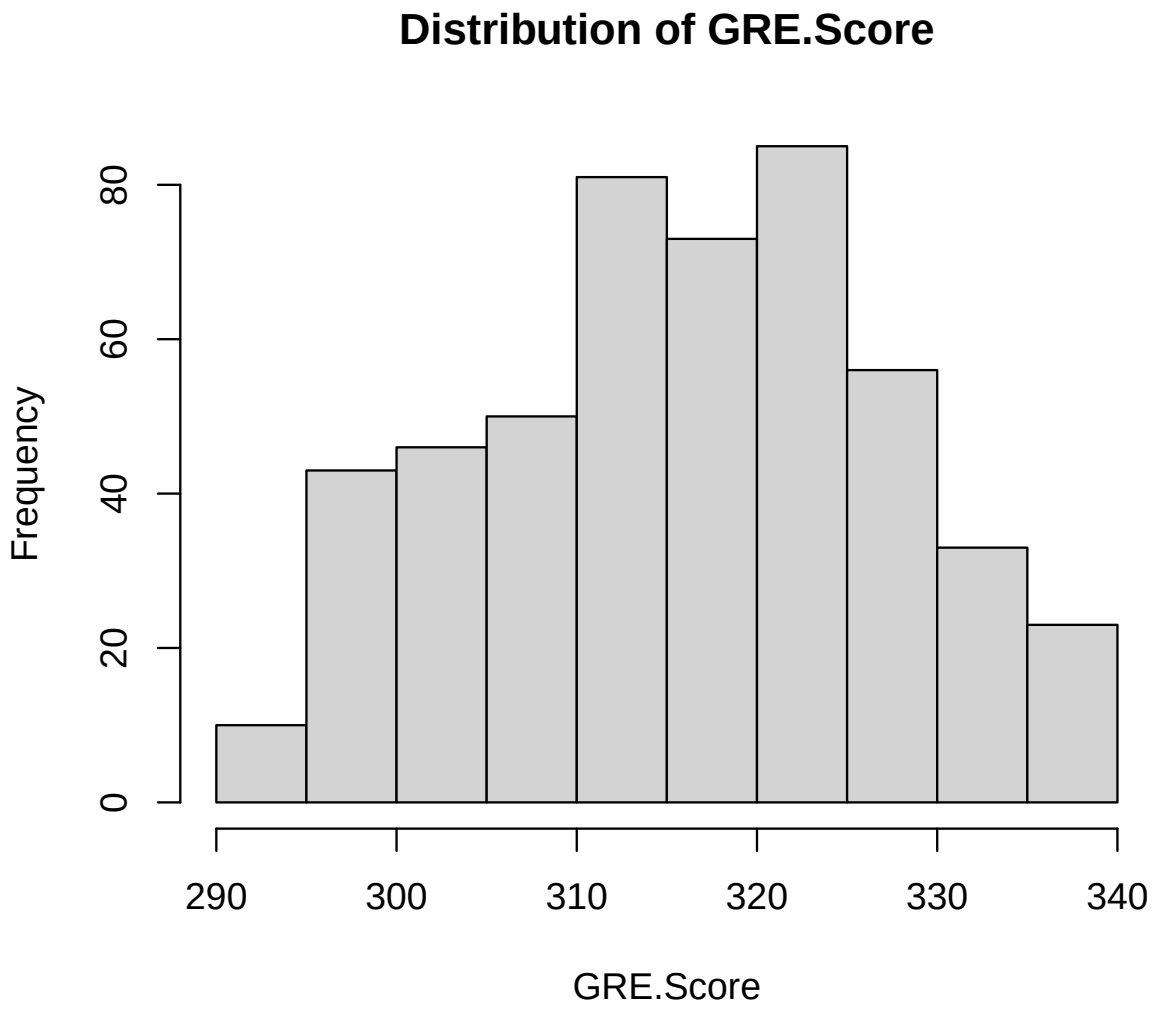

Figure 1: Distribution of GRE Score

from university rating 1. Figure 6 shows the distribution of the CGPA. We can say that the above half of students who are going to apply graduate school have a high college GPA, and only few student have relatively low GPA score. Figure 7 shows the distribution of chance of admission Figure 8 shows the boxplots for GRE score, TOEFL score, University Rating, SOP, LOR, CGPA, and only LOR has the outlier.

Furthermore, we can get the correlation between each factor from Figure 9. Based on the figure, we can say that GRE.Score, TOEFL.Score, and CGPA are highly related and the chance of admission is closed related to GRE.Score, TOEFL.Score, CGPA.

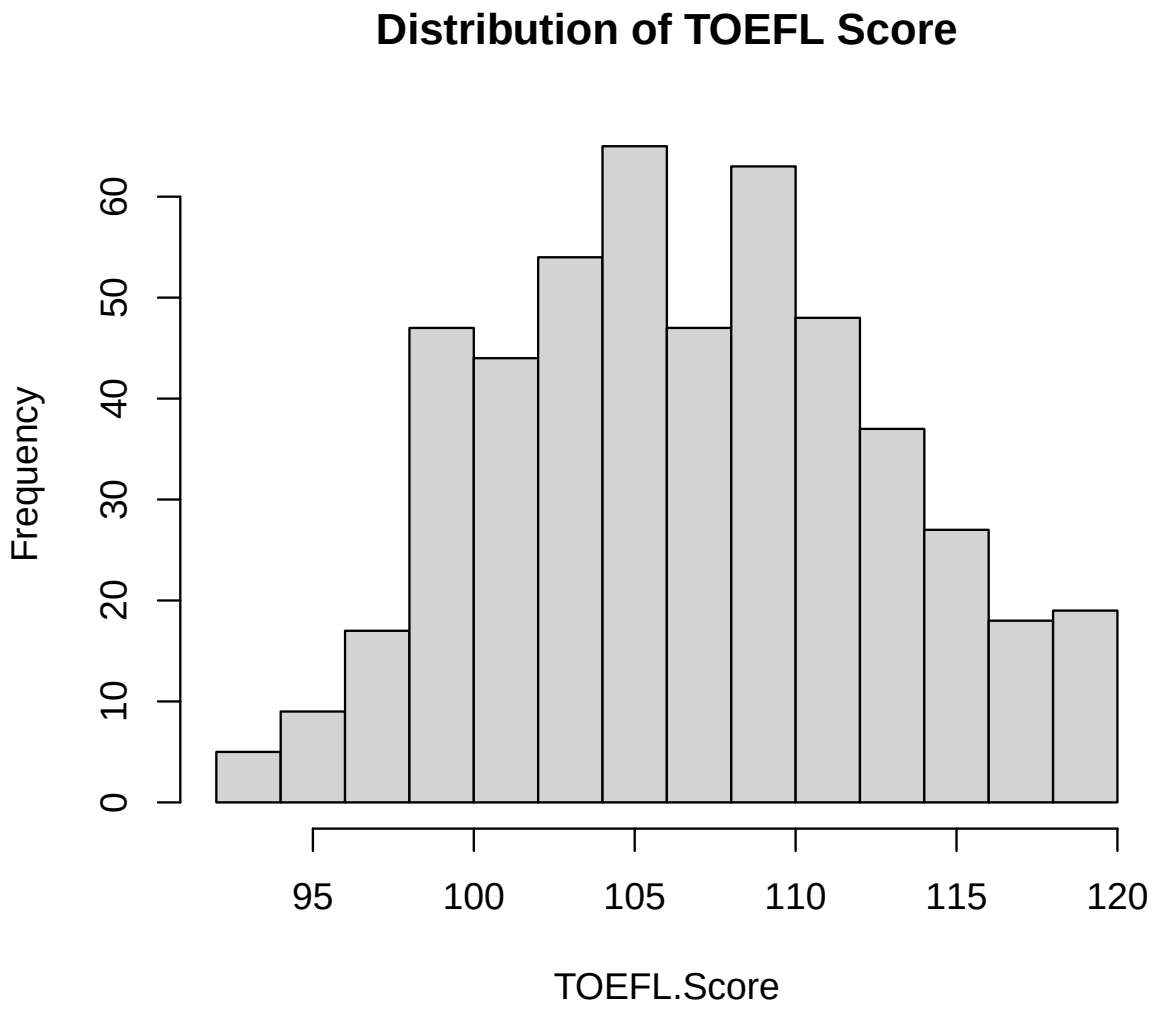

Figure 2: Distribution of TOEFL Score

3 Methods

First, we want to check whether the multicollinearity exists using VIF so that it will affect our analysis. To calculate the VIF, we have the equation

$$VIF = \frac{1}{1 - R_i^2}, \tag{1}$$

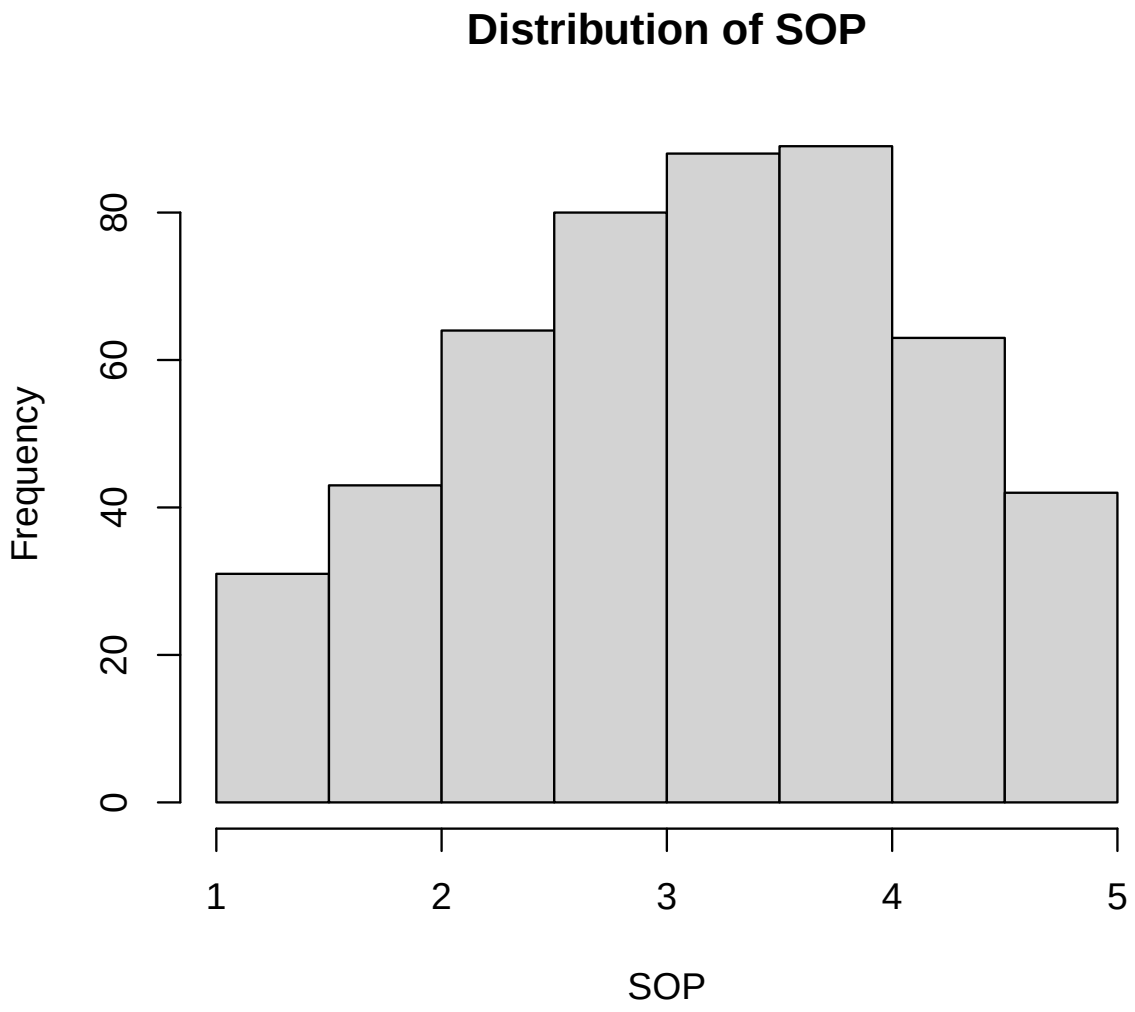

Figure 3: Distribution of SOP

,where R^2 is unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones(Wikipedia contributors, 2023).

First, we are going to use Logistic Regression to analyze the data, and the logistic regression has the form

$$\ln \frac{\pi_i}{1-\pi_i} = \eta_0 + \eta_1 X_1 + \eta_2 X_2 + \eta_3 X_3 + \eta_4 X_4 \eta_5 X_5 + \eta_6 X_6 + \eta_7 X_7, \tag{2}$$

where $\ln \frac{\pi_i}{1-\pi_i}, i \in [1, 7]$ means the log-odds of success, $i = 0$ is the intercept and $\eta_i, i \in$

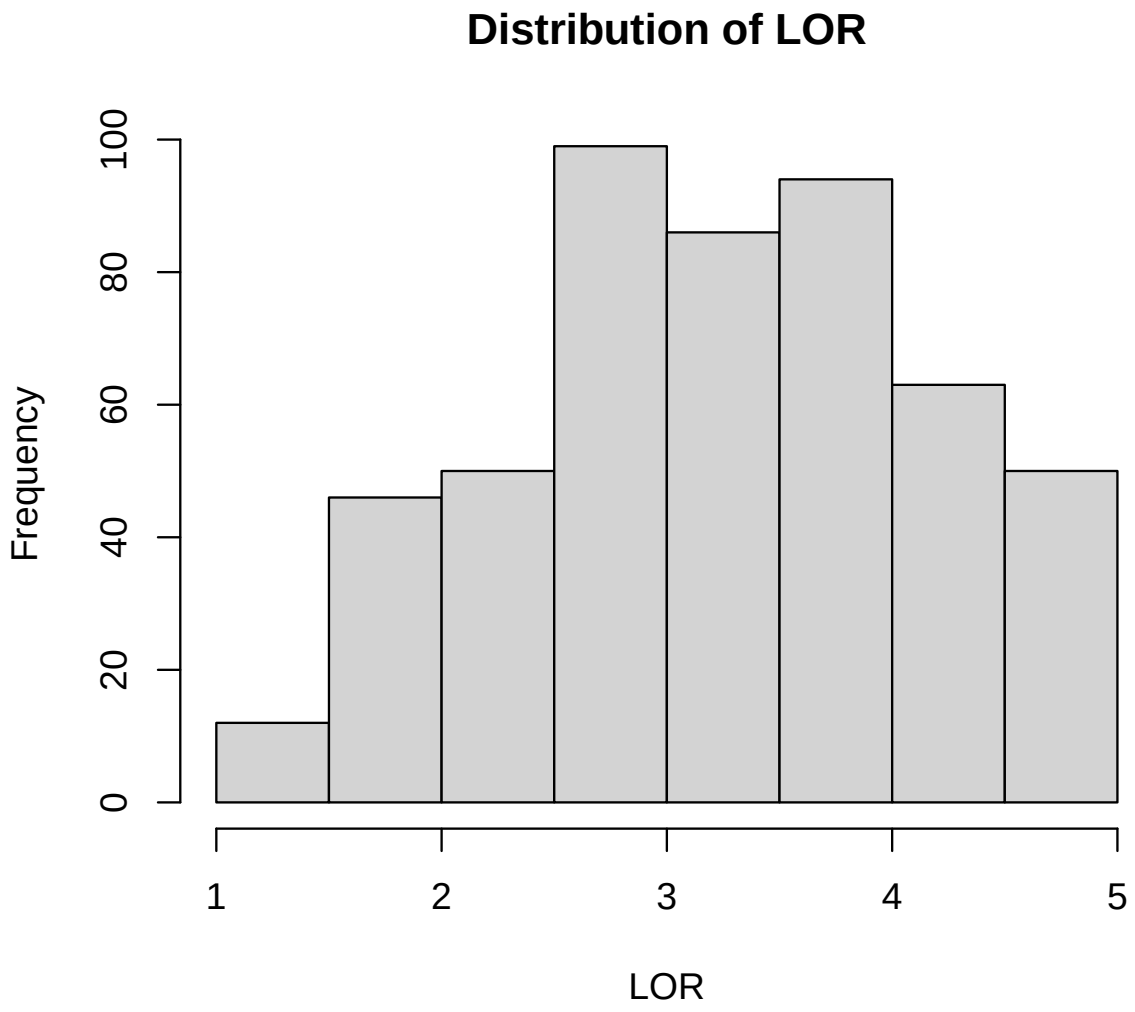

Figure 4: Distribution of LOR

[1, 7] are the coefficients for each X_i. Specifically, X_1 is GRE score, X_2 is TOEFL score, X_3 is University Rating, X_4 is SOP, X_5 is LOR, X_6 is CGPA, and X_7 is Research. Because we do not have complete data whether the students admitted or not, we assume the students who have chance of admission greater and equal 70% are more likely get the offer based on the figure 7, and more likely get rejected whose chance of admission less than 70%. If students' chance of admission greater or equal to 70%, we assign them "Yes", otherwise "No". In this method, we can use it to do the Logistic Regression. So, we have the following assumptions:

The chance of admission is binary.

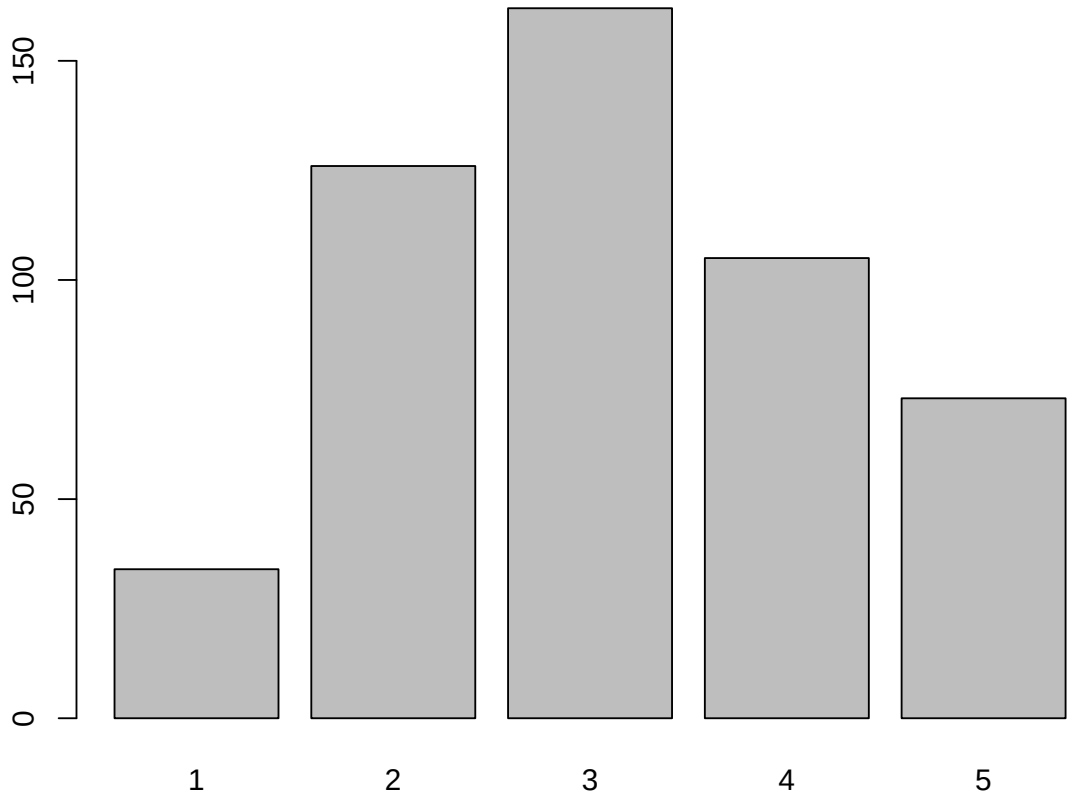

Figure 5: Bar Plot of University Rating

There is no multicollinearity between variables.

There is large sample size.

Each variable are independent.

There is a linear relationship between independent variables and log-odds.

Then, we need to test whether the drop in deviance from the null model to the fitted model is significant and find a more fitted model by compare their AICs.

H_0 : the parameters that two models under comparison do not have in common are all zero

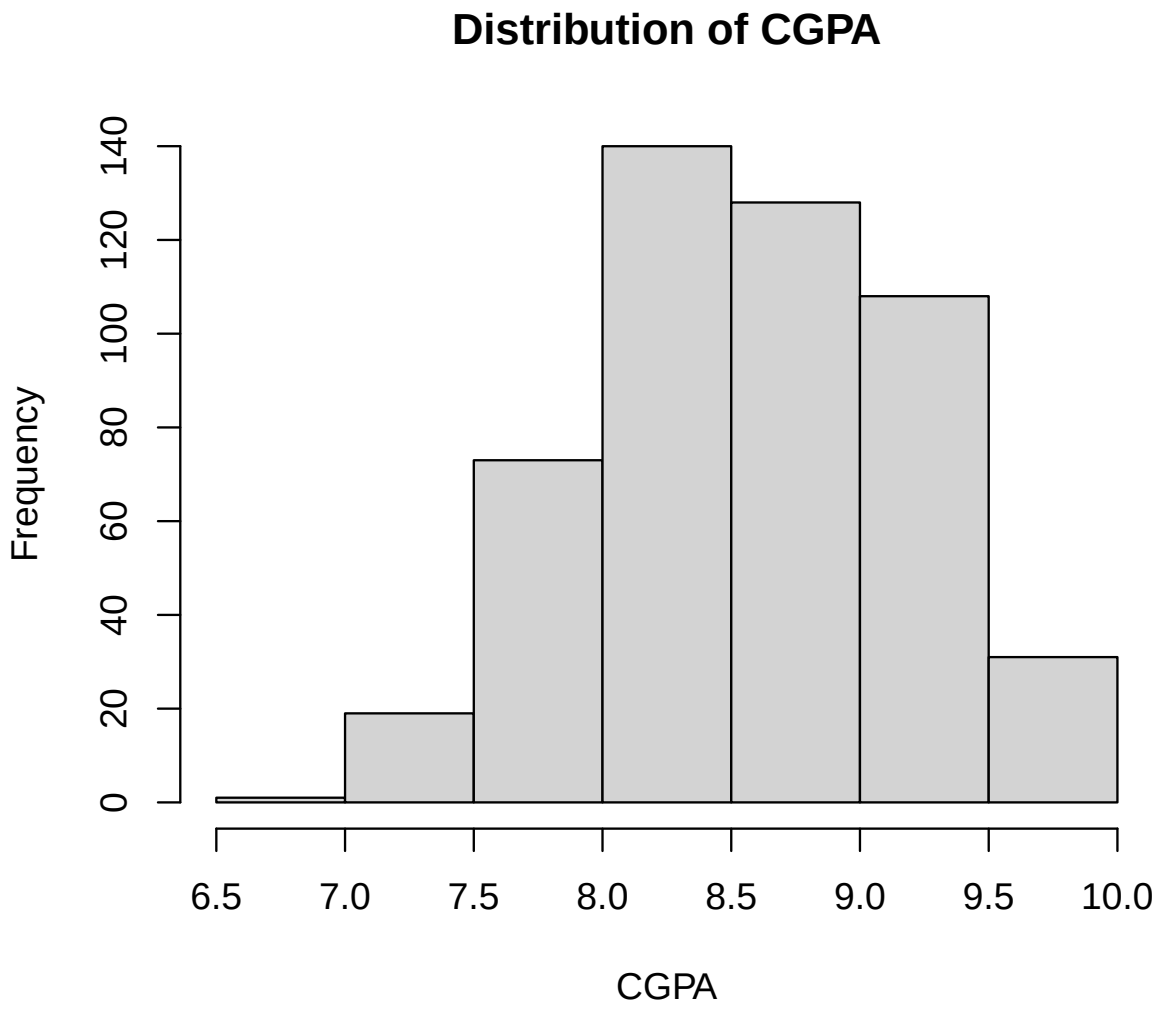

Figure 6: Distribution of CGPA

93

H_a : At lease one parameters parameters are not zero

94 After that, we need to check assumptions we listed above logistic regression using significant

95 level $\alpha = 0.05$.

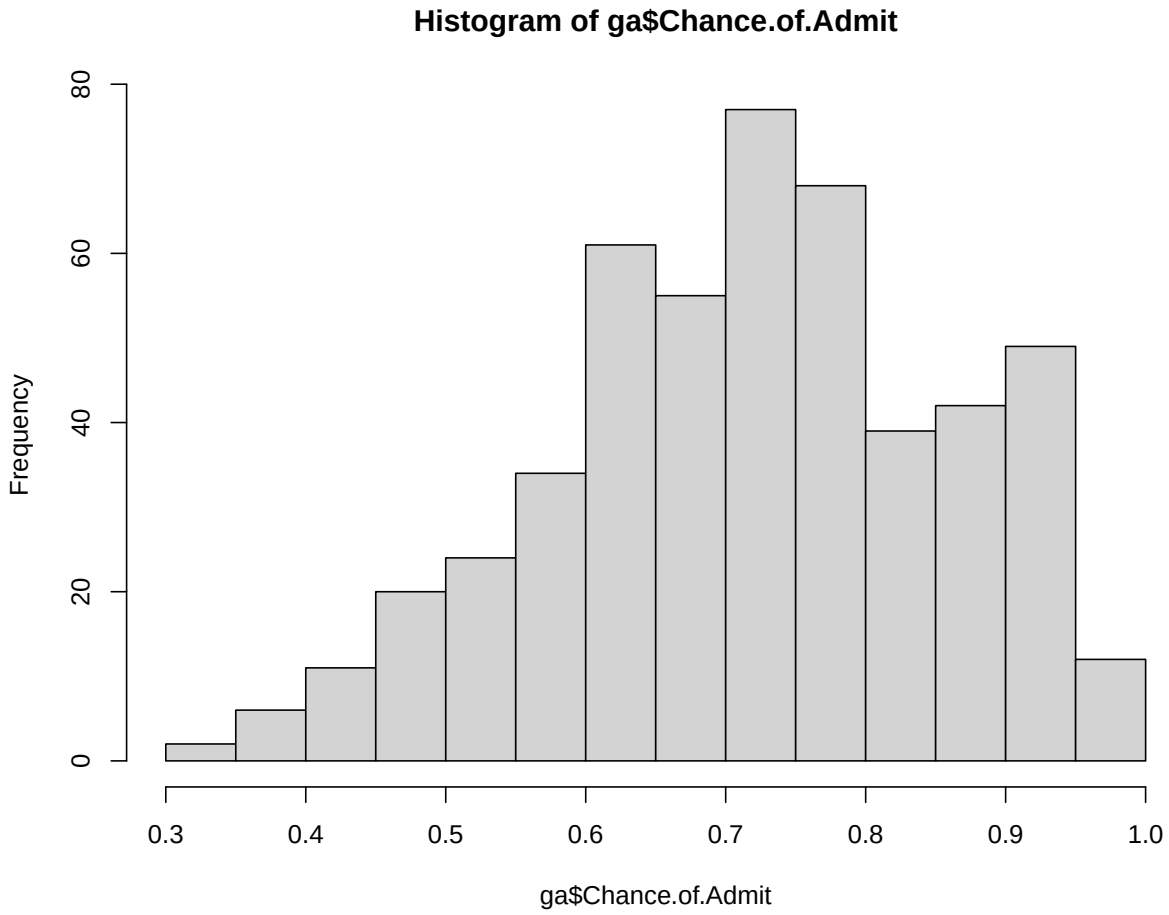

Figure 7: Chance of Admission

Table 2: Variance Inflation Factor (VIF) for each predictor

	GRE.Score	TOEFL.Score	University.Rating	SOP	LOR	CGPA	Research
VIF	1.411306	1.320238	1.261352	1.541973	1.273523	1.381430	1.086474

4 Results

First of all, the result is binary, admitted or not admitted. From the Table 2, since Variance Inflation Factor (VIF) values are less than 5, we can say that there does not exist multi-collinearity. This dataset satisfies that there is large sample size(270), and each variable are independent. Because we are conducting the logistic regression, there is a linear relationship between independent variables and log-odds from equation (3).

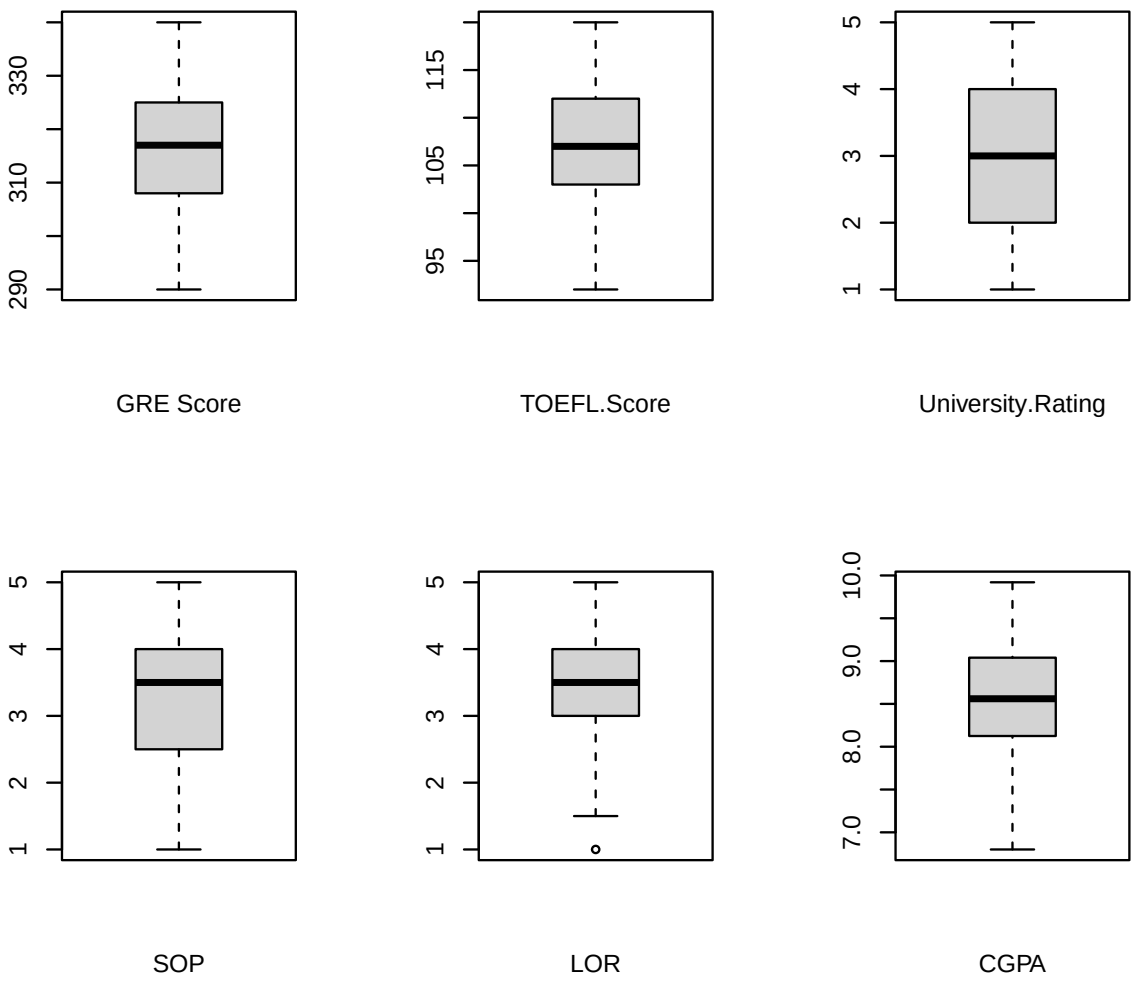

Figure 8: Boxplot of variables

	GRE.Score	TOEFL.Score	University.Rating	SOP	LOR	CGPA	Research	Chance.of.Admit
GRE.Score	1.0000000	0.8272004	0.6353762	0.6134977	0.5246794	0.8258780	0.5633981	0.8103506
TOEFL.Score	0.8272004	1.0000000	0.6497992	0.6444104	0.5415633	0.8105735	0.4670121	0.7922276
University.Rating	0.6353762	0.6497992	1.0000000	0.7280236	0.6086507	0.7052543	0.4270475	0.6901324
SOP	0.6134977	0.6444104	0.7280236	1.0000000	0.6637069	0.7121543	0.4081158	0.6841365
LOR	0.5246794	0.5415633	0.6086507	0.6637069	1.0000000	0.6374692	0.3725256	0.6453645
CGPA	0.8258780	0.8105735	0.7052543	0.7121543	0.6374692	1.0000000	0.5013110	0.8824126
Research	0.5633981	0.4670121	0.4270475	0.4081158	0.3725256	0.5013110	1.0000000	0.5458710
Chance.of.Admit	0.8103506	0.7922276	0.6901324	0.6841365	0.6453645	0.8824126	0.5458710	1.0000000

Figure 9: Correlation between each factor

Figure 10 shows the summary of the data using Logistic Regression Model. From the Figure, we can say that GRE.Score, University.Rating, Letter of Recommendation, College

```
Call:
glm(formula = admid ~ GRE.Score + TOEFL.Score + University.Rating +
    SOP + LOR + CGPA + Research, family = binomial(), data = ga)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6726  -0.3947   0.0531   0.3342   2.4209

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -63.51001    7.58287  -8.375  < 2e-16 ***
GRE.Score           0.10636    0.02679   3.970 7.18e-05 ***
TOEFL.Score         0.07267    0.04701   1.546   0.1222
University.Rating   0.48395    0.20629   2.346   0.0190 *
SOP                -0.13855    0.23303  -0.595   0.5521
LOR                 0.52759    0.22313   2.364   0.0181 *
CGPA                2.34709    0.54100   4.338 1.44e-05 ***
Research            0.59321    0.31444   1.887   0.0592 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 673.01  on 499  degrees of freedom
Residual deviance: 288.49  on 492  degrees of freedom
AIC: 304.49

Number of Fisher Scoring iterations: 6
```

Figure 10: Summary of Logistic Regression Model

GPA play a relatively important role in the chance of admission.

Based on the hypothesis test result from Figure 11, since the p-value are less than 0.05, we can say that the tiny p-value suggests that at least one of the predictors is associated with the outcome. From 12, we can say that the predicted values and residuals follow some patterns even though there are some outliers, and there are not any influential points in our regression model.

```
Analysis of Deviance Table

Model 1: ynadmid ~ 1
Model 2: ynadmid ~ GRE.Score + TOEFL.Score + University.Rating + SOP +
    LOR + CGPA + Research
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       499     673.01
2       492     288.49  7   384.52 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11: Analysis of Deviance table

So,we can drop the predictors which have the tiny p-values. We can try to fit a reduced model without some predictors, TOEFL.Score, SOP and Research because their $p-$values are greater than 0.05 and coefficient are not significant enough, and they may not have any contributions to the model. Follow the Figure 13, The The Akaike Information Criterion (AIC) for the reduced model(304.36) is a bit smaller than the full model(304.49). Thus, we have the Logistic Model based on equation (2),

$$\ln \frac{\pi_i}{1-\pi_i} = -65.244 + 0.13407 \cdot \text{GRE.score} + 0.48808 \cdot \text{University.Rating} + 0.49561 \cdot \text{LOR} + 2.42041 \cdot \text{CGPA} \tag{3}$$

From the figure 14, we can get the similar checking result as full model. Thus, we can say that the two models are fit for this dataset. And we can get the prediction model from equation (3).

5 Discussion

The main point of this paper is to predict the chance of admission to a graduate program using Logistic Regression and to find the relations among the variables. However, there are some limitations. First of all, this data is not accurate enough so we cannot use Logistic Regression properly. Whether a student is accepted or not is an inference based on the

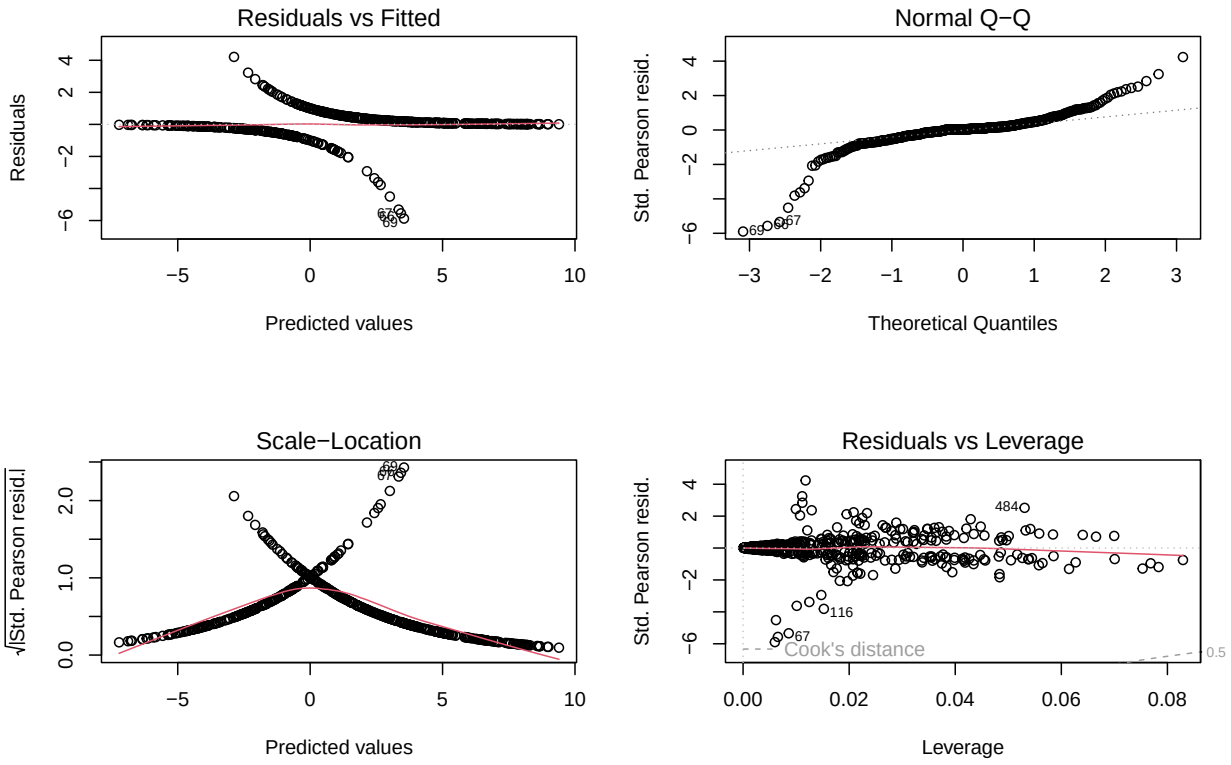

Figure 12: Summary of Full Logistic Regression Model

median and mean of our available data, not the data we actually have. Acharya et al. (2019) using other linear regression and compare with them and get the conclusion said multiple linear regression will be better to predict the chance of admission.

Futhermore, even if there does not exists multicollinearity problem based on the VIF we have shown in result part, the correlation are extremely high from figure 9. In order to solve this, we dropped some variables such as TOEFL.Score. However, we have seen that the GRE score are highly correlated with the College GPA, which we didn't drop because it may play some important role in the graduate admission as well.

Moreover, the chance of admission could depends on more variables, not restricted to the 7 variables we have shown. For example, sometimes the admission committee will consider gender, race, extracurricular activities, contribution to the community, competition, etc. Evaluation of students does not only depend on their academic achievement. Also,

```
Call:
glm(formula = admid ~ GRE.Score + University.Rating + LOR + CGPA,
    family = binomial(), data = ga)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.69580  -0.40229   0.06288   0.36678   2.34588

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -65.24400    7.09543  -9.195  < 2e-16 ***
GRE.Score           0.13407    0.02405   5.574 2.49e-08 ***
University.Rating   0.48808    0.18824   2.593  0.00952 **
LOR                 0.49561    0.20939   2.367  0.01794 *
CGPA                2.42041    0.51529   4.697 2.64e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 673.01  on 499  degrees of freedom
Residual deviance: 294.36  on 495  degrees of freedom
AIC: 304.36

Number of Fisher Scoring iterations: 6
```

Figure 13: Summary of Reduced Logistic Regression Model

the weights are different when applying to different programs. For example, when someone applies to a PhD program rather than a Masters program, the research part would likely plan a more important role. In future studies, researchers could focus more on the effects of the multifaceted variables mentioned above, rather than focusing solely on for academic achievement. This would greatly improve the prediction accuracy of applicants being admitted.

Even if this data has some limitations, it can still give us some suggestions for graduate program application, especially for international students.

Standardized testing is as important as college grades. They are all closely related and

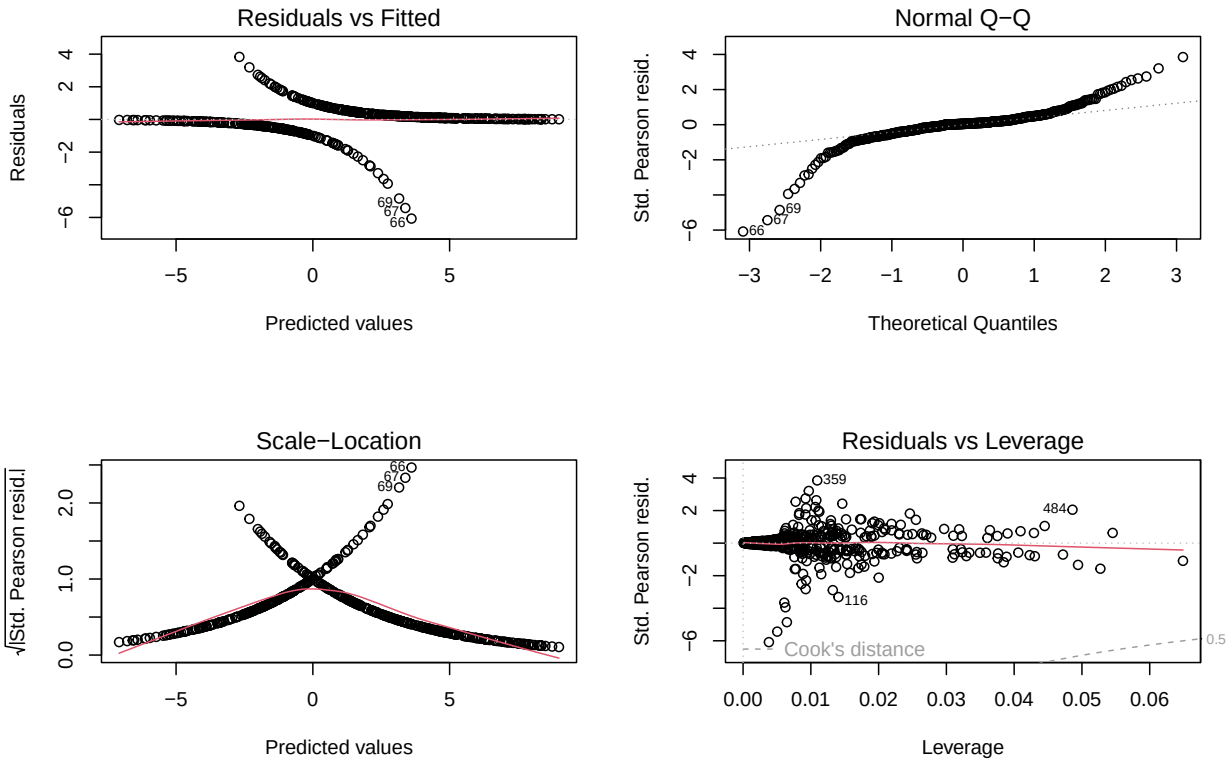

Figure 14: Summary of Reduced Logistic Regression Model

students should prepare early so they can have enough time to get a higher score as earlier as they can.

References

M. S. Acharya, A. Armaan and A. S. Antony (2019). "A comparison of regression models for prediction of graduate admissions." In *2019 international conference on computational intelligence in data science (ICCIDS)*, pp. 1–5. IEEE.

G. Chellaraj, K. E. Maskus and A. Mattoo (2008). "The contribution of international graduate students to us innovation." *Review of International Economics* **16**, 444–462.

N. Gupta, A. Sawhney and D. Roth (2016). "Will i get in? modeling the graduate admission process for american universities." In *2016 IEEE 16th international conference on data mining workshops (ICDMW)*, pp. 631–638. IEEE.

A. J. Jaeger (2003). "Job competencies and the curriculum: An inquiry into emotional intelligence in graduate professional education." *Research in higher education* **44**, 615–639.

G. Mason, G. Williams and S. Cranmer (2009). "Employability skills initiatives in higher education: what effects do they have on graduate labour market outcomes?" *Education Economics* **17**, 1–30.

A. S. A. Mohan S Acharya, Asfia Armaan (2019). "Graduate admission." https://www.kaggle.com/datasets/mohansacharya/graduate-admissions.

M. G. Powers (1990). "Marketing and recruitment for graduate programs." *Council of Graduate Schools Communicator* **24**.

C. Wendler, B. Bridgeman, R. Markle, F. Cline, N. Bell, P. McAllister and J. Kent (2012). "Pathways through graduate school and into careers." *Educational testing service* .

Wikipedia contributors (2023). "Variance inflation factor — Wikipedia, the free encyclopedia." [Online; accessed 23-April-2023].