Predicting Graduate Admission Chances

Aryan Agarwal, Ayush Srivastava, Pratyush Pandey, Ritvik Ajaria, and Uddharsh Kotahwala

ABSTRACT

College admissions are a rigorous process and various parts of a students' profile/portfolio are analyzed and weighed. Every university has its set of parameters and requirements which it prioritizes. Students are often anxious as to which fields they need to improvise on their portfolio. We, in our term project, intend to predict the chances of a students' admissions in a college based on these parameters and college ratings. In this part of the paper, we have analyzed the data set available to obtain the statistics, graphically represent the data, find out the dependence of different parameters and estimate the probability distribution function.

Keywords: Descriptive statistics, Estimation, Parametric dependence

INTRODUCTION

Exploratory Data Analysis (EDA) is the initial investigation of data to find out patterns, or any problems in the data to further test hypotheses and assumptions. This is done by calculating the summary statistics, making relevant graphical representations of the data, finding parametric dependencies of the dependent variable, making estimates of the probability distribution function, etc. In this part of the project, we perform EDA on the data set to find out the summary statistics and graphical representations so that in the later stages, we are equipped well to test the hypotheses we propose, perform further analysis and able to make prediction model for predicting the chances of graduate admissions.

DESCRIBING THE DATA SET

The data set comprises of the following information gathered from the portfolios of 500 different students and the university rating in which s/he has applied.

1. GRE Scores: Measured out of 340 marks.
2. TOEFL Scores: Measured out of 120 marks.
3. University Ratings: Measured on a 5-point scale (1 - Lowest, 5 - Highest).
4. SOP:A subjective measure of a student's statement of purpose scored on a 5-point scale.
5. LOR:A subjective measure of the LOR's possessed by a student scored on a 5-point scale.
6. CGPA: Marked on a 10-point scale.
7. Research: The number of research projects a student has done.
8. Chances of Admission: Probability of admission.

All these parameters are weighed differently by different colleges based on either their ratings or the stream of students they want to admit in their institution. In the later part of the project, we aim to use these as inputs and predict a students' admission chances in a college.

ANALYSES PERFORMED

We have performed the following analyses on the data sets mentioned above, using python and R:

Descriptive Statistics

We calculated the mean, Standard Deviation, Min. Value, Max. Value, 25^{th}, 50^{th}, 75^{th} Percentiles, Coefficient of Kurtosis and Coefficient of Skewness for each columns of the data.

Graphical Representation

We represented the data of university ratings, TOEFL scores, SOP strength, Research experience, LOR strength, GRE scores, chances of admission, and CGPAs using **bar graphs** and **histograms**.

Parametric Dependencies

We plot the dependencies of various parameters and their relation with each other, like, LOR strengths v/s CGPA, SOP v/s TOEFL scores, GRE scores v/s CGPA and so on. Moreover, we plot 2 different best-fit lines for these, as we observe that the entire data has only 2 categories in research experience, 0 or 1. We, thus bifurcate the plots on the basis of research experience and plot 2 best-fit lines in the same graph for further analysis.

Estimation

We tried estimating the underlying distribution of the individual columns using MLE and other type of estimators on parametric distributions like multi modal gaussians . We also distributed the data into two classes one with chances of admission less than 0.5 and one with chances of admission greater than 0.5, and also estimated underlying distributions for these two classes.

RESULTS AND OBSERVATIONS

The following are the results of our analyses:

Descriptive Statistics

The following table shows the results for mean, standard deviation (SD), min. value (min), max. value (max) and k^{th} percentile (k%).

	GRE	TOEFL	Univ. Rating	SOP	LOR	CGPA	Research Exp.	Adm. Prob.
mean	316.472	107.192	3.114	3.374	3.484	8.57644	0.56	0.72174
SD	11.295	6.082	1.144	0.991	0.925	0.605	0.497	0.141
min	290	92	1	1	1	6.8	0	0.34
25%	308	103	2	2.5	3	8.1275	0	0.63
50%	317	107	3	3.5	3.5	8.56	1	0.72
75%	325	112	4	4	4	9.04	1	0.82
max	340	120	5	5	5	9.92	1	0.97

Table 1. Summary Statistics

The following table is on the results obtained for coefficient of skewness and kurtosis.

	Skewness	Kurtosis
GRE Scores	-0.03972223277299966	-0.7159497473139949
TOEFL Scores	0.09531393010261811	-0.6587072628939645
Univ. Rating	0.09002387212374935	-0.8139775647199539
SOP	-0.22828490586525177	-0.7106555639324257
LOR	-0.14485407992929378	-0.7502879246086085
CGPA	-0.02653261314181717	-0.5676573553864674
Research Exp.	-0.24174688920761442	-1.9415584415584415
Adm. Prob.	-0.28909558547899383	-0.4621237427062441

Table 2. Skewness and Kurtosis

Graphical Representation

The following are the graphs we obtained for the data.

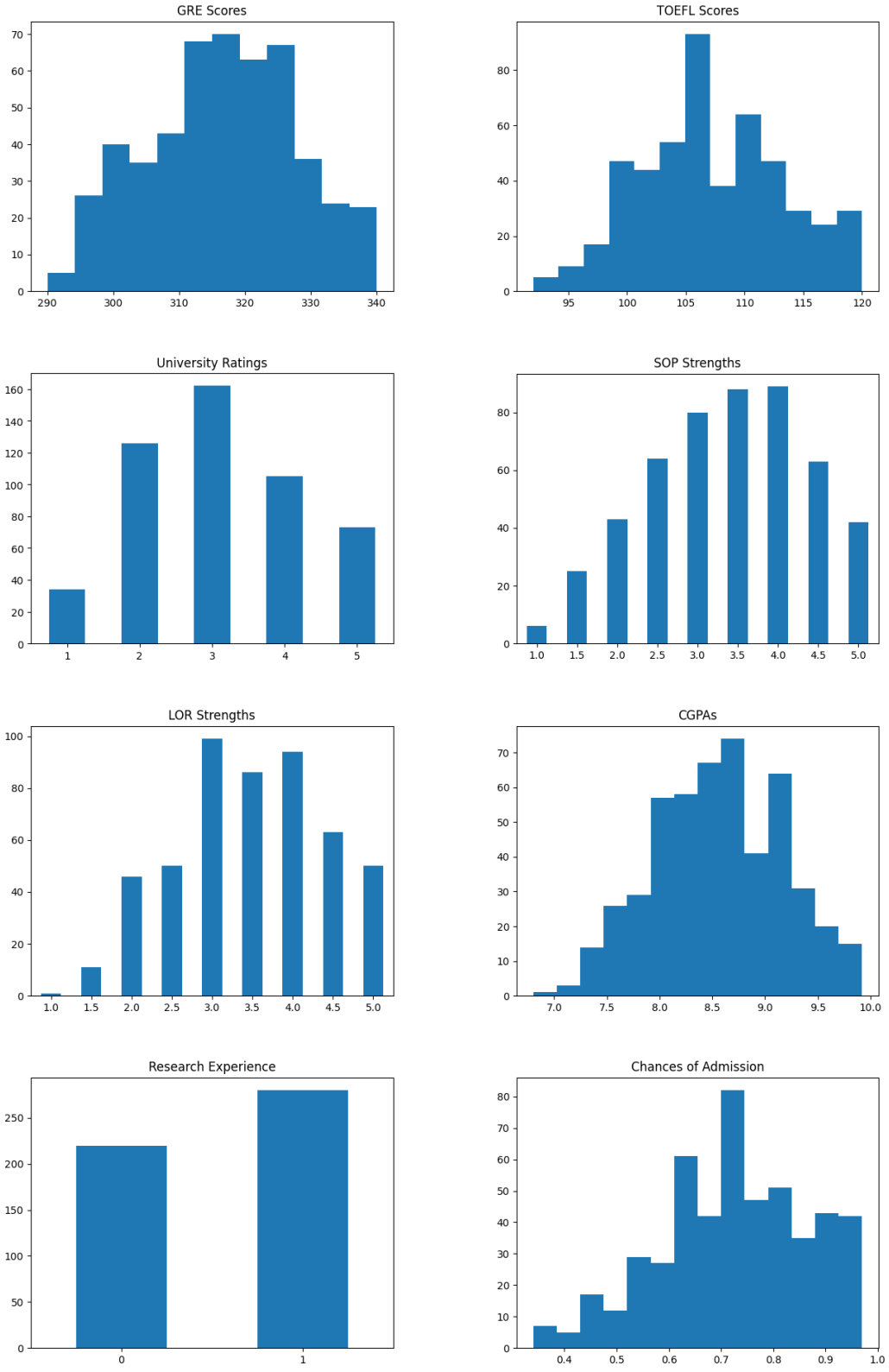

Figure 1. Graphical representation

Parametric dependencies

The following are the plots for dependencies of parameters on one another. In these, we have not bifurcated the data based on research experience.

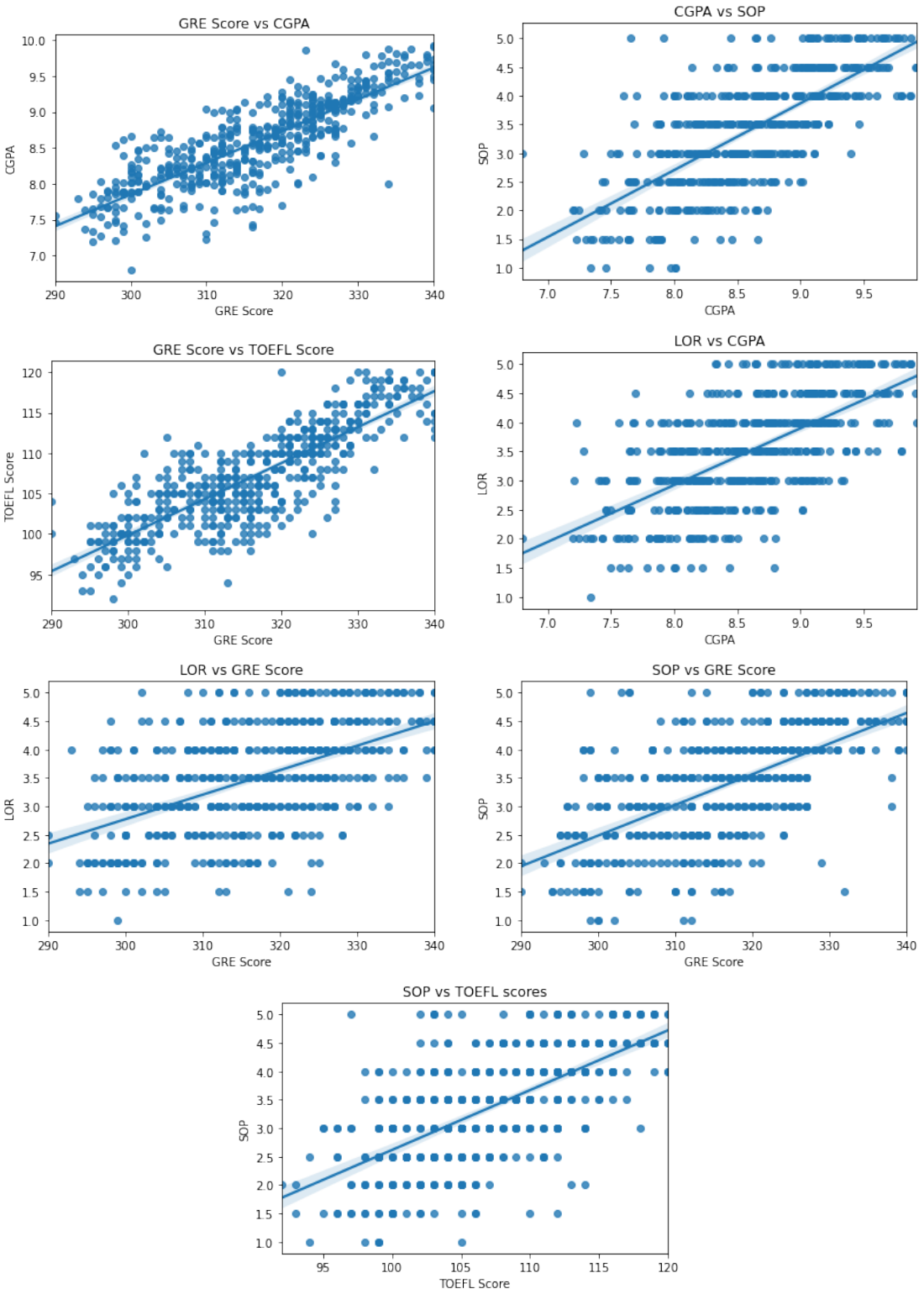

Figure 2. Dependencies of parameters on each other

The bar graph for chances of admission v/s TOEFL and GRE scores are:

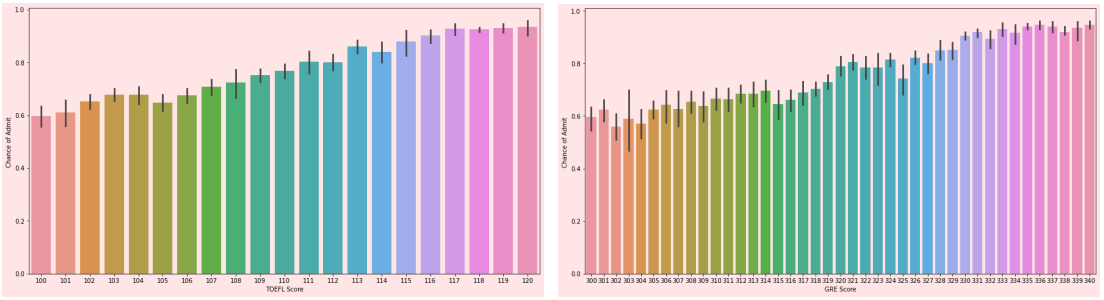

Figure 3. Chances of admission v/s test scores

The distribution of CGPA according to the university rating the student is applying in is:

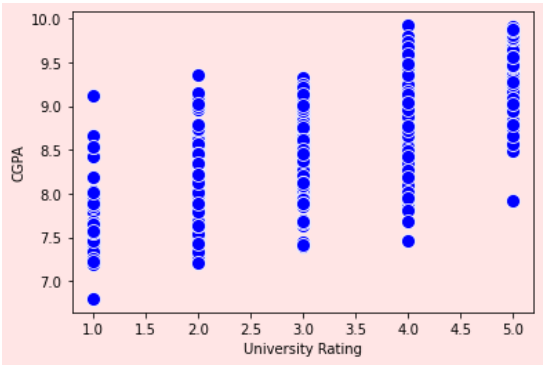

Figure 4. CGPA Distribution

The normalized histograms for distribution of TOEFL and GRE scores are:

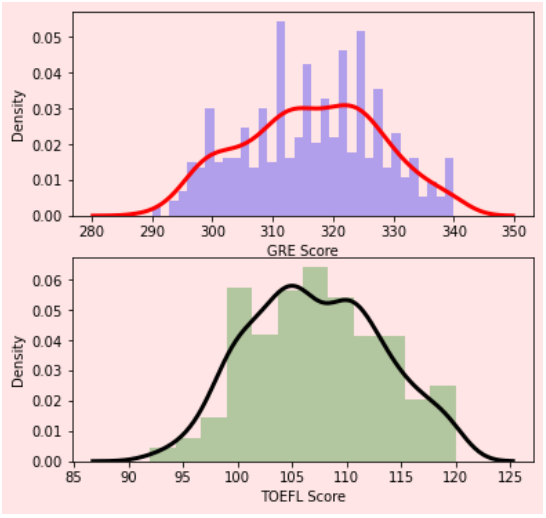

Figure 5. Normalized Histograms

We have summarized the correlation coefficients and following are the results:

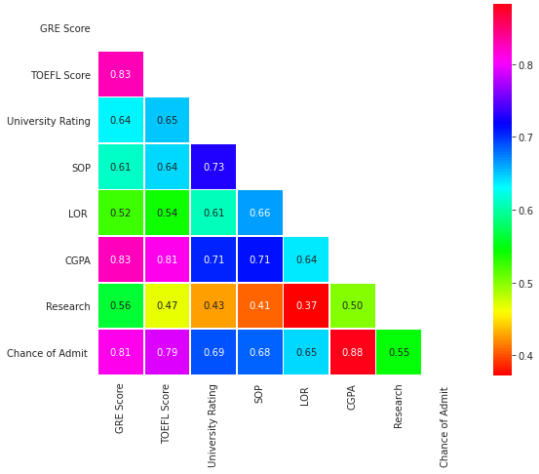

Figure 6. Correlation Coefficients

We can make a variety of interesting observations from these results:

- The TOEFL and GRE scores of the students are highly (and positively) correlated to each other, and both of them are positively correlated to the student's CGPA. From this we can infer that students who are good at academics in their undergraduate tend to perform better at these competitive exams as well.
- Students who study at better universities tend to get a better CGPA, and write a better SOP.
- There is a very little correlation between research experience and LOR strength of students, implying that doing research need not guarantee a better LOR.
- The chance of admit is highly positively correlated towards the CGPA and exam scores, and comparatively less towards the other factors, with the lowest being research experience. This means that grad schools do not value research experience as much as they do academics.

Estimating the underlying distribution

We hope to gain some insights of the data and how they behave so we have tried to estimate these distributions through appropriate type of estimators. This will hopefully help us in Hypothesis testing. We have estimated the underlying distributions using multimodal Gaussians and Discrete Distributions. We have shown various graphs and also mentioned the RSS values.

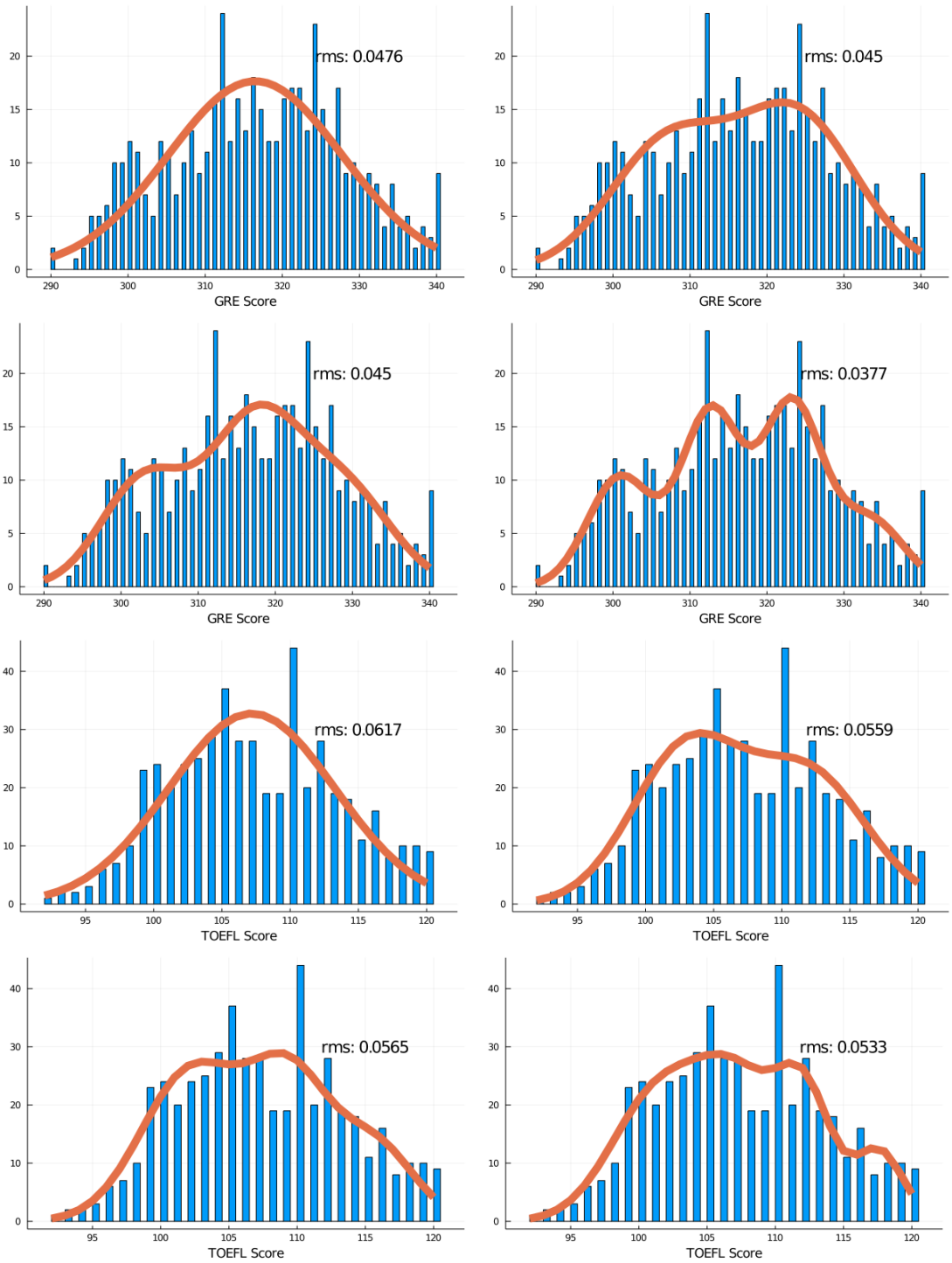

Figure 7. Estimated Probability distribution with RSS values

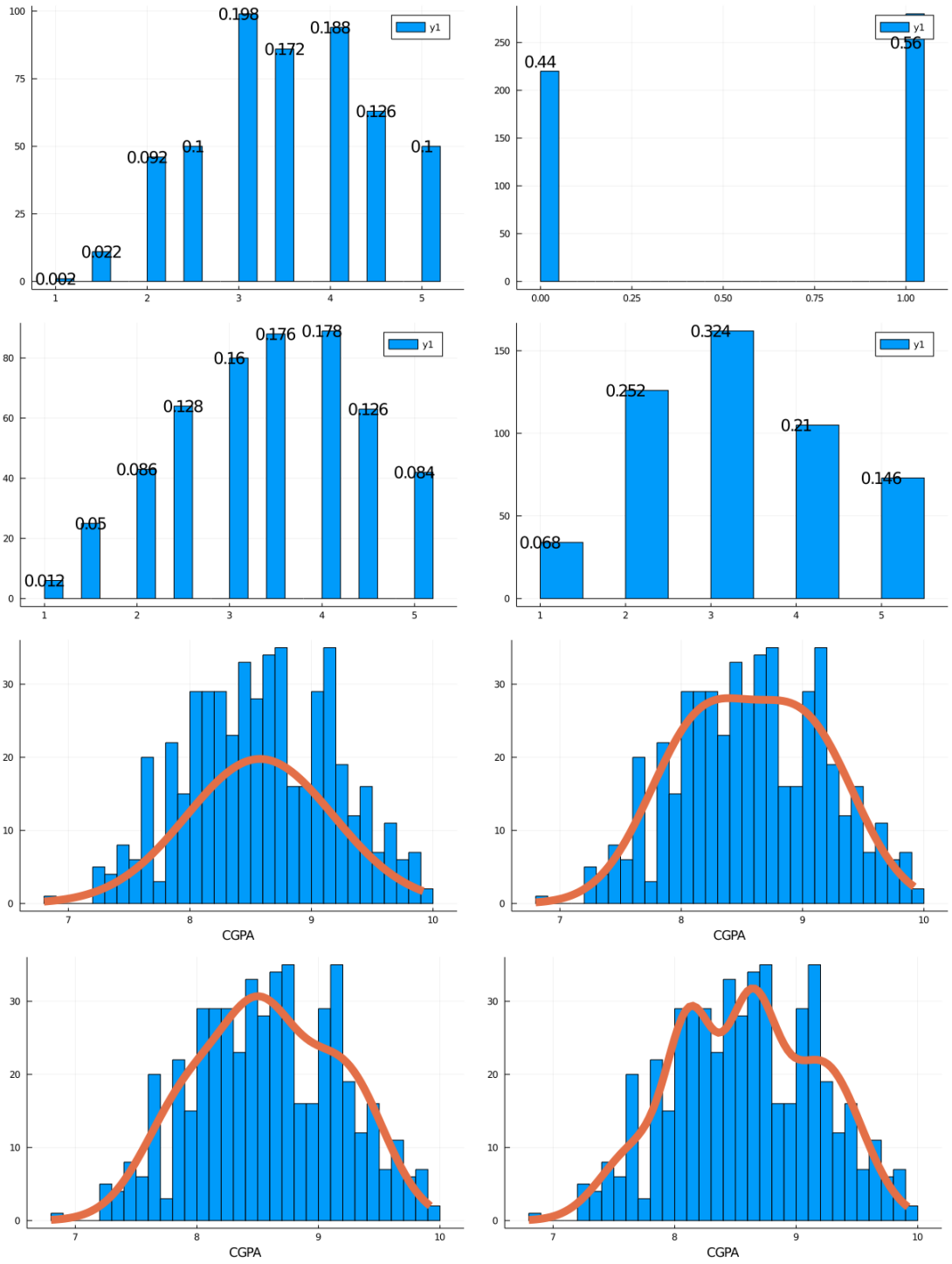

Figure 8. Other Distributions

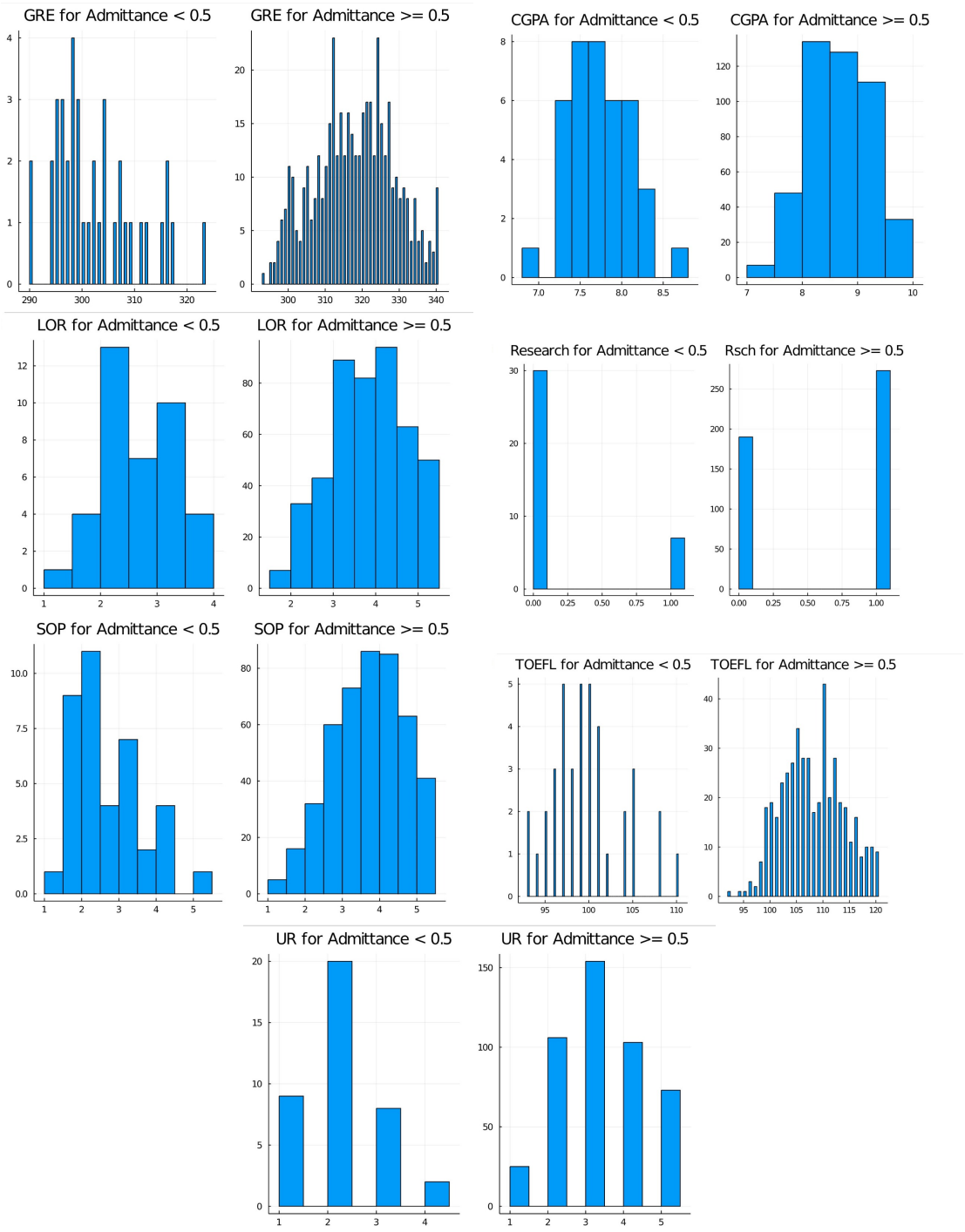

Figure 9. Distribution of classes based upon chances of admission

FUTURE SCOPE AND TARGETS

We have performed the analysis to get an insight into the data and their underlying distributions, and their mutual dependencies. We hope to create several hypothesis with this with the help of results obtained in this paper. Through these hypothesis we hope to answer the original question that we hope to tackle which was to help students identify fields they need to improvise on their portfolio to improve their chances of acceptance. We also aim to ultimately make a linear regression model to predict with some confidence, the chances of admission of a student based on his portfolio.