Data Science Final Project

Graduate Admissions in India and Taiwan

Group2

Members:

0513201 黃子軒 0513230 陳奕婷

0513403 陳昀萱 0513456 林芸如

0513424 李周辰浩

Table of Contents

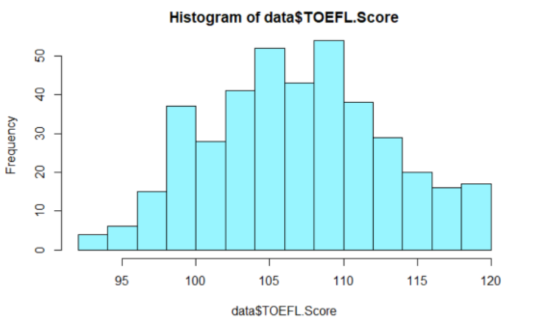
1. Introduction to the dataset: Graduate Admissions in India

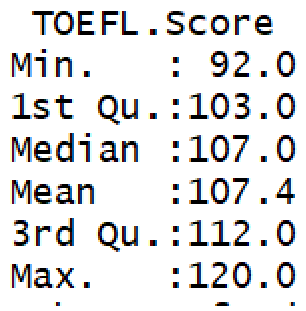
With all five members in our group being senior students, we are deciding what to do after graduation. Most of our classmates, including ourselves, are going to pursue graduate studies. Therefore, we decided to use this dataset and have some fun with it.

This dataset is created for prediction of graduate admissions from an Indian perspective. The dataset contains several parameters which are considered important during the application for Masters Programs. The parameters included are : 1. GRE Scores (out of 340) 2. TOEFL Scores (out of 120) 3. University Rating (out of 5) 4. Statement of Purpose and Letter of Recommendation Strength (out of 5) 5. Undergraduate GPA (out of 10) 6. Research Experience (either 0 or 1) 7. Chance of Admit (ranging from 0 to 1). Our data contains 400 records (in this data it means 400 applicants) and 9 variables. The dataset source comes from this site(https://www.kaggle.com/mohansacharya/graduate-admissions/)

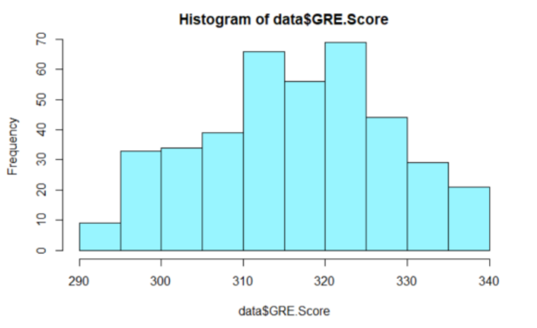
2. Exploratory Data Analysis

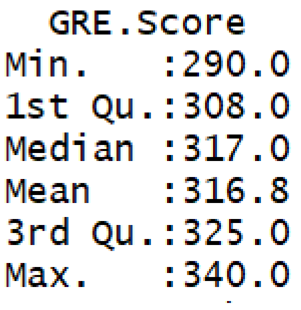
(a) Attributes Distribution and Stats



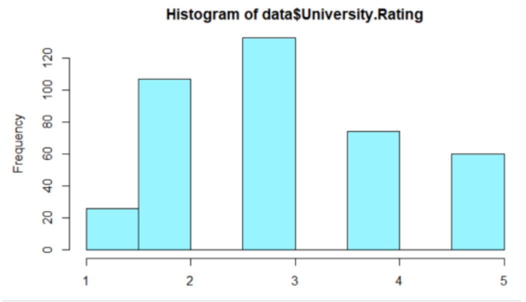


```
 TOEFL.Score
Min.   : 92.0
1st Qu.:103.0
Median :107.0
Mean   :107.4
3rd Qu.:112.0
Max.   :120.0
```

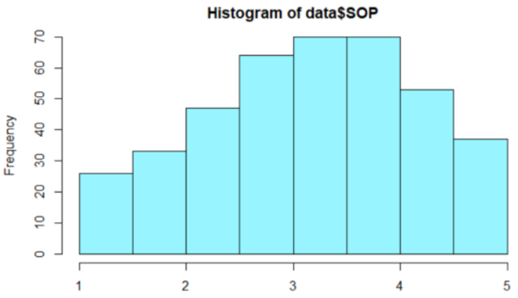




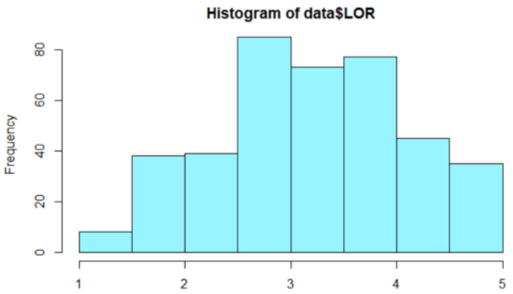
```
  GRE.Score
Min.   :290.0
1st Qu.:308.0
Median :317.0
Mean   :316.8
3rd Qu.:325.0
Max.   :340.0
```



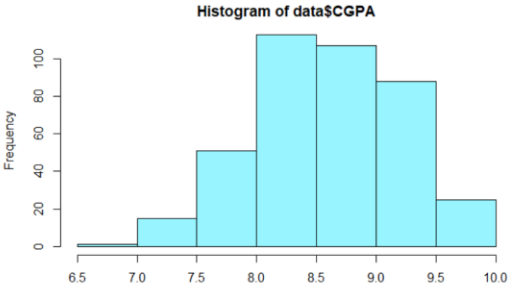
```
University.Rating
Min.   :1.000
1st Qu.:2.000
Median :3.000
Mean   :3.087
3rd Qu.:4.000
Max.   :5.000
```



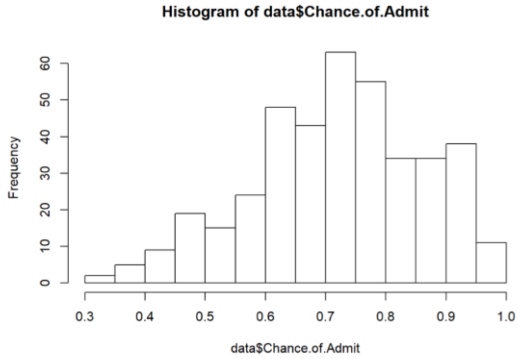
```
      SOP
Min.   :1.0
1st Qu.:2.5
Median :3.5
Mean   :3.4
3rd Qu.:4.0
Max.   :5.0
```



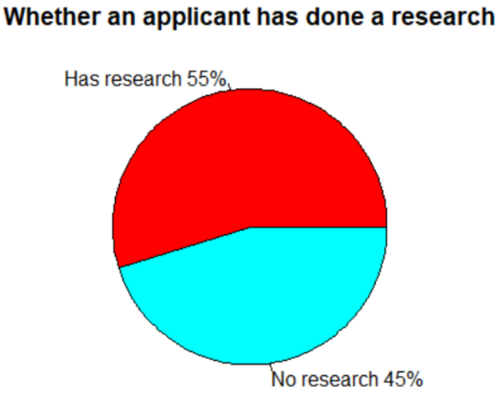
```
      LOR
Min.   :1.000
1st Qu.:3.000
Median :3.500
Mean   :3.453
3rd Qu.:4.000
Max.   :5.000
```



```
      CGPA
Min.   :6.800
1st Qu.:8.170
Median :8.610
Mean   :8.599
3rd Qu.:9.062
Max.   :9.920
```



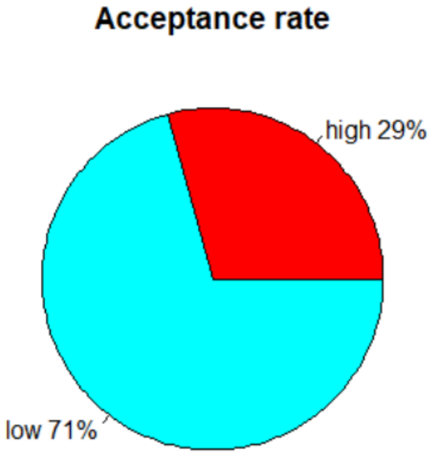
```
Chance.of.Admit
Min.   :0.3400
1st Qu.:0.6400
Median :0.7300
Mean   :0.7244
3rd Qu.:0.8300
Max.   :0.9700
```



If you look at the GRE Score, the lowest score and highest score observed is 290 and 340(full score), and the average score is 316. Then take a look at TOEFL scores, the lowest and highest score observed is 92 and 120(full score), with the average score of 107. The overall chance of admit among these applicants was 72%. The average CGPA is 8.599, which is around 3.7 in the 4.0 scale. Only 55% of the students have done a research in their undergraduate studies.

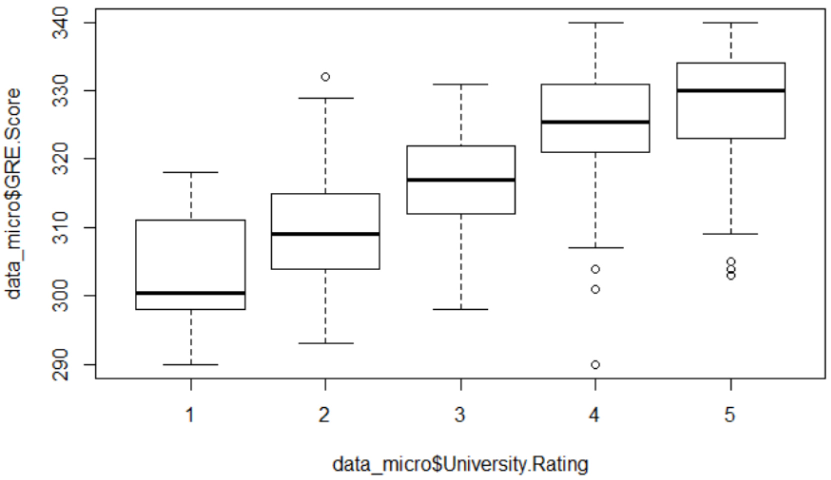
(b) Adding a new column: highacceptance

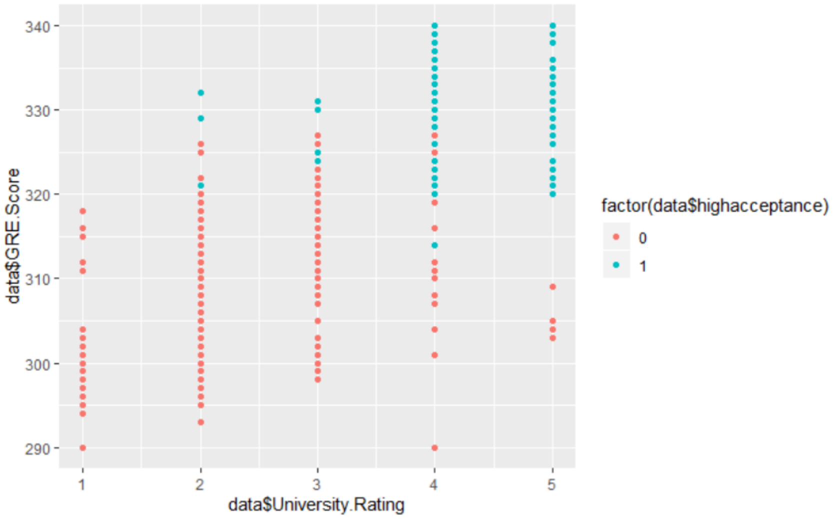
If chance.of.admit is larger than 0.8, then the new column is 1, otherwise, it is set to 0. By doing this conversion, the continuous variable is changed to a binary variable. After this conversion, there are only 29% of the dataset have an acceptance rate higher than 80%.



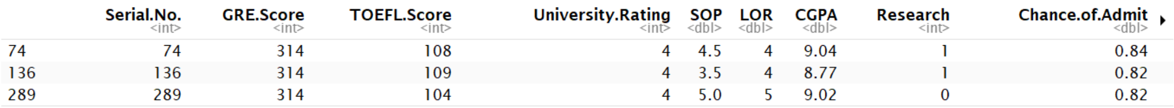
(c) Grouping by University Ratings

c1. University Ratings, GRE scores, and acceptance rate





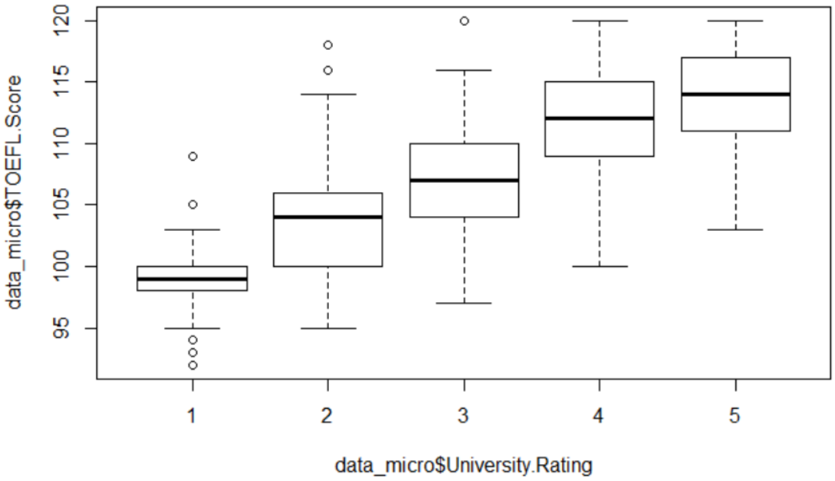
GRE Scores tends to be higher as the applicant’s university’s rating goes higher. But there are still some outliers, showing that there are still some students who work hard in lower-ranking universities and some students who didn’t prepare as well in higher-ranking universities. To see how important the GRE Score is and how high is the threshold for a high admission rate, If you look at the scatterplot, you can see that most scores above 320 has an admission rate higher than 80%, but there is still an applicant that has a score below 320 and still has high admission rate. Let’s see what he’s/she’s got.

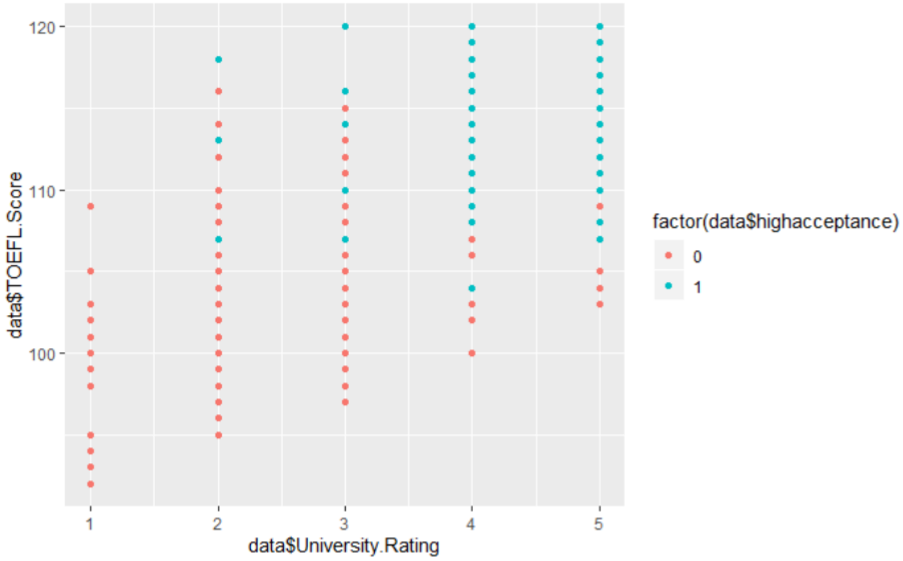


	Serial.No.	GRE.Score	TOEFL.Score	University.Rating	SOP	LOR	CGPA	Research	Chance.of.Admit
	<int>	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	<int>	<dbl>
74	74	314	108	4	4.5	4	9.04	1	0.84
136	136	314	109	4	3.5	4	8.77	1	0.82
289	289	314	104	4	5.0	5	9.02	0	0.82

According to the observation, if you have a GRE score lower than 320, your SOP rating has to be above average(3.4) and your LOR rating has to be above 4 in order to have a chance of admission higher than 80%.

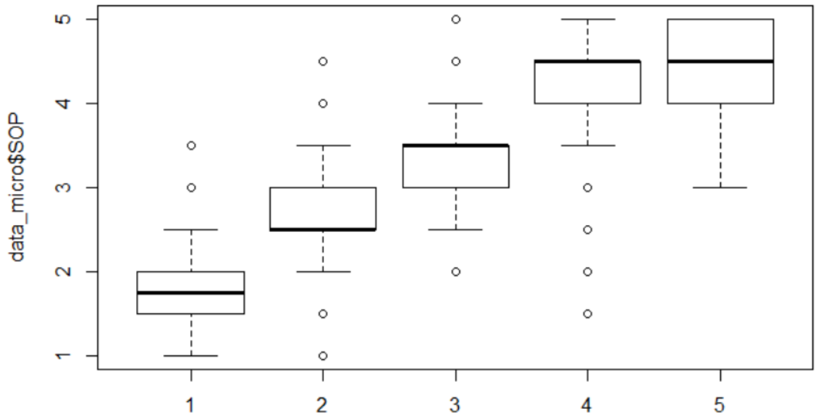
c2. University Ratings, TOEFL scores, and acceptance rate

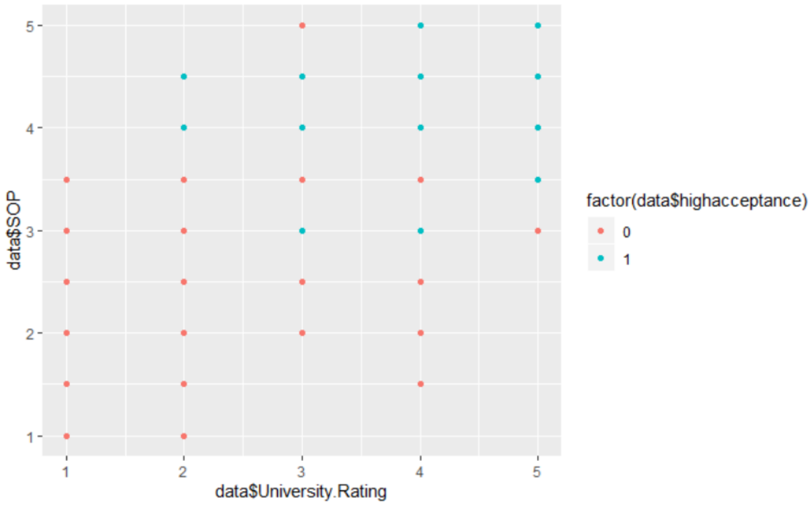




TOEFL also tends to go higher as the university rating gets higher. Although there are still some outliers, where applicants in universities rated 2 and 3 can get scores higher than 115. Compared to GRE scatter plot, there isn’t an obvious threshold where you can say that as long as you receive a TOEFL score higher than 110, you have a higher admission rate. I can only conclude that if you study in a university ranked 4 or 5, and you also have a TOEFL score higher than 110, you will have a higher admission rate.

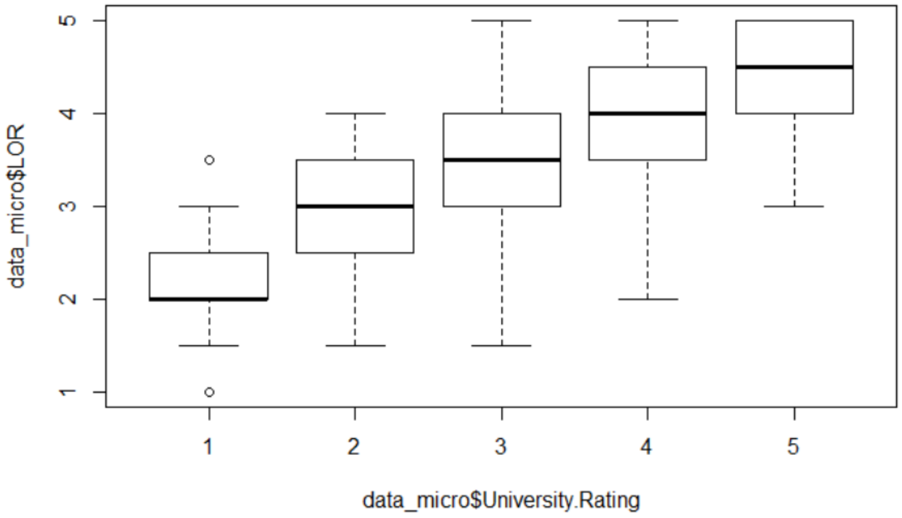
c3. University Ratings, SOP strength, and acceptance rate

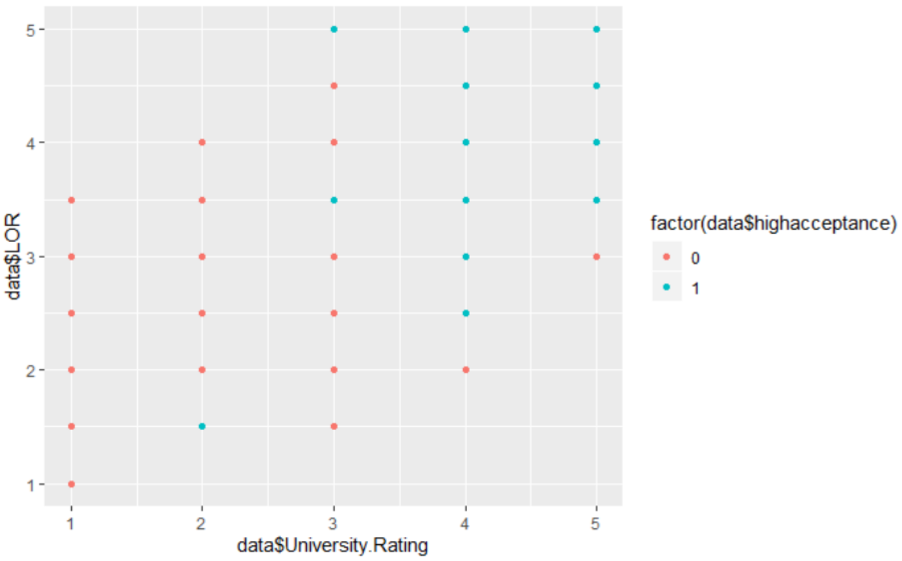




Comparing the average SOP rating of different universities grouped by their ratings, most applicants from universities with higher ratings can write better SOPs. But there are still some students who can write excellent SOPs in lower-ranking universities and some students who write poor SOPs in higher-ranking universities.

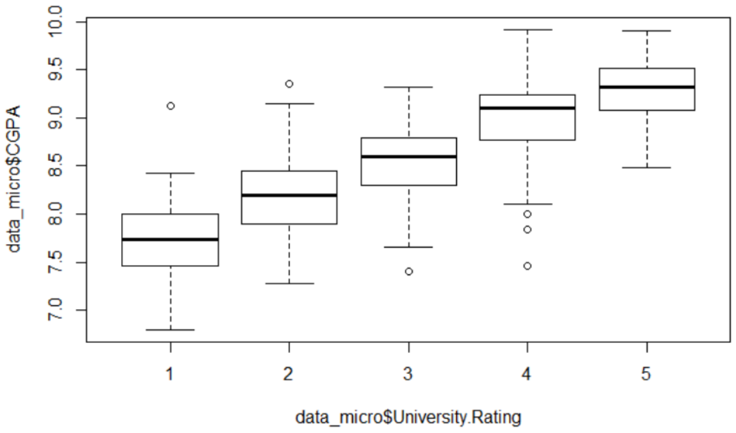
c4. University Ratings, LOR strength, and acceptance rate

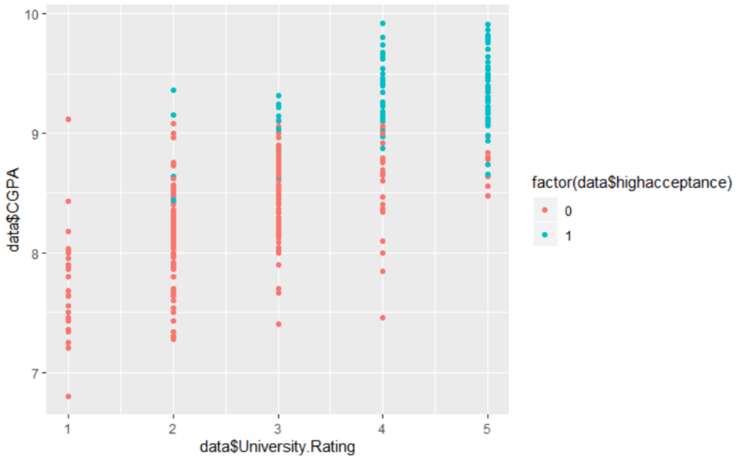




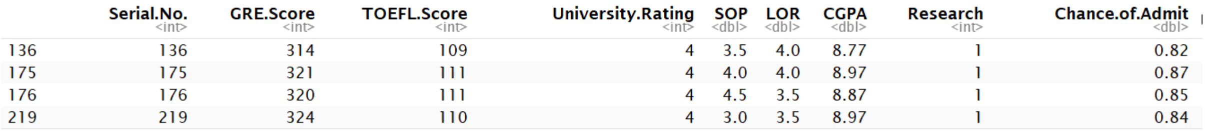
Comparing the average LOR rating of different universities grouped by their ratings, most applicants from universities with higher ratings have better LORs. Something interesting that I noticed is that LOR ratings don't have as much outliers than SOP ratings.

c5. University Ratings, CGPA and acceptance rate

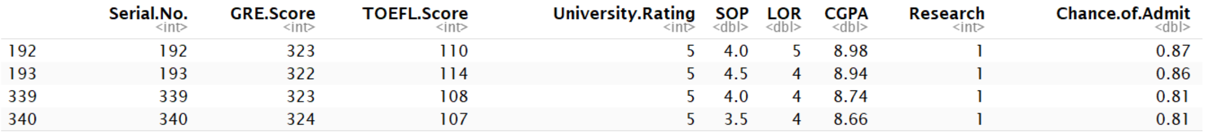




Applicants in universities with higher ranking tends to have higher cgpa. If the applicant's university's ranking is higher than 2 and his/her cgpa is higher than 9, than this applicant has a high possibility to have a high admission rate. What if you don't have a high GPA? We took out the applicants who have CGPAs lower than 9 and still have a high acceptance rate.



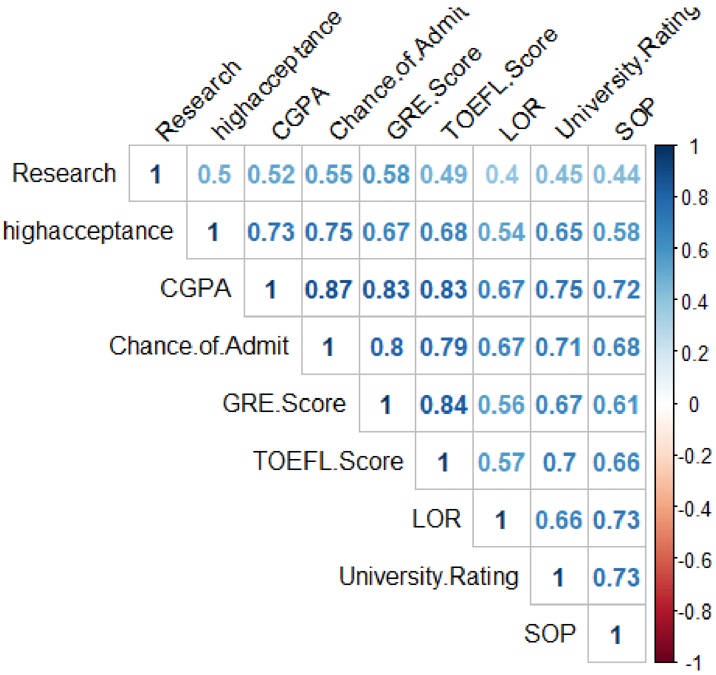
	Serial.No. <int>	GRE.Score <int>	TOEFL.Score <int>	University.Rating <int>	SOP <dbl>	LOR <dbl>	CGPA <dbl>	Research <int>	Chance.of.Admit <dbl>
136	136	314	109	4	3.5	4.0	8.77	1	0.82
175	175	321	111	4	4.0	4.0	8.97	1	0.87
176	176	320	111	4	4.5	3.5	8.87	1	0.85
219	219	324	110	4	3.0	3.5	8.97	1	0.84



	Serial.No. <int>	GRE.Score <int>	TOEFL.Score <int>	University.Rating <int>	SOP <dbl>	LOR <dbl>	CGPA <dbl>	Research <int>	Chance.of.Admit <dbl>
192	192	323	110	5	4.0	5	8.98	1	0.87
193	193	322	114	5	4.5	4	8.94	1	0.86
339	339	323	108	5	4.0	4	8.74	1	0.81
340	340	324	107	5	3.5	4	8.66	1	0.81

In order to have a high acceptance rate, students need to have a high TOEFL and GRE score. And make sure you do research in your undergraduate studies.

(d) Correlation Matrix



	Research	highacceptance	CGPA	Chance.of.Admit	GRE.Score	TOEFL.Score	LOR	University.Rating	SOP
Research	1	0.5	0.52	0.55	0.58	0.49	0.4	0.45	0.44
highacceptance		1	0.73	0.75	0.67	0.68	0.54	0.65	0.58
CGPA			1	0.87	0.83	0.83	0.67	0.75	0.72
Chance.of.Admit				1	0.8	0.79	0.67	0.71	0.68
GRE.Score					1	0.84	0.56	0.67	0.61
TOEFL.Score						1	0.57	0.7	0.66
LOR							1	0.66	0.73
University.Rating								1	0.73
SOP									1

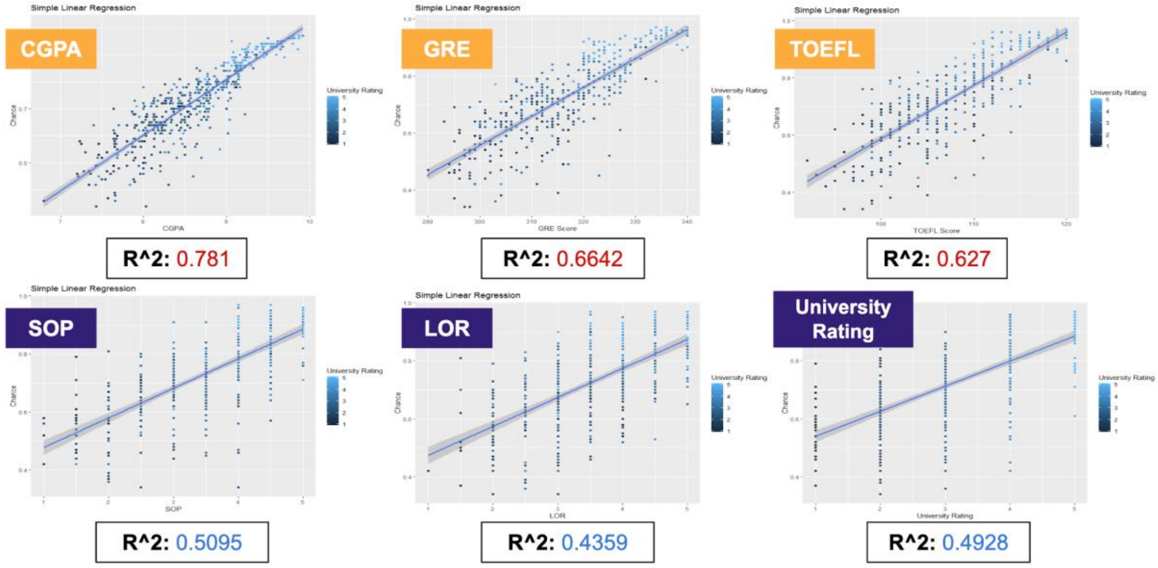
All attributes are positively correlated. Looking at the ones that have a correlation higher than 0.8, they are CGPA, GRE score, TOEFL score, and Chance.of.Admit. We can observe that these three attributes are especially important in affecting the chance of admission.

3. Regression and Classification Results

(a) Regression

After doing the exploratory data analysis, we decided to run the data on some regression models, in order to observe the relationship between the attributes and “Chance of Admit”. We use different combinations of attributes as inputs, letting admission chance (scaling from 0 to 1) regress on those permutations, and here are some results:

- **Simple Linear Regression:**

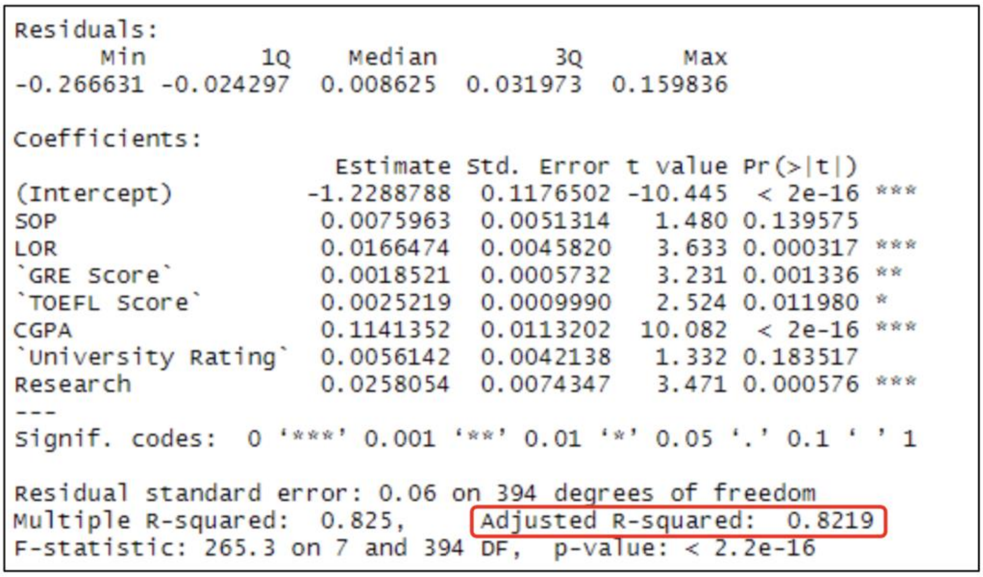


Our first experiment in this section is using merely one feature at a time as our input, and compare their R-square results ultimately. From the graphs above, we can see that the attributes with an orange tag (CGPA, GRE, TOEFL) show relatively better results than those with a purple tag (SOP, LOR, University Rating). We can assume that orange-tagged attributes are more important when it comes to graduate admission. (We skipped binary attributes in this section.)

In order to improve regression R-square result, we decided to increase our input features at a time.

- **Multiple Linear Regression:**

All attributes



```
Residuals:
      Min        1Q    Median        3Q       Max 
-0.266631 -0.024297  0.008625  0.031973  0.159836 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -1.2288788  0.1176502 -10.445  < 2e-16 ***
SOP                  0.0075963  0.0051314   1.480 0.139575    
LOR                  0.0166474  0.0045820   3.633 0.000317 ***
`GRE Score`          0.0018521  0.0005732   3.231 0.001336 ** 
`TOEFL Score`        0.0025219  0.0009990   2.524 0.011980 *  
CGPA                 0.1141352  0.0113202  10.082  < 2e-16 ***
`University Rating`  0.0056142  0.0042138   1.332 0.183517    
Research             0.0258054  0.0074347   3.471 0.000576 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06 on 394 degrees of freedom
Multiple R-squared:  0.825,	Adjusted R-squared:  0.8219 
F-statistic: 265.3 on 7 and 394 DF,  p-value: < 2.2e-16
```

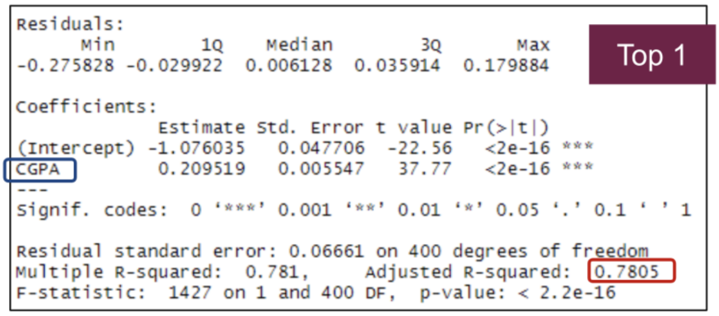
As expected, we poured all the features in the model during our first try. The R-square immediately rose to 0.825, in comparison to the best result from simple linear regression (CGPA: 0.781), it is a big progress.

However, it’s easy to get high accuracy by throwing all the attributes in the model, but in real case people don’t usually do that. So instead of R-square, we decided to optimize “Adjusted R-square”, which gives punishment on the increase of number of using attributes, by selecting only part of the attributes.

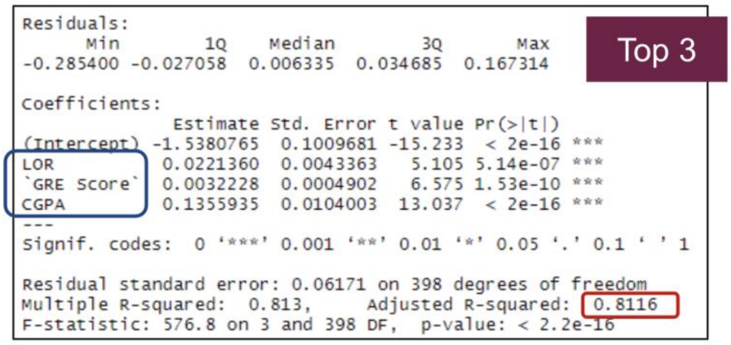
$$Adj.R^2 = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

- **Multiple Linear Regression (variable selection):**

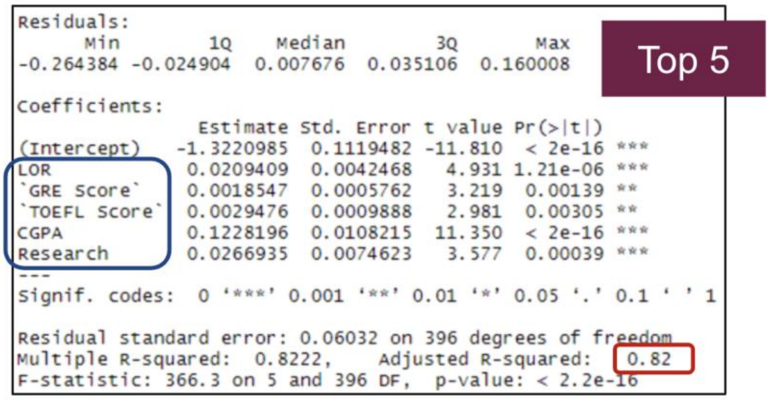
So as to cut down the number of input features, we picked those with higher t-value (t-value in some ways represents the importance of this attribute to the model), and put them into combinations.



```
Residuals:
      Min        1Q    Median        3Q       Max
-0.275828 -0.029922  0.006128  0.035914  0.179884

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.076035   0.047706  -22.56   <2e-16 ***
CGPA         0.209519   0.005547   37.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06661 on 400 degrees of freedom
Multiple R-squared:  0.781,	Adjusted R-squared:  0.7805
F-statistic:  1427 on 1 and 400 DF,  p-value: < 2.2e-16
```



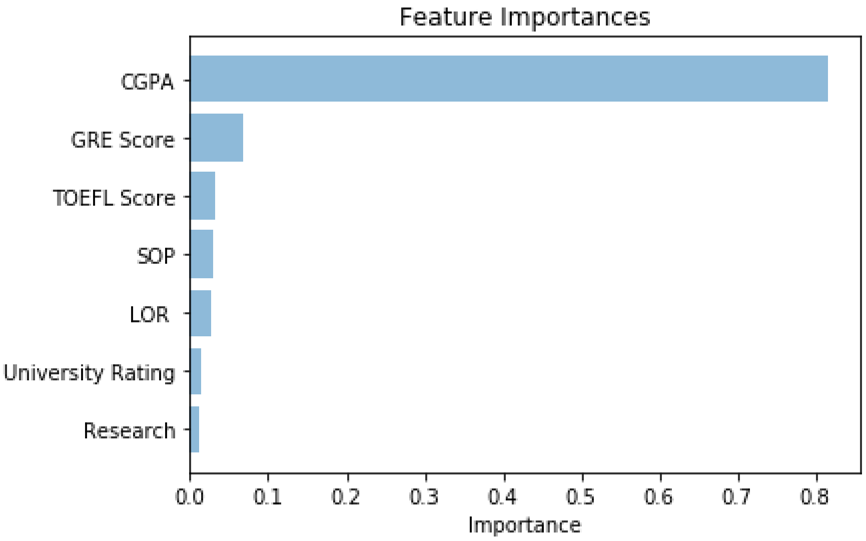
```
Residuals:
      Min        1Q    Median        3Q       Max
-0.285400 -0.027058  0.006335  0.034685  0.167314

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.5380765  0.1009681 -15.233  < 2e-16 ***
LOR          0.0221360  0.0043363   5.105 5.14e-07 ***
`GRE Score`  0.0032228  0.0004902   6.575 1.53e-10 ***
CGPA         0.1355935  0.0104003  13.037  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06171 on 398 degrees of freedom
Multiple R-squared:  0.813,	Adjusted R-squared:  0.8116
F-statistic: 576.8 on 3 and 398 DF,  p-value: < 2.2e-16
```



```
Residuals:
      Min        1Q    Median        3Q       Max
-0.264384 -0.024904  0.007676  0.035106  0.160008

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.3220985  0.1119482 -11.810  < 2e-16 ***
LOR           0.0209409  0.0042468   4.931 1.21e-06 ***
`GRE Score`   0.0018547  0.0005762   3.219  0.00139 **
`TOEFL Score` 0.0029476  0.0009888   2.981  0.00305 **
CGPA          0.1228196  0.0108215  11.350  < 2e-16 ***
Research      0.0266935  0.0074623   3.577  0.00039 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06032 on 396 degrees of freedom
Multiple R-squared:  0.8222,	Adjusted R-squared:   0.82
F-statistic: 366.3 on 5 and 396 DF,  p-value: < 2.2e-16
```

We selected attributes with top 1, 3, 5 highest t-value as three different input combinations. Although the adjusted R-square didn't improve, the one using only 5 features also showed a 0.82 outcome, which is really close to the one using all the features.

To sum up, we can know the level of importance of attributes by the graph below.



(b) Classification Models

Furthermore, we ran this data on some classification models. To implement "classification", we converted our target -- Chance of Admit, into binary data. If the original probability is larger than 0.8, we labeled it as 1, otherwise 0. We also split our data into 80-20 training and testing sets.

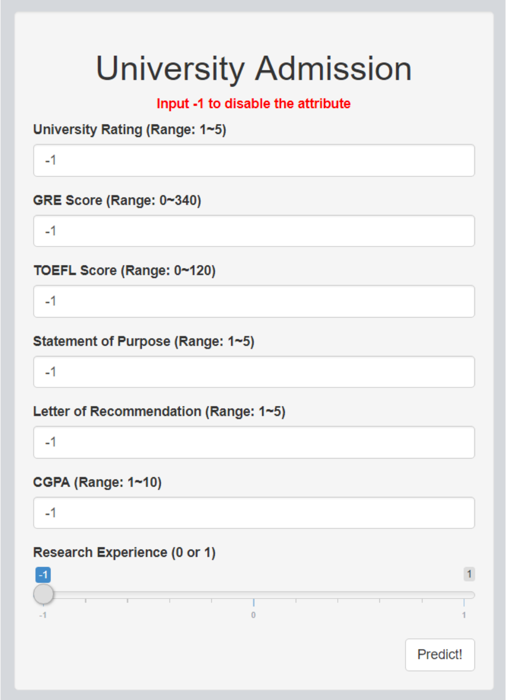
Model	Accuracy
Logistic Regression - All variables	**91%**
Logistic Regression - Top 5	90%
Logistic Regression - Top 3	87%
Logistic Regression - Top 1(CGPA)	87%
Naive Bayes	88%
Decision Tree	83%

Model	Accuracy
KNN: k = 10, 14, 16, 17, 18	90%
k = 11, 13, 15, 19, 20	89%
k = 2, 4, 6, 7, 8, 9, 12	86% - 88%
k = 1, 3, 5	78% - 83%
SVM - linear	88%
SVM - radial	88%

As we can see, all the models performed pretty well, yet "logistic regression model with all-variable input" reached a slightly higher rate of accuracy (91%).

4. Application: Shiny Web

After running the data through different classification models. The result shows that logistic regression model has a fairly high accuracy in predicting the data, so we decided to implement a simple Shiny Web App with logistic regression model.



University Admission

Input -1 to disable the attribute

University Rating (Range: 1~5): -1

GRE Score (Range: 0~340): -1

TOEFL Score (Range: 0~120): -1

Statement of Purpose (Range: 1~5): -1

Letter of Recommendation (Range: 1~5): -1

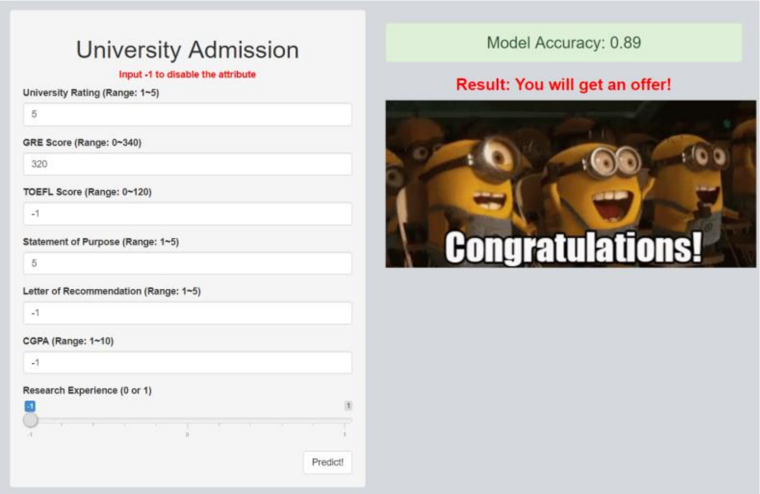
CGPA (Range: 1~10): -1

Research Experience (0 or 1): -1

- **Input**: Any attribute combination the user chose to input

- **Output**: The prediction of `highacceptance` (1 or 0)

The web app allows users to input any attribute combination they want, those panels which are input "-1" are disabled. After pressing the "predict" button, the app reactively trains different logistic regression models depending on the attributes chosen by the user and shows if he/she will get an offer.



University Admission

Input -1 to disable the attribute

University Rating (Range: 1~5): 5

GRE Score (Range: 0~340): 320

TOEFL Score (Range: 0~120): -1

Statement of Purpose (Range: 1~5): 5

Letter of Recommendation (Range: 1~5): -1

CGPA (Range: 1~10): -1

Research Experience (0 or 1): -1

Model Accuracy: 0.89

Result: You will get an offer!

As shown on the right, if you input your University Rating, GRE Score and Statement of Purpose, the app result shows whether you will get an offer (in this case, yes!) and also shows the model accuracy, which is 89% if you train a logistic regression model depending on these 3 attributes. Of course, if you input more attributes to train the model, the model accuracy will be higher, you get more accurate predictions.

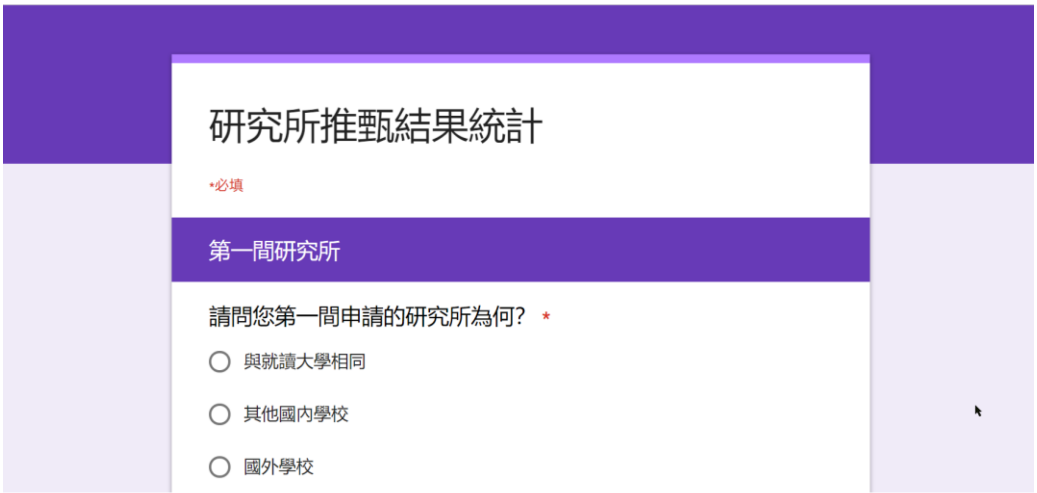
5. Extended Research: Graduate Admission in Taiwan

Recently, senior students are applying for graduate schools, which arouses our curiosity about the graduate admission results in Taiwan, so we decided to conduct a survey by issuing online-questionnaire on Facebook.

(a) Online Questionnaire Design

There are two parts to our questionnaire. First part is about graduate school, and the second part is about personal information. What follows is a description of questions in first part.

- Which graduate school did you apply?
- Which department did you apply?
- Did you need to attend an interview?
- How was the result?



研究所推甄結果統計

*必填

第一間研究所

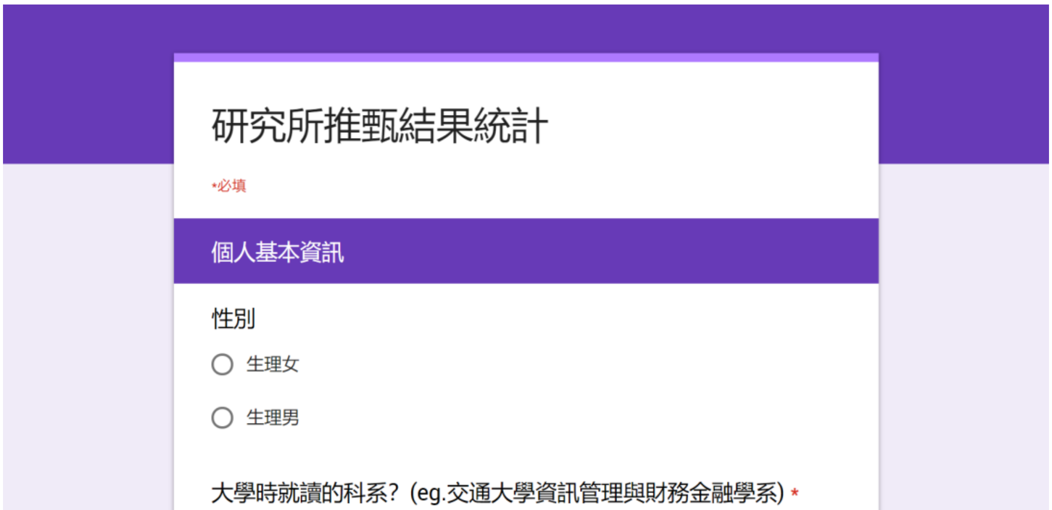
請問您第一間申請的研究所為何？ *

○ 與就讀大學相同

○ 其他國內學校

○ 國外學校

In addition, it is important to understand more about the respondents. Therefore, we ask for personal information about gender, English level, major, performance in university (average score and average ranking) and whether have some specific experiences such as thesis, internship, competition and group/team.



研究所推甄結果統計

*必填

個人基本資訊

性別

○ 生理女

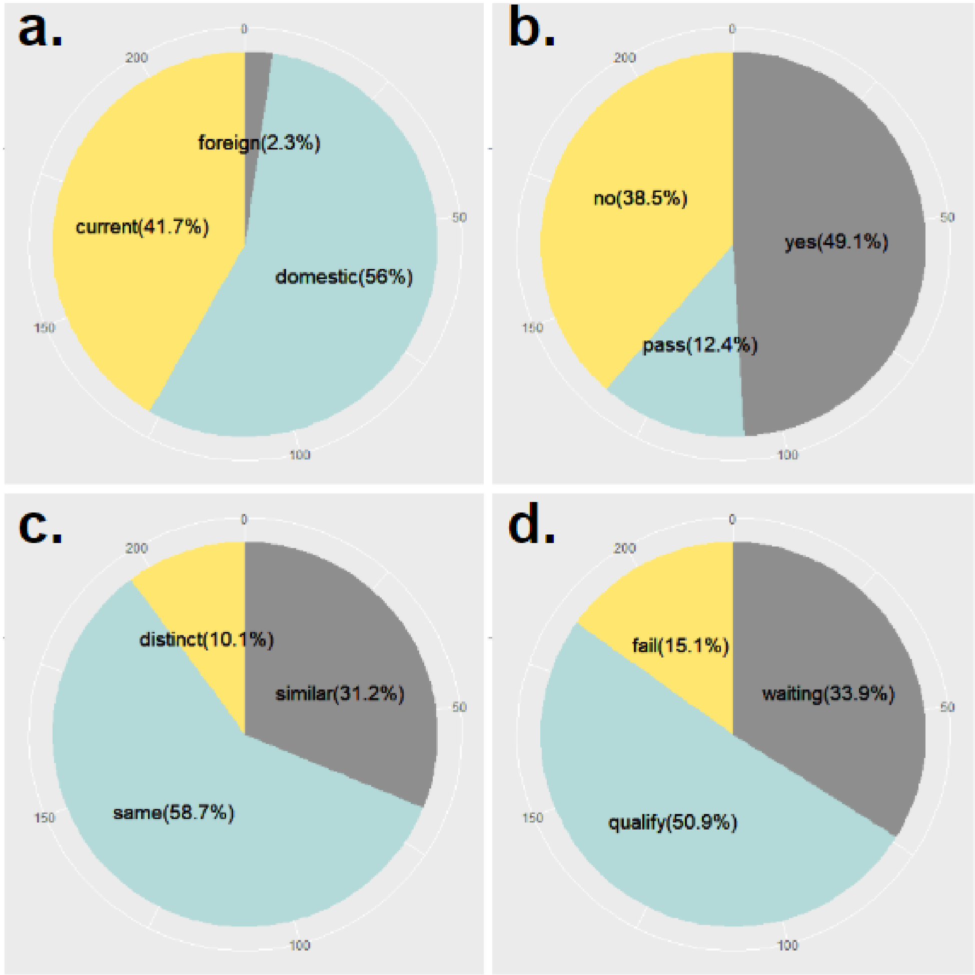
○ 生理男

大學時就讀的科系？(eg.交通大學資訊管理與財務金融學系) *

(b) Results

Finally, we receive 222 replies, and 218 of them are effective samples. Then, we convert the responses into 16 attributes, including a. GraduateSchool, b. Interview, c. Department, d. Outcome, e. Gender, f. Major, g. AverageScore, h. Ranking, i. EnglishLevel, j. Thesis, k. Internship, l. Competition, m. SchoolTeam, n. DepTeam, o. Association, p. Club.

In pie chart a. GraduateSchool, 42% of respondents decide to apply for graduate school in their current university; 56% of them choose to go to other domestic university, and only 2% of students try to study abroad. In pie chart b. Interview, “yes” means the respondent need to attend an interview; “no” means there is no interview, and “pass” means there is an interview, but the respondent is directly admitted. In pie chart c. Department, 59% of respondents decide to apply for the same department in graduate school; 31% of them choose to go relative department, and only 10% of students want to switch the academic field completely. In pie chart d. Outcome, it was happy to know that more than half of respondents are qualified, and one-third of them are waiting for the final consequence.



Bar chart e. Gender, shows that there are more female than male in our samples.

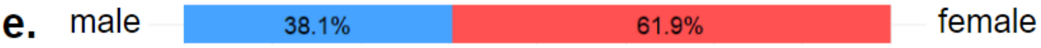
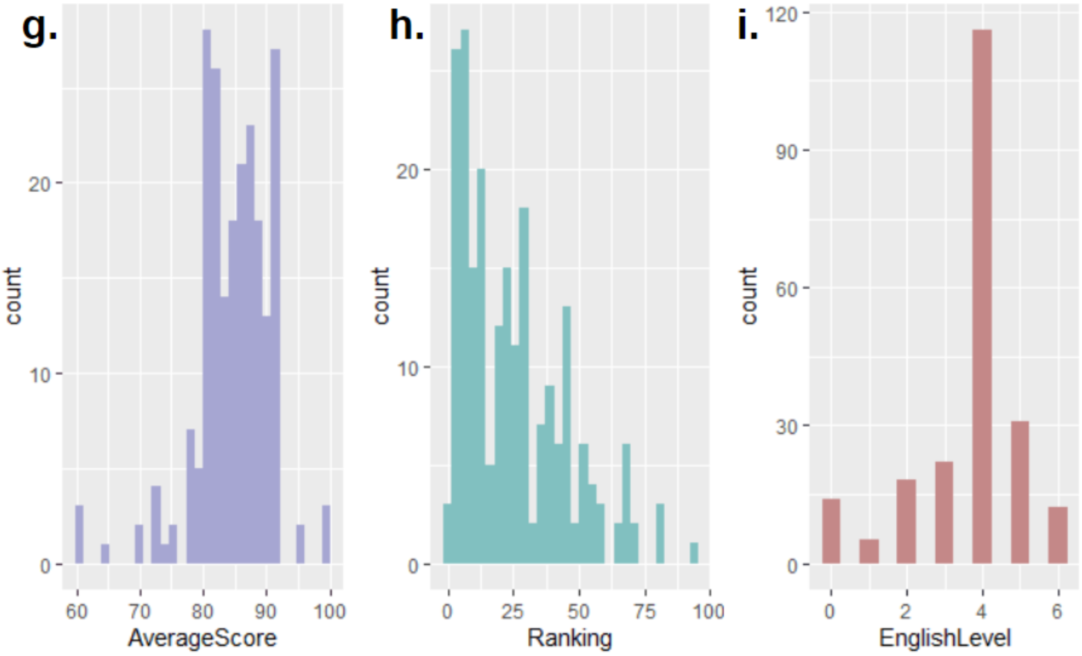
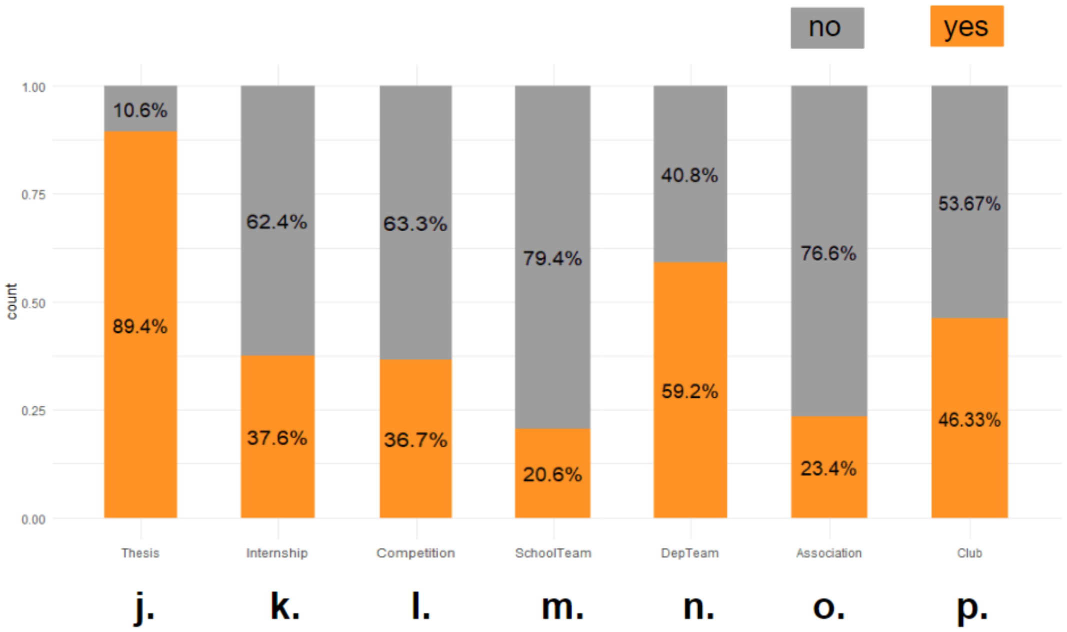


Chart g. AverageScore, h. Ranking, i. EnglishLevel, reflects the specific distribution. In chart g. AverageScore, the great majority of respondents’ average score in university are between 80 to 90. In chart h. Ranking, most of the students are top 30%. Chart i. EnglishLevel indicates that most of the students’ English level are in B2(4).

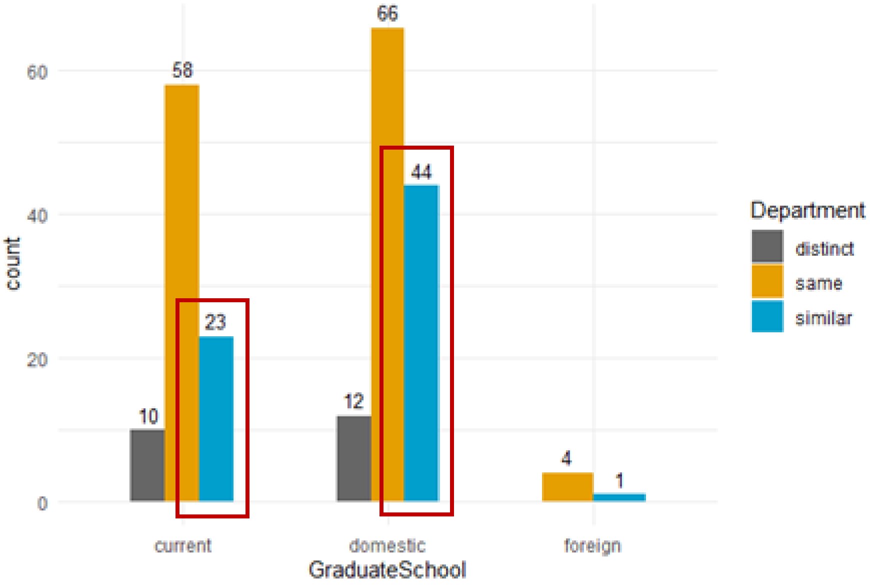


These bar charts below show the percentage of students who have specific experience. 89% of students have done a least one research; 38% of them have worked as an intern; 37% of them have participated in an academic competition; 21% of them are school team member; 59% of them join their department team; 23% of them take part in regional association ,and 46% of them are club participant.



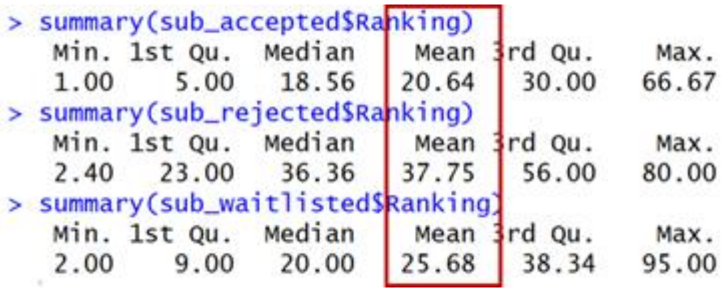
- **Preference of Graduate School**

The bar chart below, as we see, illustrates students' preference of applying for their graduate school. Compared the highlighted columns, we can say that if students plan to study similar major but not the same as their major in the university, they would like to apply for the graduate school which is different from the university they've studied.



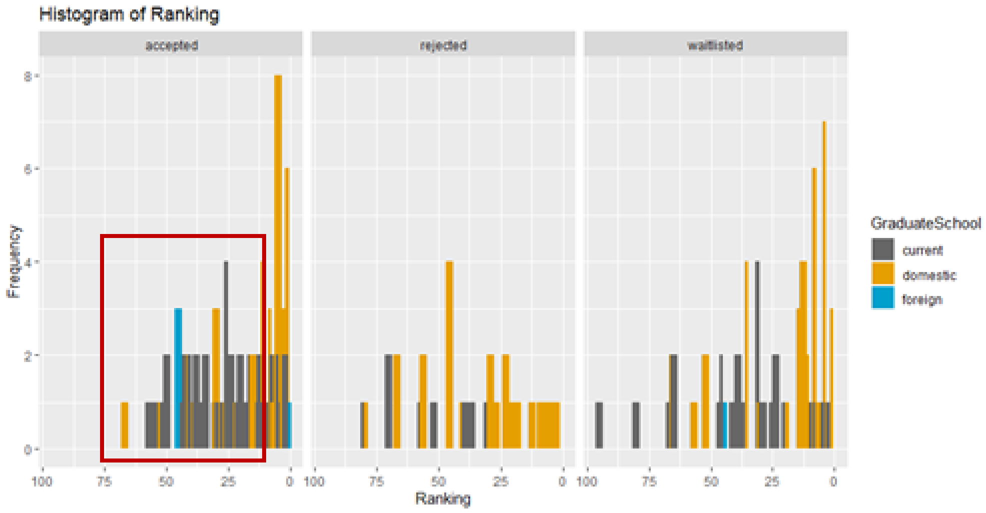
- **Histogram of Ranking**

If looking at three types of outcomes separately, the average ranking of accepted students is 20.64, then that of waitlisted students is 25.68, and the average ranking of students who are rejected is 37.75. This shows that students who are accepted by the graduate school would have higher ranking than others in average.



```
> summary(sub_accepted$Ranking)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    5.00   18.56   20.64   30.00   66.67
> summary(sub_rejected$Ranking)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.40   23.00   36.36   37.75   56.00   80.00
> summary(sub_waitlisted$Ranking)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00    9.00   20.00   25.68   38.34   95.00
```

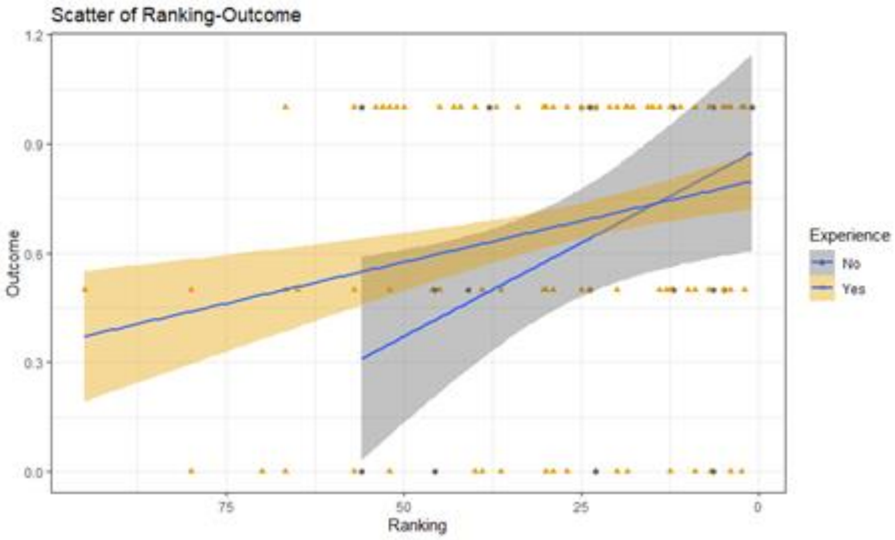
By the bar chart below, the highlighted area reveals that if students plan to apply for the graduate school same as their university, the required ranking to be accepted would be lower. That is, if your ranking wouldn't be the top 1%, you still have the chance to be accepted.



- **Scatter plot of Ranking-Outcome**

First of all, the variable "Experience" is redefined which is different from it mentioned previously. If applicants never join alumni association, school club, school ball team or department ball team, the variable "Experience" is defined as "No". Otherwise, it's defined as "Yes".

From the scatter plot below, we can see the relation between students' ranking and their final outcome. It reveals that students need to have better ranking if they don't have any experiences in order to get better outcome.

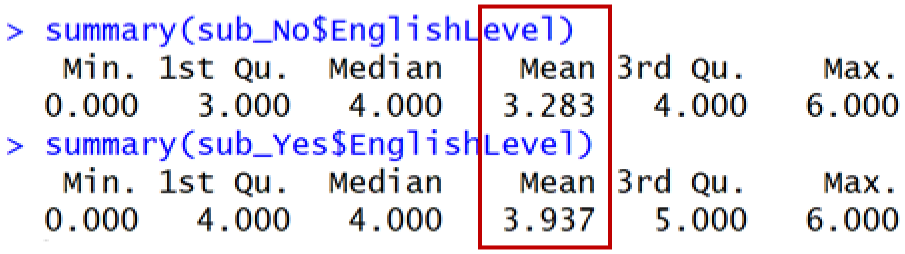


- **Bar Chart of English Level**

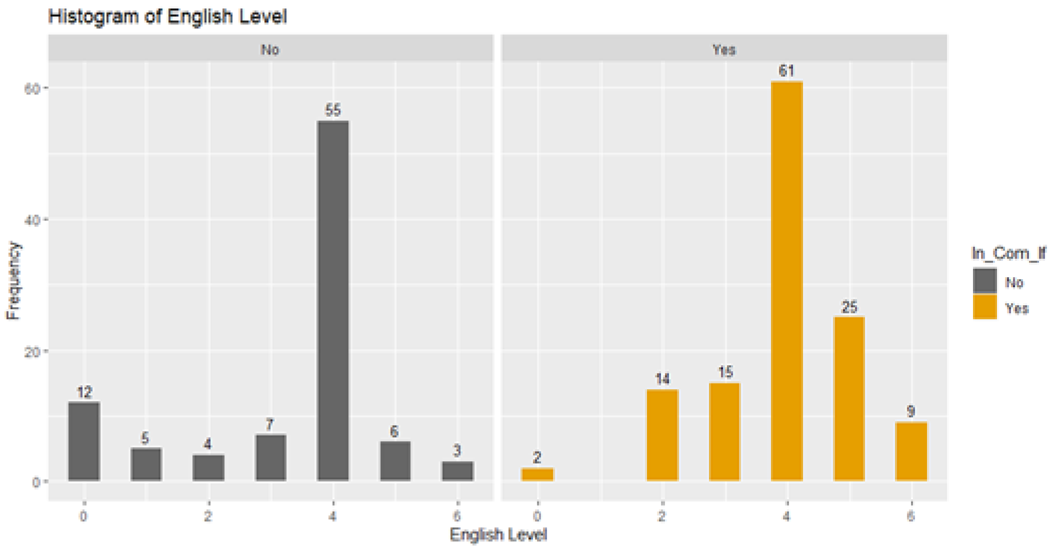
According to students' intern or competition experience, if students who don't have any intern or competition experiences, we would define as type "No", otherwise "Yes".

If looking at the histogram below, we can easily see better English level of students in type "Yes".

Also, if looking at the data below, the average English level of students with intern or competition experiences is 3.937, and that of students with no intern or competition experiences is 3.283. Apparently, students who have taken part in a competition or an intern would have higher English level.



```
> summary(sub_No$EnglishLevel)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   3.000   4.000   3.283   4.000   6.000 
> summary(sub_Yes$EnglishLevel)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   4.000   4.000   3.937   5.000   6.000 
```



6. Conclusion

After a period of hard study in the university, we finally can graduate from the school. However, what are we going to do after graduation would be the next trouble. Then, going to graduate school can be one of our choices and most of us would actually pick this option. Therefore, this project is about what really matters when we applying a graduate school.

In the US, “GRE score”, “TOFEL score” and “GPA” would be the three key factors for them to apply for a graduate school. If students can have high GRE score, TOFEL score and GPA, they mostly would be accepted by the graduate school. While in Taiwan, according to our questionnaire survey, what really matters is the graduate school students applying for, what department they’ve studied and their average score during university.

To sum up, no matter in the US or in Taiwan, students’ scores in the university seems to be the most decisive factor when making out the application for graduate school. Consequently, if there’s still time, be hardworking and never slack off during your time in the university and you would have higher chances to get the admission from the graduate school.