Name: Lo Kai Yeung
UID: 20608482

Report on the Correlation between the University Ranking and other factors

Project Objective

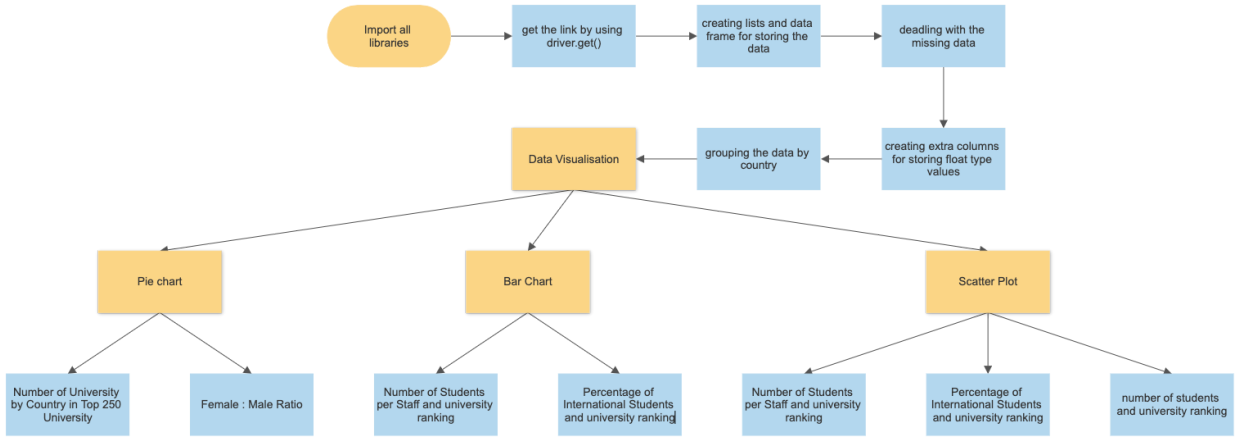
The problem of the project is regarding the characteristics of top 250 universities around the world. The board of directors may want to improve their university ranking. They may wonder what factors are correlated with the university ranking. In this project, the percentage of international students, number of students per staff and number of students are examined. The board of directors may want to improve the university ranking by making different policy decisions. University ranking offers a shorthand list of universities which provide prospective students with some information for looking for a decent university. A higher university ranking may attract more talented students to pursue their higher education in that university.

Descriptions of the Data

The data is extracted from a webpage namely, World University Rankings 2021 | Times Higher Education (THE) (url = https://www.timeshighereducation.com/world-university-rankings/2021/world-ranking). Given that the webpage on qs ranking does not contain much information regarding the university, Times Higher Education is used for the project. The data of university names, country, number of students, number of students per staff, the percentage of

international students and female male ratio. There are a total of 250 rows and 6 columns. Given that the majority of the data are in string type, they are converted into float type for further analysis. The final data frame contains 250 rows and 12 columns. The female male ratio is also splitted into 2 columns, one is the female ratio and another one is the male ratio. The data indicates that the United States has the most universities which has 69; the United Kingdom ranks in second place which has 35; Germany ranks in the third place which has 24.

Main Analyses



There are 3 major challenges that I met. The first one was that the webpage required users to sign up. It means when I use the urlopen() function, I cannot get the complete html code. Therefore, I tried another function namely, driver.get(), to get the complete html code of the webpage. Another difficulty that I encountered was dealing with the missing value. When I tried to convert the string to float, it occured error occurred as np.nan was not a string but a float. As a result, I used an if-statement to skip the np.nan. The last challenge was when I found the correlation between the university ranking and

number of students per staff. Some universities have the same ranking but different number of students per staff. To put it differently, using the same x value can come up with different y values. When I plot a scatter plot for the data, an error will occur. The solution to this problem is to treat them as different rankings. For example, Duke University and Tsinghua University are both ranked 20 so I changed their ranking. Duke University is ranked 20 and Tsinghua University is ranked 21.

Result

A total of 3 scatter plots are plotted. They find the correlation between the university ranking and number of students per staff, percentage of international students and the total number of students. For the first scatter plot, the correlation coefficient between the university ranking and number of students per staff is approximately 0.17. For the second scatter plot, the correlation coefficient between the university ranking and percentage of international students is approximately -0.20. For the third scatter plot, the correlation coefficient between the university ranking and total number of students is around -0.15. The aforementioned result shows the university ranking and the three factors are weakly correlated. It indicates the fact that if the board of directors aim to improve the university ranking, it may not be feasible to change the three factors which are number of students per staff, percentage of international students and the total number of students.

Conclusions

The problem is what factors should the board of directors consider when they would like to improve the university ranking. Although this report only takes into consideration three factors including number of students per staff, percentage of international students

and the total number of students. The result shows that there is a weak correlation between them. Put differently, if the board of directors would like to enhance the university ranking, changing the three stated factors may not be feasible. Given the fact that only three factors are examined in this report, other factors may also be investigated. The striking example would be the number of thesis written in each year, the average salary of graduate students and the number of students per professor.