Data Wrangling Final Project - Top University Record

Kanya Kreprasertkul

4/4/2020

Introduction

University rankings record is beneficial to every person who is related to education systems. For example, students can use this data to consider which university they should apply. Academic personnel can use it to improve their quality of education. High school teacher can also use it to guide their children. Nowadays, there are many online sources that publish university rankings record. They use several types of methodology to calculate ranking, such as learning environment, research performance, academic award received by faculty member, etc. As I have interested in education system, I want to demonstrate this record into graphics to make it easier to interpret and to extract some insights from it.

Datasets

The first data source is World University Rankings 2019-2020 data from Center for World University Rankings website. This website uses quality of education, alumni employment, quality of faculty and research performance to indicate the rankings. It provides top 2000 world ranking and location of each university. Also, it shows national ranking of each university.

The second data source is diversity and pay dataset from GitHub. Diversity dataset contains diversity record of 4,574 universities. It shows the number of total enrollment and the number of enrollments of each race. Pay dataset provides early career salary, mid-career salary, percent of student body in STEM and percent of alumni who think they are making the world a better place.

The last data source is University Statistics from Kaggle which is in JSON file. This dataset includes many interesting variables, such as acceptance rate, tuition fee, percent of students who received aid, etc. However, there are only 311 universities included in this dataset.

Data Preprocessing

First of all, I imported all dataset and stored them in data frames. After I investigated them, I found out that I need to perform many data cleaning. As I want to join data frames, I need to change university name which is a key that I used to join them. There are several universities that have different way to write their name. For example, “Virginia Polytechnic Institute and State University” and “Virginia Tech”. Also, each data source used different format to write university name. For example, “Rutgers University-New Brunswick”, “Rutgers University–New Brunswick” and “Rutgers University at New Brunswick”.

There are more than hundred rows of university name that I need to change. So, I decided to focus on only university in top 100 USA university ranking but I still need to clean around hundred rows in total. I used setdiff() to see what university name I need to change. Apart from changing rows name, I used str_replace() to replace some format. For example, replace “-” with “at” or “–” with “at”. After I cleaned all data and join

them into one data frame, I changed some of university name to be the official name to make it useful for further investigation.

Results

After I downloaded top 2000 world university rankings, I calculated total number of universities from each country. Table 1 shows top 10 ountries in top 2000 world university rankings. We can see that United States of America has the highest number, following by China and Japan. So, I determined to look deeper into education system in USA.

Country	Count
United States of America	358
China	249
Japan	130
France	95
United Kingdom	94
Germany	69
India	68
Italy	66
South Korea	64
Turkey	61

Table 1. Top 10 countries in top 2000 world university rankings.

Then, I changed column name and some of country name to make my data compatible with the choroplethr package. Following, I attached region data from the choroplethrMaps package and merged it with my data and then I created a choropleth map of my data as shown in Figure 1.

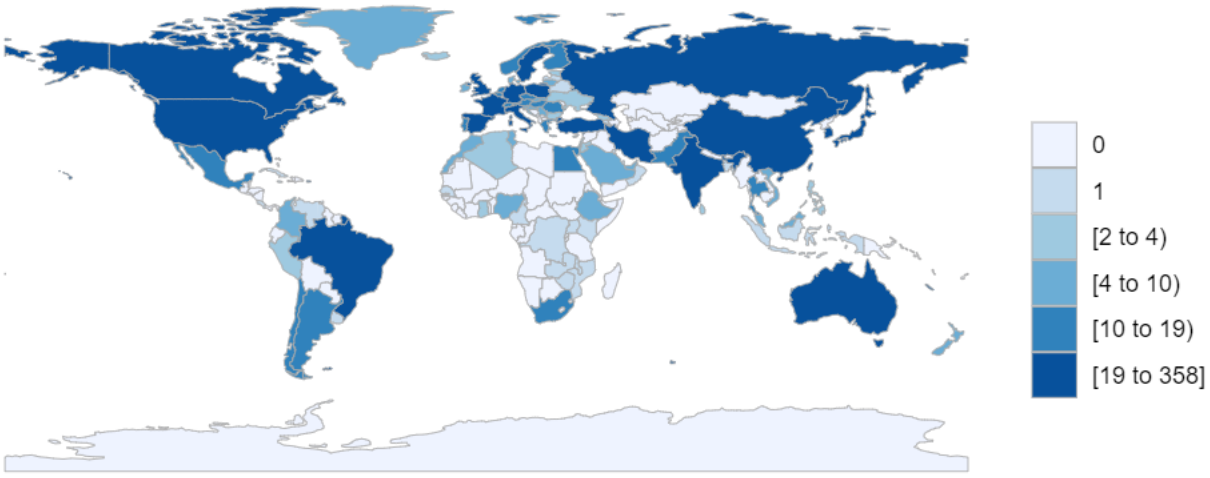

Figure 1. The number of university in top 2000 ranking by country.

After I finished cleaning data and merge all of my data source into one data frame, I calculated total number of universities from each state. Then, I ranked only top 10 state in top 100 USA national university rankings and showed it in Table 2.

State	Count
California	12
New York	8
Texas	8
Massachusetts	6
Florida	4
Illinois	4
Indiana	4
North Carolina	4
Pennsylvania	4
Georgia	3

Table 2. Top 10 states in top 100 USA national university rankings.

Then, I changed column name and some of state name to make my data compatible with the choroplethr package. Following, I attached state name from the choroplethrMaps package and merged it with my data and then I created a choropleth map of my data as shown in Figure 2.

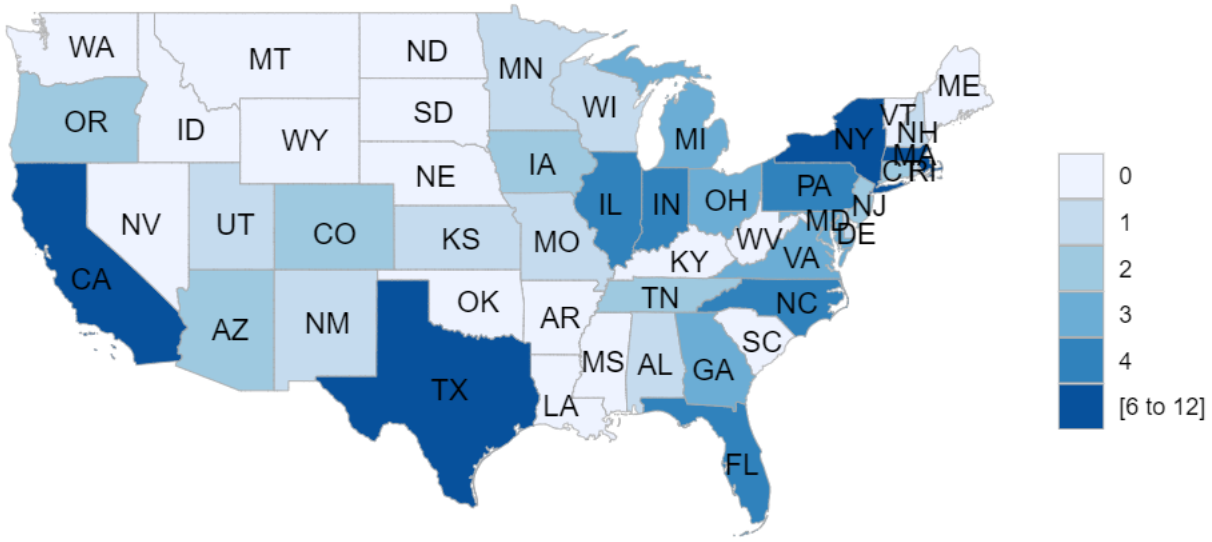

Figure 2. The number of university in top 2000 ranking by state.

As shown in Figure 3, I plotted the relation between acceptance rate and national ranking, and used stat_smooth() to see their linear relationship. Acceptance rate is the rate at which applicants are accepted. The lower acceptance rate, the harder applicants get admitted. As expected, university with better ranking has lower acceptance rate. Also, it is significantly low for university within top 10 ranking. For the relation between tuition fee and national ranking, university with better ranking has higher tuiton fee which is interesting. This might mean that higher tuition can provide better education quality. So, I want to see more about tuition fee.

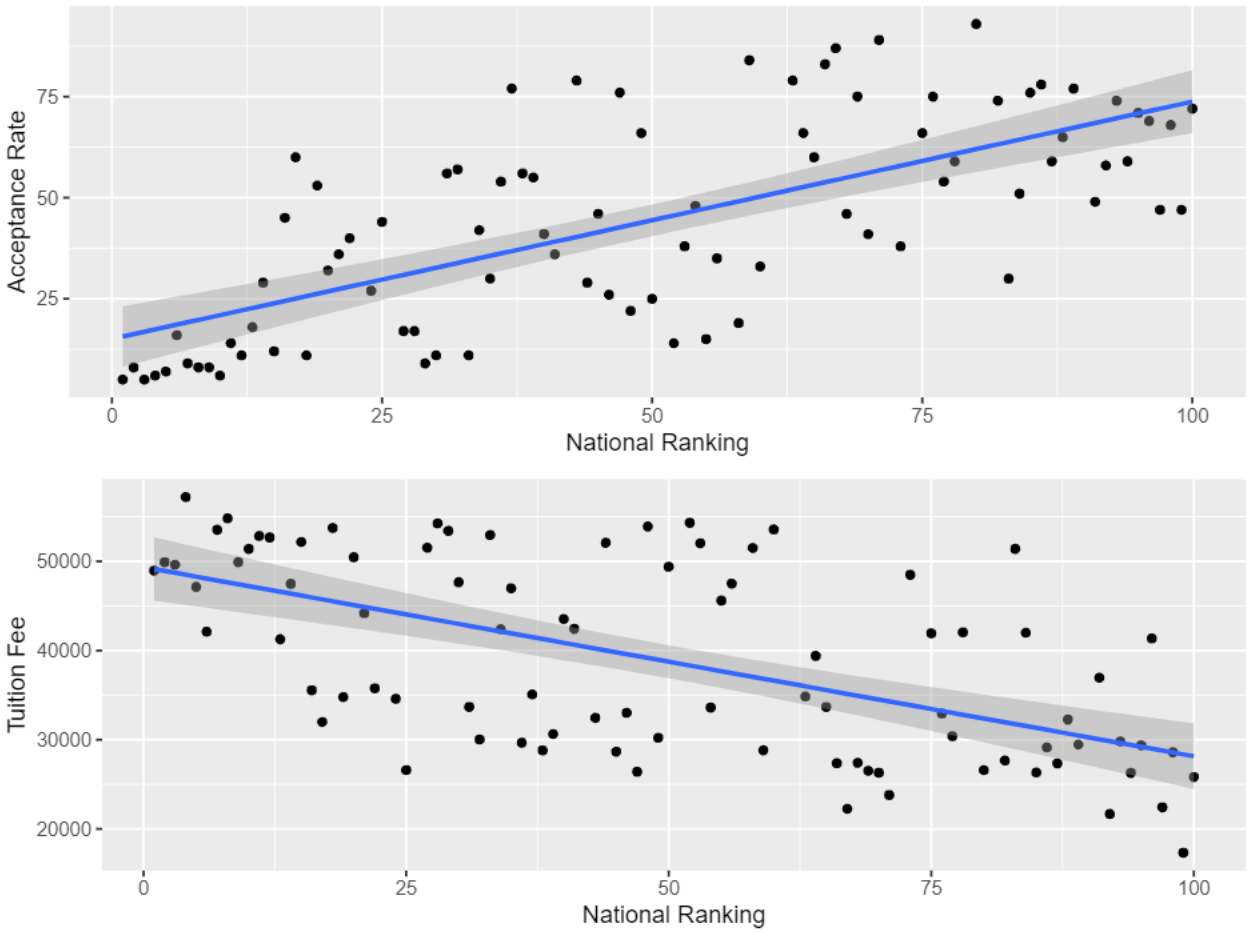

Figure 3. Upper: Relation between acceptance rate and national ranking. Lower: Relation between tuition fee and national ranking.

Next, I plotted the relation between percent of students who received aid and national ranking. We can see from Figure 4 that university with better ranking offers aid to their students more than university in lower ranking. When looking into tuition fee after receiving aid, total tuition fee of university with better ranking is lower than total tuition fee of university in lower ranking. As they give our more funding, they can attract many talented students. Also, some students are really smart, but they do not have sufficient money. This might be the reason why university in high ranking have higher education quality. So, I wanted to look more into a type of university because private university and public university have different type of funding system.

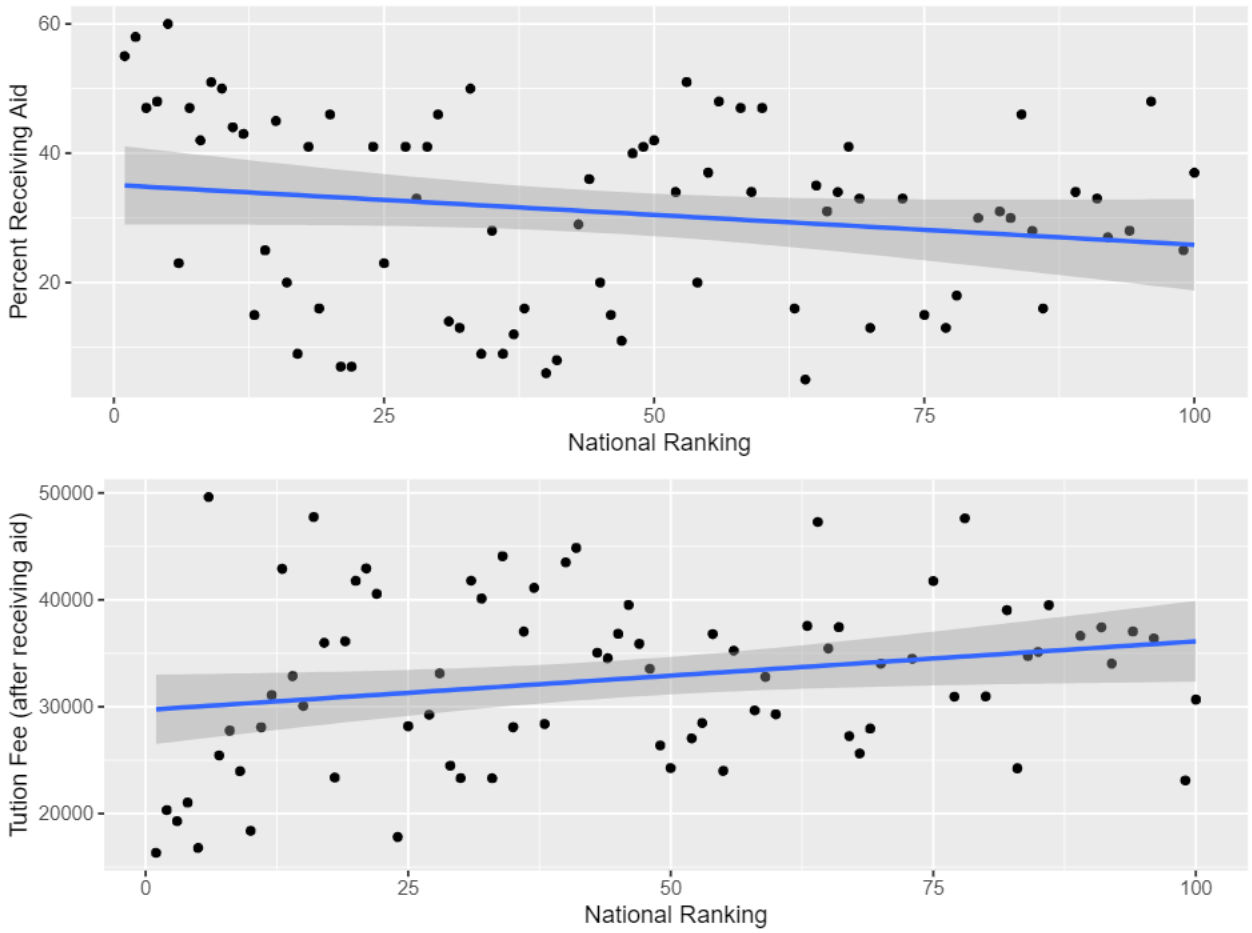

Figure 4. Upper: Relation between percent of students who received aid and national ranking. Lower: Relation between tuition fee after receiving aid and national ranking.

Public university is funded by state governments. In contrast, private university relies on private contributions. So, private university tuition fee is generally higher. To prove this claim, I plotted to see differences between private university and public university. In data preprocessing, there are 8 missing data in the column indicating private or public. So, I manually filled NA because this information can be easily searched from internet (unlike acceptance rate or percent of students who received aid). There are 33 private university and 63 public university within the 100 national ranking. As usual, private university tuition fee is higher than public university tuition fee. But, average percent of students who received aid from private university is two times higher. This means private university in my dataset receives plentyful private contributions.

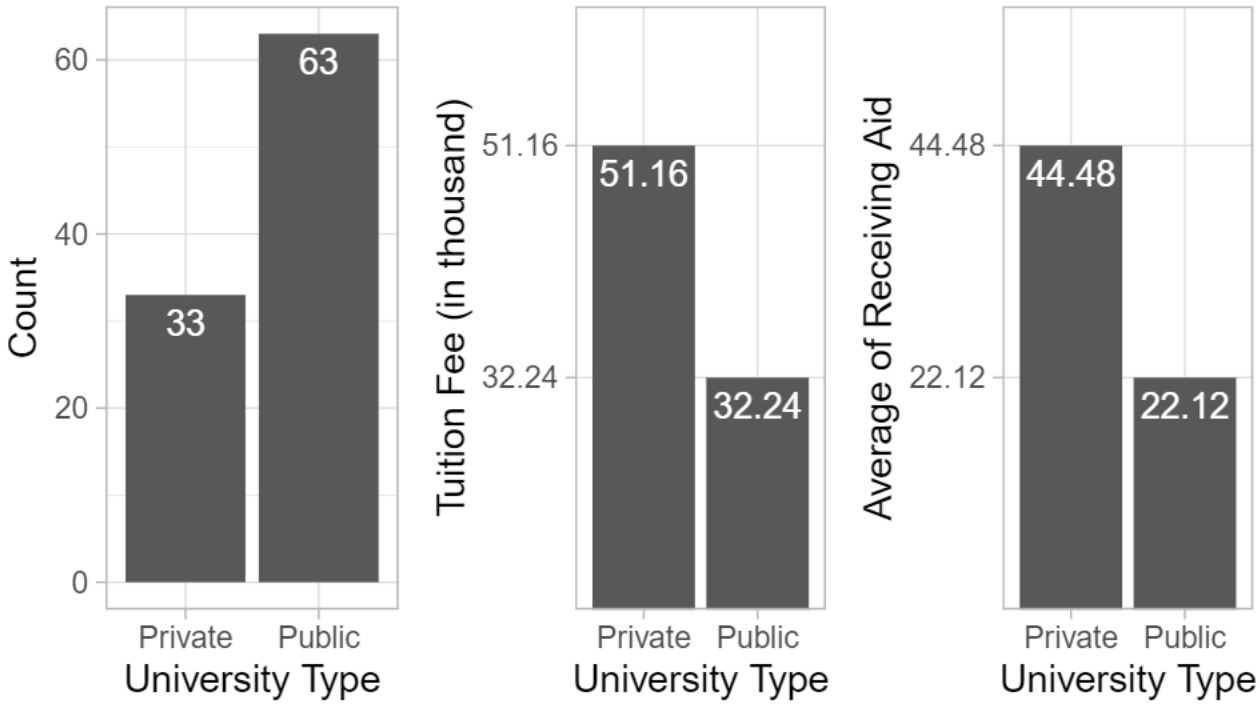

Figure 5. Left: total number of private university and public university in top 100 ranking. Middle: average tuition fee of private university and public university. Right: average percent of students who received aid of private university and public university.

Next, I plotted to see the relation between tuition fee and national ranking and the relation between percent of students who received aid and national ranking again. But this time, I colored each point by university type because I want to see their distinction. We can see that private university tends to have better rank than public university. Almost all of the top 10th university are private university. Most of private university tuition fee are higher than public university tuition fee. But, most of private university offers more funding to their students. We can see that almost half of the students in private university received aid. This might be the reason why private university has many intelligent students even though their tuition fee is high.

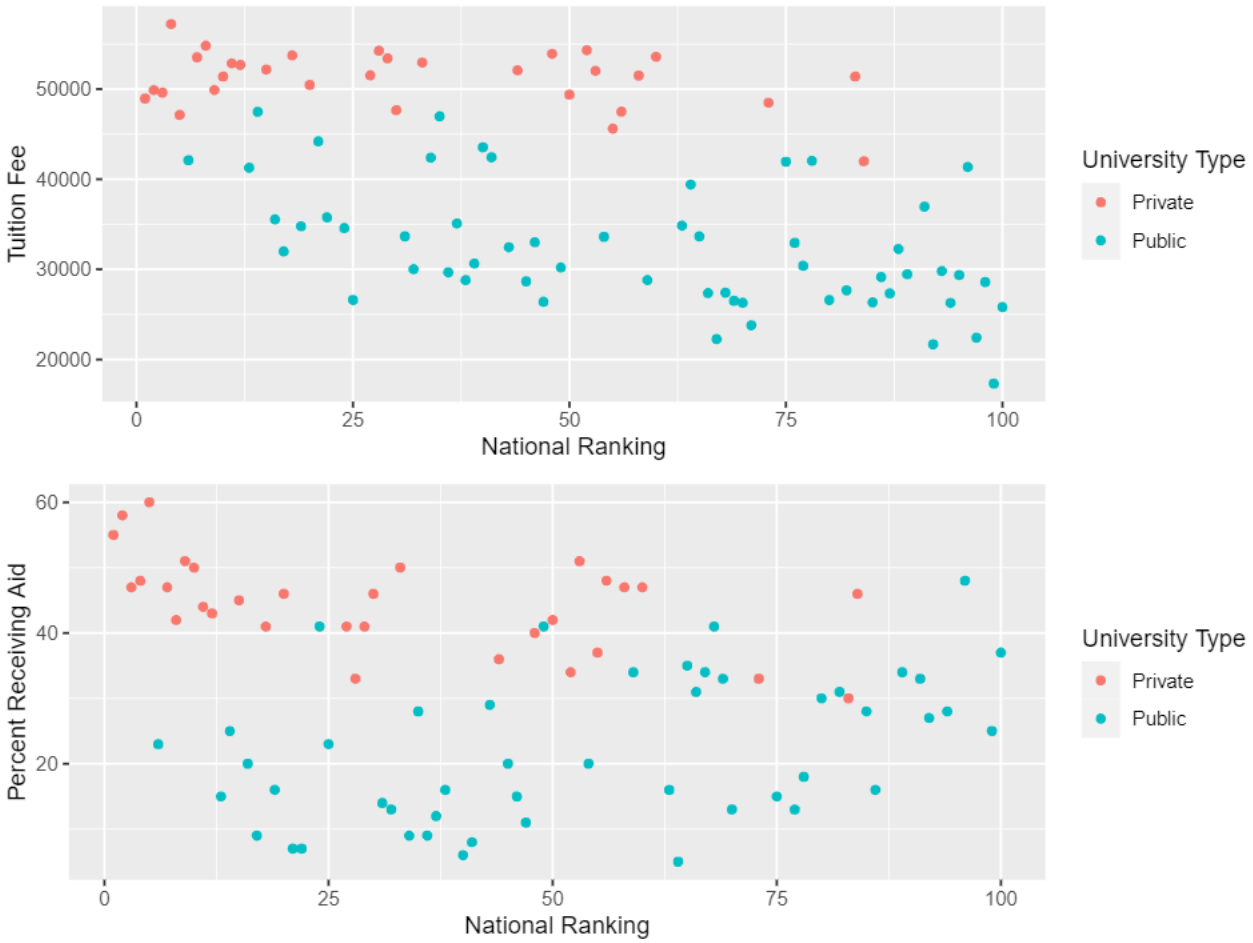

Figure 6. Upper: Relation between tuition fee and national ranking colored with university type. Lower: Relation between percent of students who received aid and national ranking colored with university type.

Diverity dataset provides many races, i.e. White, Black, Hispanic, Asian, American Indian / Alaska Native, Native Hawaiian / Pacific Islander, Two Or More Races, Non-Resident Foreign, Total Minority and Unknown. So, I chose only major race, i.e. White, Black, Hispanic and Asian. Then, I re-scaled the number to make total of 4 races to 100%. Next, I mutated the numbers and then used gather() for futher barplot.

There are many articles stating about how diversity benefit education quality. To prove this claim, I plotted diversity percentage of each university in 100 ranking. University in top 50 ranking is shown in Figure 7 and the rest is shown in Figure 9. As we found out earlier that California has the higher number in the top 100 ranking, I plotted Figure 8 to examine only university in California. There are 12 universities and we can see that they have high diversity. Also, they have high percentage of Asian students comparing to other races, which might be because many Asian people live in California.

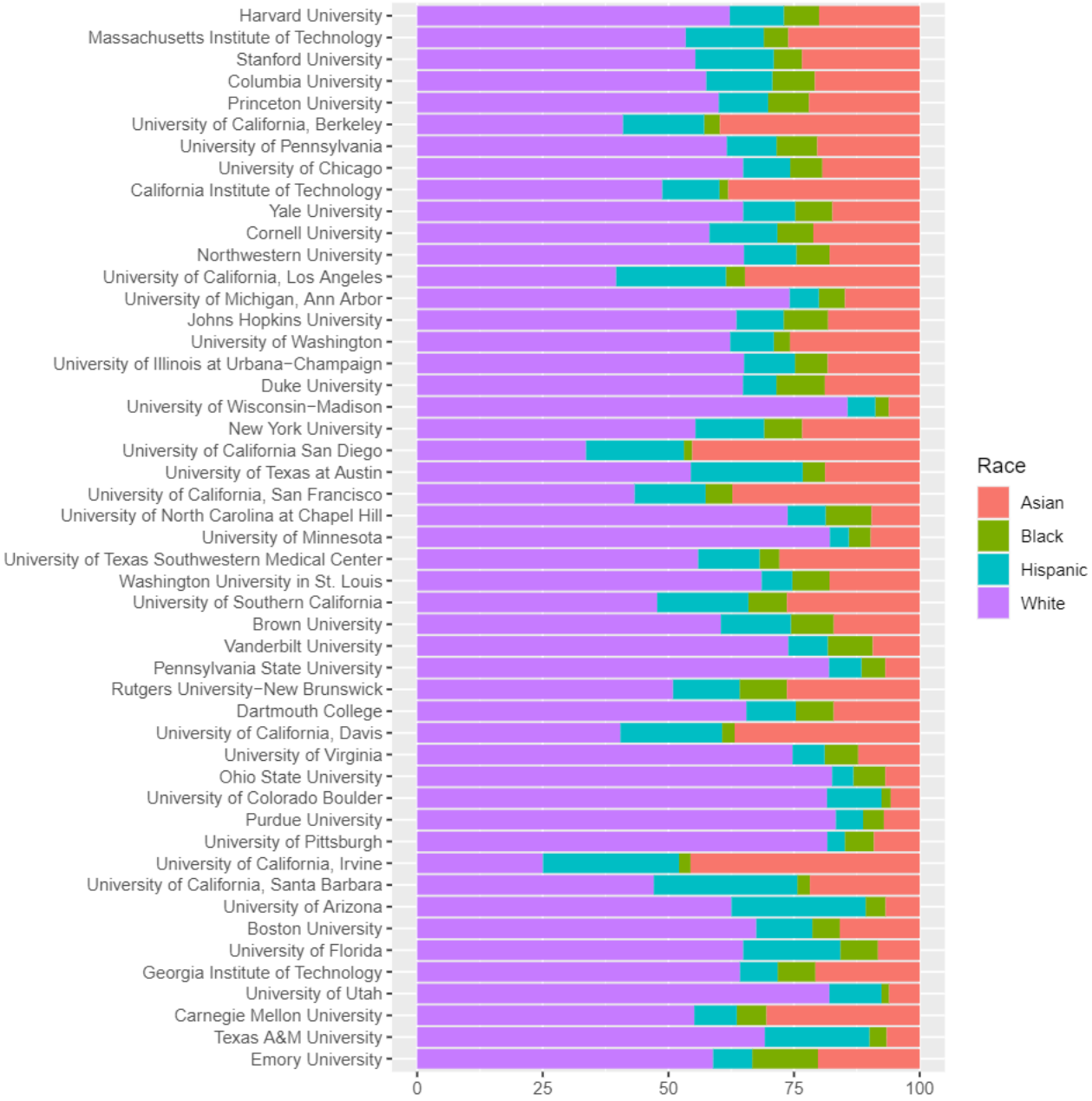

Figure 7. Diversity percentage of university in 1-50 ranking.

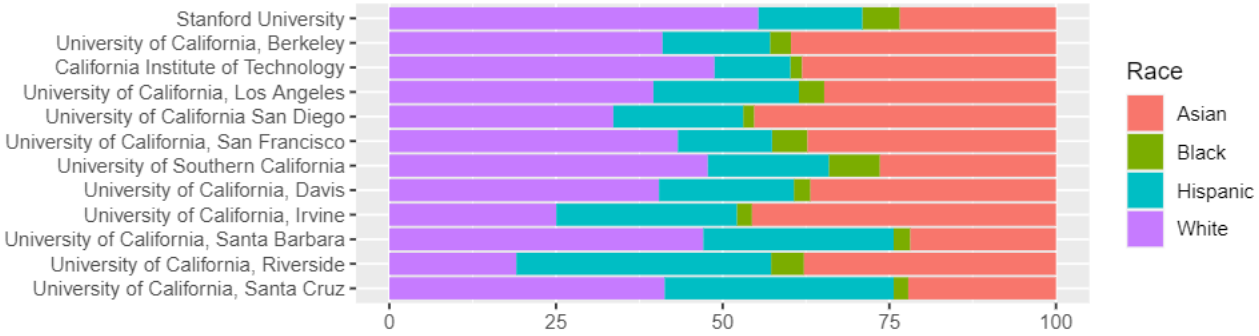

Figure 8. Diversity percentage of university in California.

For Figure 9, we can see that these universities have less diversity than top 50 university. Studying in diverse environment provides opportunity to learn from people with different backgrounds and this might lead to collaboration and innovation. It is always nice to learn from people who have difference experience and backgrounds. Also, diversity encourage students to challege themselves. Consequently, diversity may actually be the main reason for the better ranking.

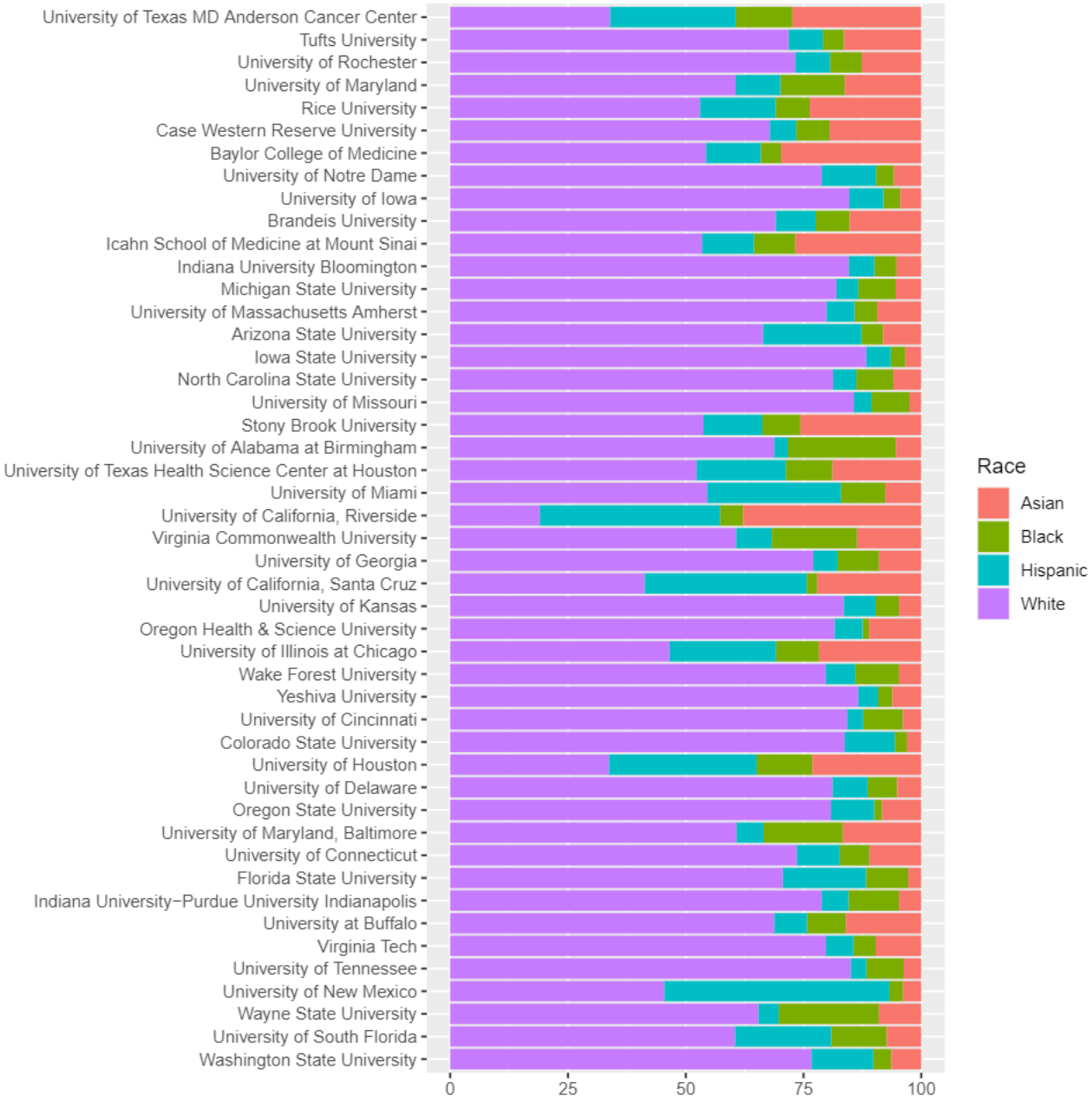

Figure 9. Diversity percentage of university in 51-100 ranking.

SAT score and ACT score are a standardized test used for college admissions in the United States. As expected, students who got into university in better ranking have higher test score. This information is useful for students who is considering applying for college admissions. As admission has a fee and application process need a letter of recommendation, students need to consider whether they will be admitted or not. The record can inspire students to keep pushing themselves if they want to get in a high ranking university. Also, high school teachers can use this data to encourage their students and make a better preparation for them.

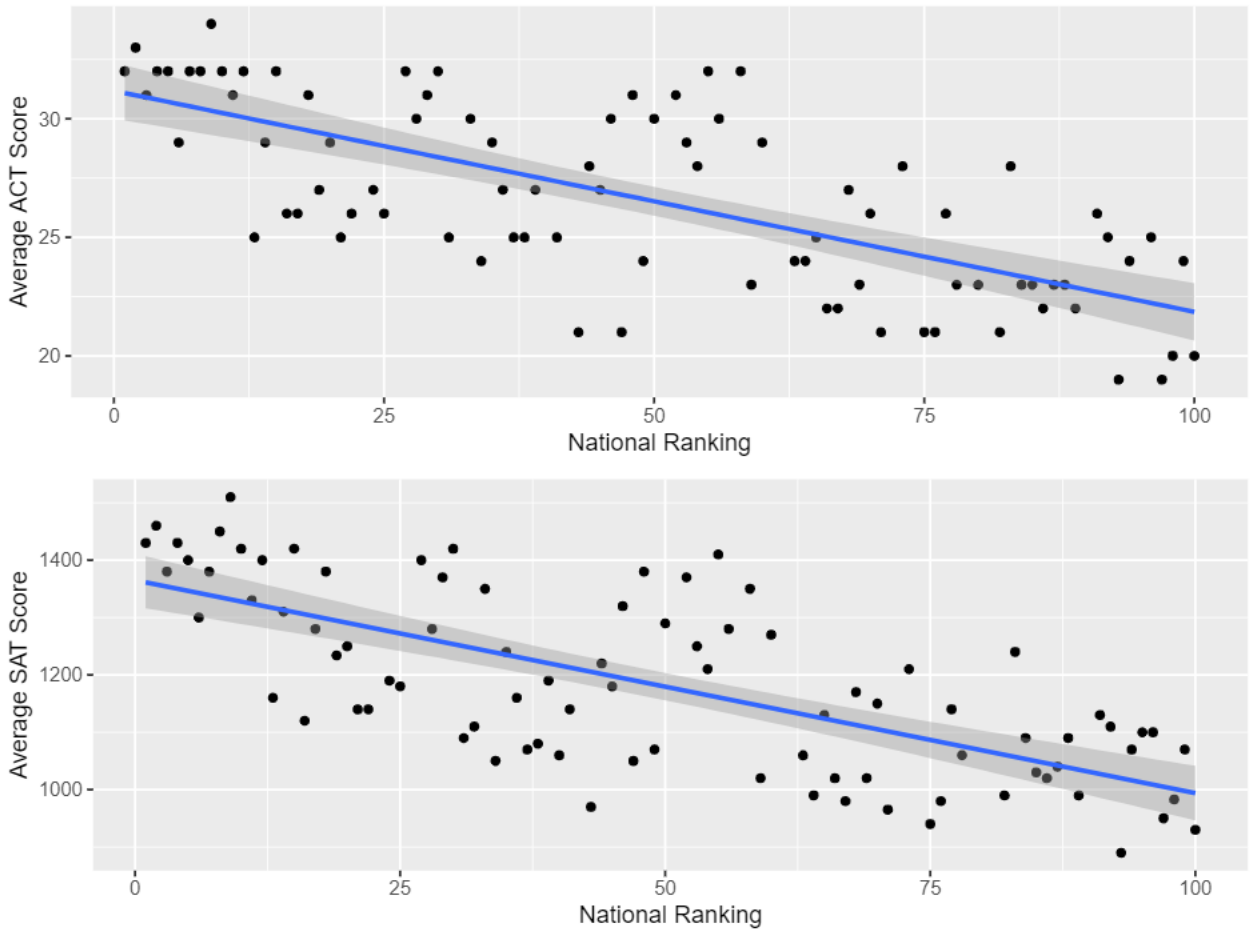

Figure 10. Upper: Relation between ACT score and national ranking. Lower: Relation between SAT score and national ranking.

Next, I plotted to see the relation between early career salary and national ranking, and the relation between mid-career salary and national ranking. We can see that people who graduted from better ranking university tend to have higher salary which is not surprising. Students graduated from famous university might be able to ask for higher salary. Also, it might be easier for them to get a job in the top biggest company. Even when they are at the middle of their career, they still tend to get better pay. This might indicate that they tend to get promotion and be succesful in their career.

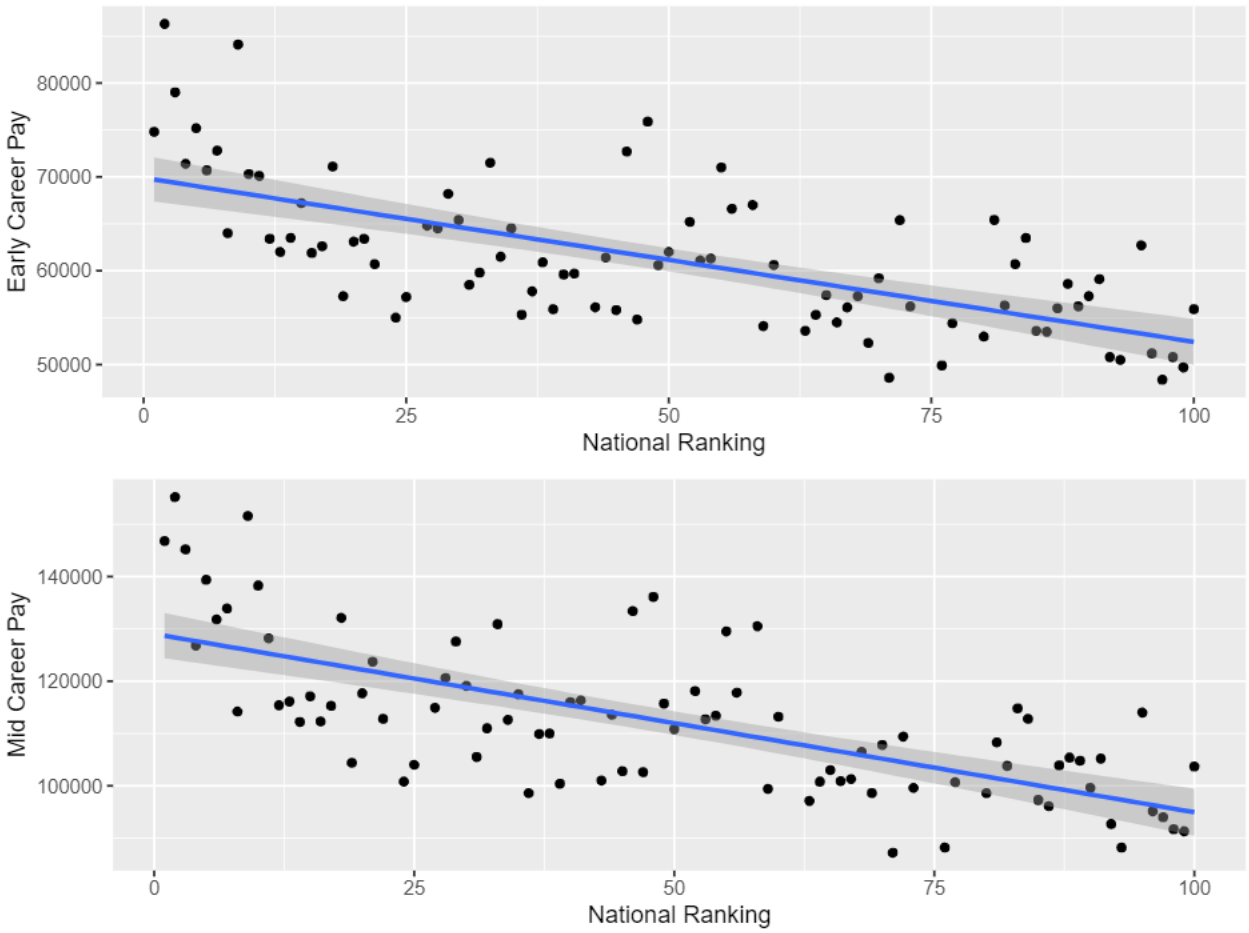

Figure 11. Upper: Relation between early career salary and national ranking. Lower: Relation between mid-career salary and national ranking.

Then, I plotted to see the relation between percent of student body in STEM and national ranking. University in higher ranking tends to have higer percent of student body in STEM, and this might be the reason why students graduted from higher rank university get higher salary. However, looking into percent of alumni who think they are making the world a better place graph, we can see that higher number of students graduating from lower rank university think they are making the world a better place. This might be able to infer that those students are happier and less stressful.

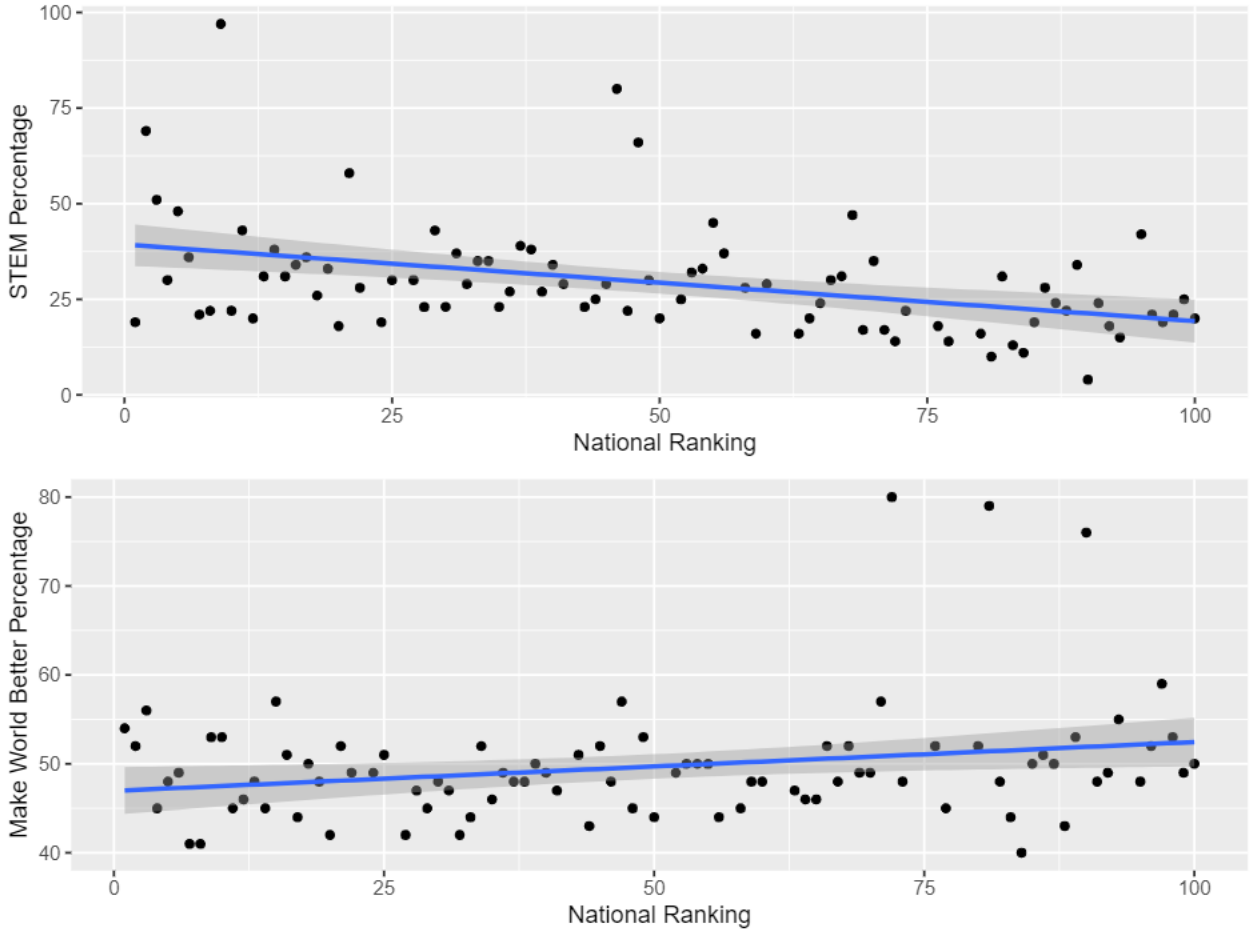

Figure 12. Upper: Relation between percent of student body in STEM and national ranking. Lower: Relation between percent of alumni who think they are making the world a better place and national ranking.

Conclusion

University ranking record is widely publish. Anybody can access this record. Although, these data need a lot of preprocessing, they provide many insights. For example, students can use this record to estimate which university they can get admitted. As every univrsity wants to develop and get a better rank, faculty member can use this data to make action, such as provide more funding to students, increase diversity environment, etc. In addition, there are many factors contributing to the ranking. By investigating all plots, funding and diversity seem to be the most important factors. However, further research should be conducted.

Reference

- (n.d.). CWUR - World University Rankings 2019-2020. Retrieved from https://cwur.org/2019-2020.php
- Thomas Mock, Jessie Mostipak. (2020, Mar 9). College tuition, diversity, and pay. Retrieved from https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-03-10/readme.md
- Christopher Lambert. (2018, Jan 21). University Statistics. Retrieved from https://www.kaggle.com/theriley106/university-statistics