Factors that Affect Graduation Rate at Universities in Washington State

Adam Klingler and Kayla Perez

Summary of Research Questions and Results

We had two big questions that motivated this project. Those questions being:

1. What is the relationship between competitiveness of an institution and its graduation rate?
    a. Competitiveness being the acceptance rate for an institution. Low acceptance rates means a high competitiveness and high acceptance rates means a low competitiveness. We found that there was a negative correlation between the two. As the graduation rate increased, acceptance rate decreased, which meant that school applications became more competitive.
2. What is the relationship between students needing financial aid at an institution and its graduation rate?
    a. There was a negative correlation between these as well. As the graduation rate increased, the percentage of students who needed financial aid decreased.

Motivation and Background

Competitiveness in an institution's applications process is usually seen in a positive light. Popular wisdom would suggest that the more difficult a school is to get into, the more likely students are to graduate. To investigate this wisdom, we wanted to analyze the data for some of the major universities in the state of Washington. If true, then as the successful application rate decreases, the graduation rate should increase.

Financial aid is the barrier to many students going to the school of their choosing. With this aspect of our study, we wanted to investigate whether having more students on financial aid would correlate to higher graduation rates.

The results of this study are useful to both prospective students and educational institutions. For prospective students, this study will provide another tool for them to analyze a school, and to provide some data to back up what they hear from family and friends. For institutions, this will help them guide future policy towards a path with higher graduation rates.

Dataset

For our project, we needed to combine 5 datasets split into 6 parts. These datasets contain different information per post-secondary institution in Washington State. We obtained these datasets from DataPlanet.com and manually had to select the data we wanted to include in the download. Unfortunately, DataPlanet isn't very easily accessible to the general public. It requires a log in, which can be done through the UW Libraries website, but that requires you to individually select and download each dataset.

For convenience, we've linked our GitHub Repository, where we downloaded and stored the exact datasets used to create our results. The datasets themselves include a link back to their original source in the second row of the respective excel file.

GitHub Repo: https://github.com/AdamK42/cse163-project/tree/master/datasets

Applicants per Institution:
"total_applicants.xlsx"
This dataset contains the total number of college applicants per institution in Washington between the years of 2001-2018.

Admitted per Institution:
"total_admitted.xlsx"
This dataset contains the total number of admitted college students per institution in Washington between the years of 2001-2018

Undergraduate student population per Institution:
"student_population.xlsx"
This dataset contains the total college undergraduate student population per institution in Washington between the years of 2003-2018.

Graduation Rate per Institution:
"graduation_rates.xlsx"
This dataset contains the total undergraduate graduation rates per institution in Washington between the years of 2003-2018.

Undergraduate Financial Aid per Institution:
"financial_aid_public.xlsx" and "financial_aid_private.xlsx"
These datasets contain the number of students that received Title IV financial aid per institution in Washington between the years of 2011-2017.

The data we include was only the 4 year institutions, as many of the technical schools and community colleges did not have enough data available to work with. These were combined into 1 dataset as is explained below.

Methodology

The process started off by combining the datasets we got from online. Each dataset was downloaded as an excel file and was read in as a pandas DataFrame using the built in excel file reader. Then, we stripped the columns to get the names of all of the universities we had data for. Using these names and the raw DataFrames, we constructed a dictionary for each of the statistics. Then, we created new DataFrames from each of those dictionaries and merged them together by school name and year. The financial aid data was separated by public and private, so we needed to specify suffixes for the financial aid columns. Private schools have the suffix "_private" and the public schools have the suffix "_public".

To start off answering Question 0, we needed to compute the acceptance rate percentage for all of the schools. This was done at the beginning when creating the main DataFrame. Then, the DataFrame was filtered down to a smaller one that only included the columns of interest: School, Year, and percent_accepted. After an initial plotting of this filtered DataFrame, we noticed that there were some schools that were visibling lacking some or all of their data, but were still showing up in the legend on the graph. Visual inspection of the underlying DataFrame showed that major schools had at least 15 data points, so we filtered the DataFrame to drop any schools with less than 15 data points. Once we had the correct amount of data, we could move onto plotting.

To answer Question 1, we filtered our main DataFrame to include only the year, school, population, the financial aid columns, and the graduation rate. For the calculations to be valid, 5 of these columns need to be populated, so we dropped the other rows from the data that didn't meet this requirement. To move all financial aid numbers into 1 column, we filled all NaN spots with 0 and took the absolute difference of the public and private column into a new column. This new column was used to calculate the financial aid ratio, which is the percentage of the student population using financial aid. At this point, we got some weird values. Some of our data showed that more students were on financial aid than went to the school, which didn't make sense given the context. To fix this, we filtered out values that were greater than or

equal to 100 percent. Then, after a visual inspection of the data, we discovered that the major schools all had at least 7 years of data, as a lot of the smaller schools were missing several data points.
We had low confidence in including the schools with missing data points, with fears that it would skew our data. We decided to filter out schools with less than 7 years of data to fix this issue. Once this was done, we moved on to plotting.

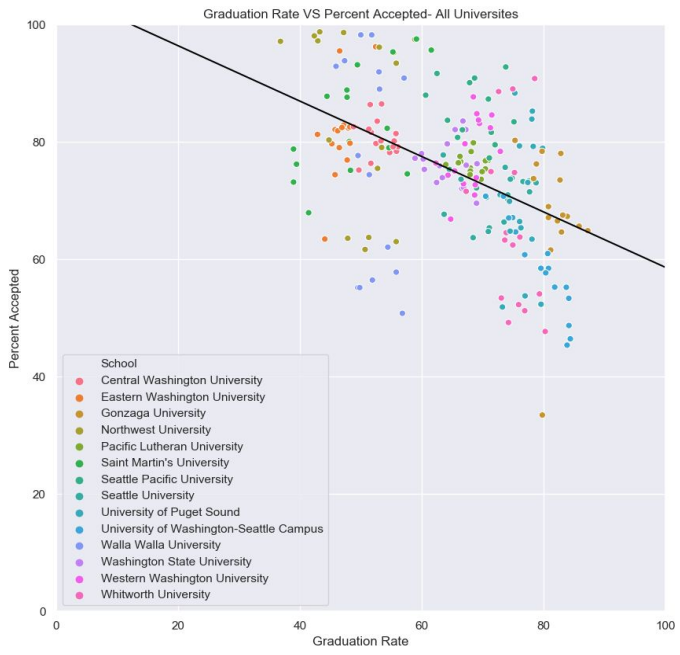
We decided on having 6 plots for each question. 3 plots to show graduation rate vs competitiveness/financial aid ratio over time for all schools, public schools, and private schools, and 3 plots to show average graduation rate vs average competitiveness/financial aid ratio for all schools, public schools, and private schools. These were plotted and labeled accordingly. Then, a trend line was plotted on top of our graphs to show the relationship between are two variables. We also calculated the Pearson's R to see how strong of a correlation there is, both mathematically and qualitatively.

Our methodology for each question was very similar, with a few minor differences. This is heavily reflected in our code, where we made use of factoring and helper functions to do most of the generic work that both analyses shared.

Results

Question 0: What is the relationship between competitiveness of an institution and its graduation rate?

For this project, we defined competitiveness to be the percentage of applicants admitted to a school. If a school's acceptance rate is **low**, they have a **high** competitiveness. If a school's acceptance rate is **high**, they have a **low** competitiveness. After plotting, we found that generally, there was a negative correlation between acceptance rate and graduation rate. This was the case for all universities, both public and private.

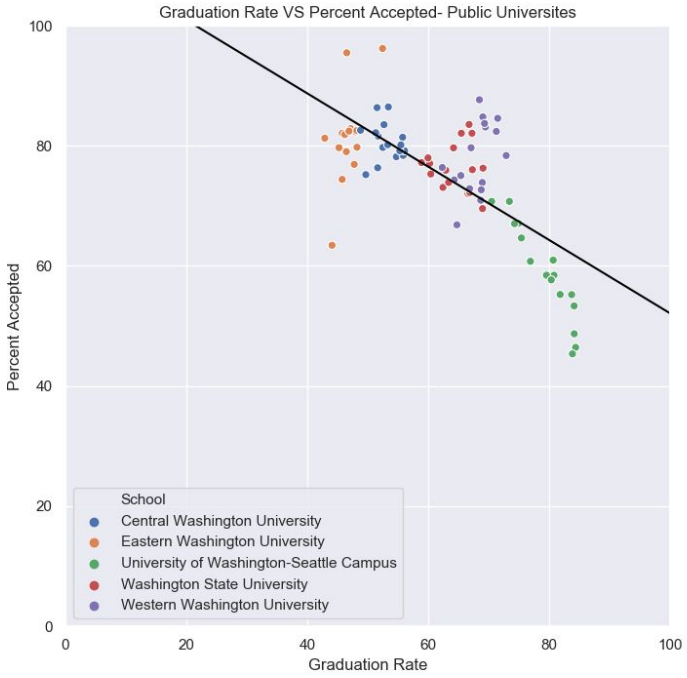


Plotted with plot_grad_rate_vs_competitiveness

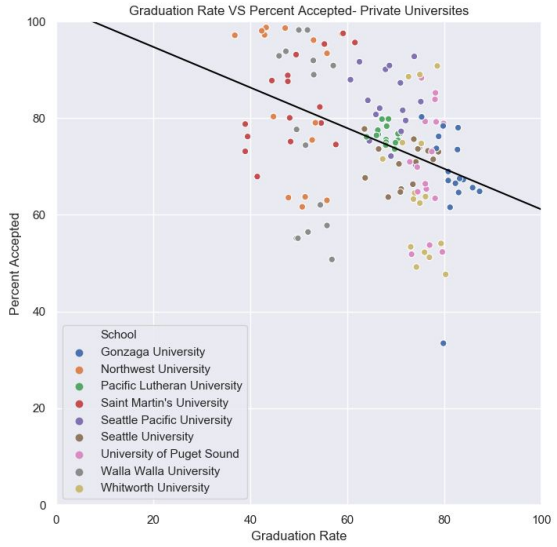
Looking at all of the universities together, it's fairly clear that this trend applies to everyone. As a school's graduation rate decreases, the percentage of students accepted tends to increase. Even though this correlation wasn't terribly high, with a Pearson's R of -0.485, this led us to believe that competitive applications in schools were effective in leading more students to graduate. Overall, this wasn't very surprising. Society and general experience tells us that if you go to a school that's more difficult to get into, you're more likely to succeed and graduate from that school. Our results show that that tends to be the case.

Things get more interesting when we look at the difference between public and private schools. While the trend is pretty much the same, looking at certains and some of the numbers tells an interesting story.

To the right is a plot of graduation rates versus acceptance rates over time for public universities. We found it incredibly surprising to see just how far away University of Washington (UW, in green) was from the rest of the schools. Out of all of the public schools in Washington, UW has the highest graduation rates, while also having some of the lowest acceptance rates. This plot also had the highest correlation, with a Pearson's R of -0.705.

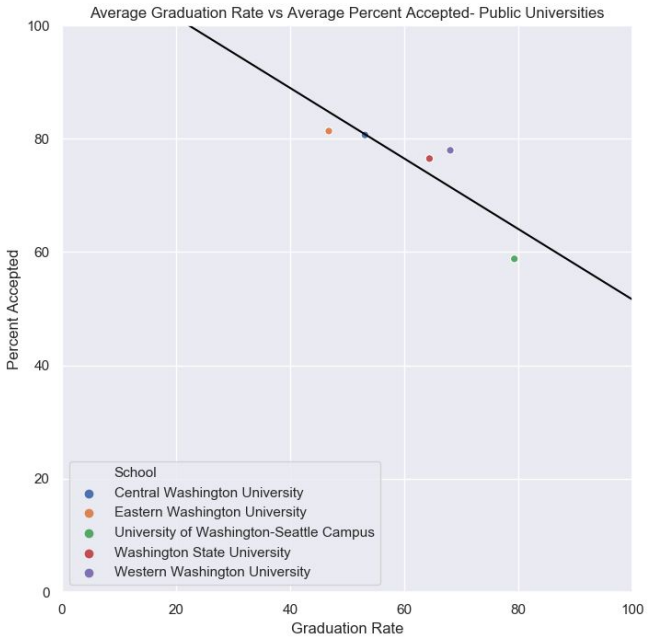


Plotted with plot_grad_rate_vs_competitiveness_public



Plotted with plot_grad_rate_vs_competitiveness_private

Our
private school plot tells a similar, but less exciting story. With a Pearson's R of -0.406, the correlation is generally the same, just not as significant. All of the private schools are more mixed together as opposed to the plot for public schools. None of the schools stand out from each other, but the general trend still stands.



Plotted with plot_average_grad_rate_vs_competitive_public

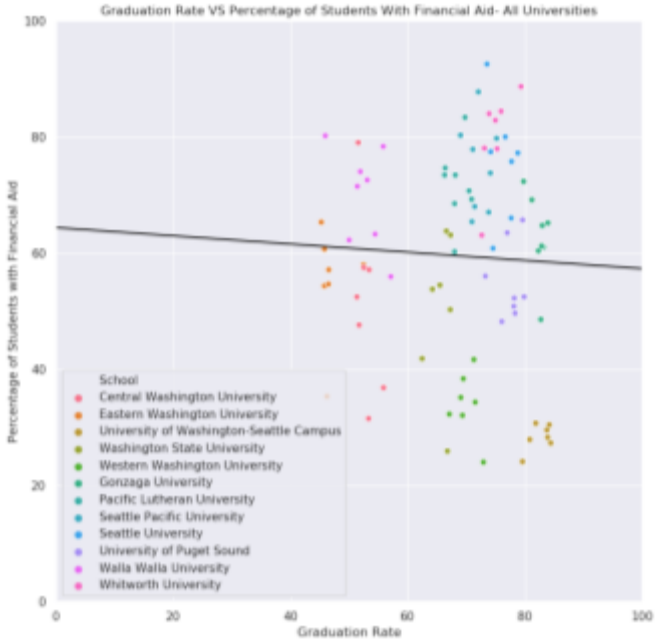
Looking at average graduation rates vs average competitiveness, our findings are a bit more concrete. Our plot for public universities had one of the highest correlations, with a Pearson's R of -0.855. It's really easy to see that UW is the outlier in this case, whereas the other schools are closer together.

Being students of UW, we know firsthand how competitive the application process is. Many students who get accepted have high SAT/ACT scores, or they take a lot of AP classes. Students who apply to UW tend to have a higher GPA than those who apply to schools on the bottom end of the plot, like Central Washington University (blue) or Eastern Washington University (orange).

That isn't to say that the other schools aren't as "good" as UW is though. A lot of what makes a school "good" is dependent on personal preference and judgement. The point of our research wasn't to find the best schools in Washington. We wanted to give prospective students some insight as to why school applications can be so competitive. As our analysis suggests, there is one good reason for schools to have a more competitive application, as it does tend to lead to higher graduation rates. But, there are many more factors that a student should consider in making their decision.

Question 1: What is the relationship between students needing financial aid at an institution and graduation rate?

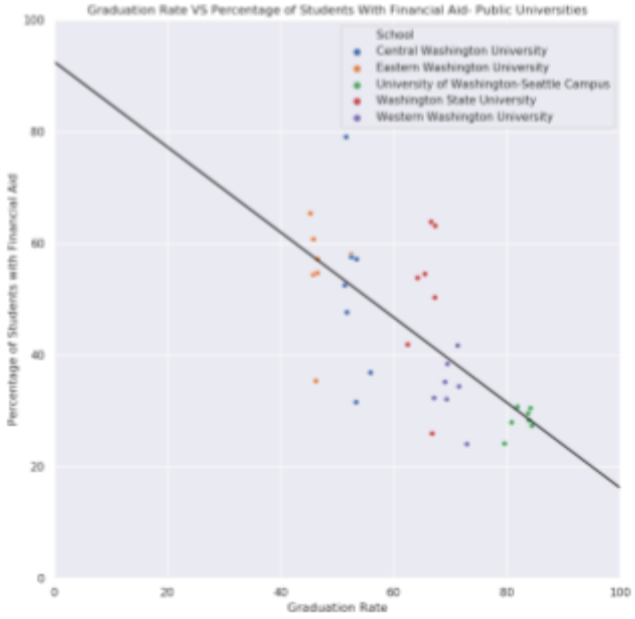
For this question, our axes are fairly straight forward. For the x-axis, we have graduation rate and for the y-axis we have the financial aid percentage. As mentioned in out Methodology, some of our data suggested more people had financial aid then went to the university. These values were dropped in all calculations, but it is important to note that they did exist.



Plotted with plot_grad_rate_vs_financial

The graph on the left shows a plot for all universities, where we can see that the best fit line does not describe the dataset very well. With a Pearson's R of -0.046, there is effectively no correlation when considering all schools. As we will see, this is primarily due to private institutions since they have such a high cost to enter.

The graph on the right is for public universities and is far more informative. With a Pearson's R of -0.679, there is a clear trend in the data. We have no definitive understanding of why this correlation is, and would probably need to do further investigation into socioeconomic backgrounds versus graduation rate to have a conclusive answer. Looking at external factors would be necessary before making any definitive calls on why this is the pattern.

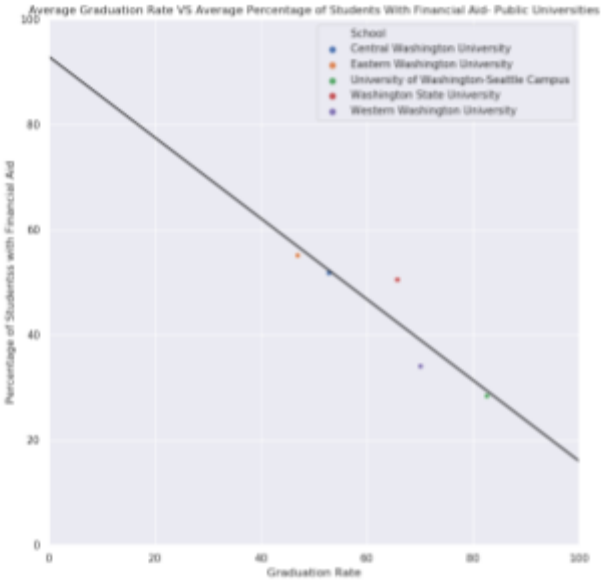


Plotted with plot_grad_rate_vs_financial_public

Plotted with plot_grad_rate_vs_financial_private

The graph on the left is the plot for the private universities. This data also has a low correlation to the best fit line with a Pearson's R of -0.180, however upon visual inspection it is interesting to note the overall very high percentage of students with financial aid. This is most likely due to the very high cost of private education, and the prevalence of academic scholarships available for these private schools. The high graduation rates of these private schools also contributes to the lack of correlation, especially when looked at in comparison to the public schools.

Finally, the plot to the right is for the average graduation rate versus the average financial aid percentage of public schools. While this plot does not provide us any new insights compared to the previous public school plot, it is interesting to note that this line is the best fit of all the plots we generated, with a Pearson's R of -0.913. This may be due to the low number of data points for this graph, but it reinforces the need for additional study into socioeconomic background versus graduation rate, as this correlation is highly intriguing.



Plotted with plot_average_grad_rate_vs_financial_public

Reproduction

The first thing you'll want to make sure you have is all 5 of our datasets, which can be downloaded from the link above. You'll also need to have our Python files downloaded, which are included with this report. Those files being "data_cleaning.py", "creating_dataframe.py", and "plotting.py". Put the dataset excel files into a folder labeled "datasets" and make sure this folder and all the Python files are together in the same spot.

Open your command prompt and navigate to the folder that has all the necessary files in it. Once your command prompt is in that folder, type "python plotting.py" to run the plotting file, which has the main method pattern in it. This will give you all 12 plots, which will save to the folder that "plotting.py" is in. It will also display the Pearson's R in the console for each of the 12 plots.

Work Plan Evaluation

Our work plan began with us trying to put our data into a MultiIndex DataFrame. This seemed like a good idea as a MultiIndex DataFrame would fit the shape of our data's space very well. However, once we got the DataFrame put together, we immediately ran into problems trying to get any meaningful work done with it. We were having trouble fully understanding functions that went with using a MultiIndex. Certain functions were convenient at times, but they were "read only", so not entirely useful. We realized that we were unable to actually use the MultiIndex DataFrame the way we intended. We decided to abandon the idea altogether and retrofit our code to construct a regular DataFrame with repeated columns. While it wasn't what we were envisioning, this solution allowed us to use more familiar functions and algorithms to produce our results and accomplish our analysis.

Unfortunately, the time spent on the data structure portion meant we ran out of time and had to cut part of our original analysis plan. Since we switched to a more conventional DataFrame structure, our calculations and plotting was simplified somewhat. We were able to get this portion of our work plan done in just 2 days or so, calculating our necessary values and producing plots using techniques already taught in the class. While we did end up fumbling with seaborn to figure out which function we needed to use to produce our plots, we did end up

with a solution that allowed us to produce our own best fit line and scatter plot on the same axis. We chose to solve our own best fit line since we did not need the error bars that the seaborn best fit produces.

Due to the limited time we had to complete our project from the MultiIndex DataFrame confusion, we had limited time to finalize our report. We decided to use Google Docs instead of Jupyter Notebook so that we could write the report in parallel easier, and because we decided that seeing the code along with our interpretations was unnecessary. Since we are just importing the images of our plots instead of importing code, we were able to save some time in this section as well.

Testing

We did almost all of our initial development in Google Colab. We wanted to use an interactive Python shell in order to see the output our code created. Since data analysis is a very visual process, we thought this would be helpful in testing our code.

The only "testable" functions we wrote for this analysis were the two inside the "data_cleaning.py" file. This being that if these two functions did not work in the way we expected them to, our entire analysis would be incorrect. These functions made up the basis of our main DataFrame, so it was crucial that they worked correctly.

While writing the code, we tested as we went. However, instead of writing traditional tests, we wrote small chunks of code and visually inspected the output it produced. This is the reason why we chose to develop in Google Colab; it made looking at the output quick and easy. We knew what we wanted our output to look like, so if it looked incorrect, we'd step through the code and find out what went wrong. This ended up being an effective way to test, as we managed to get our expected output fairly quickly.

We have provided some test functions using the assert_equals function from "cse163_utils.py" in the "test_data_cleaning.py" file to show you that our code does work as expected. If these two functions did NOT work as expected, our DataFrame we used for our analysis wouldn't be an accurate representation of our data, and thus our analysis would be invalid.

Collaboration

This project was completed entirely by Adam and Kayla. We had no outside help, aside from various users on Stack Overflow and authors of Pandas and Seaborn documentation.