World University Rankings

Ashvini Patel
Data Science

ABSTRACT

With the urge to out shine one another and the extreme competitiveness in education, it has become an important aspect in a global standing. It's extremely difficult and challenging to choose from all the prestigious universities across the country, hence the reason for coming up with world ranking of universities. Some of the most reliable sources come from Times Higher Education and Center for World University Rankings. It insures to look at various methodologies when determining the ranking. There are many considerations that the create the ranking based on the data that is collected from the universities.

INTRODUCTION

Over the years education has become an important aspect and especially now because most believe there is nothing without education in this world. Education is the ability to gain knowledge that is abundantly accessible and being able to apply it to the world around us and change it into something better. It develops a perspective of looking at life to make it easier to survive and sustain (Doumbia).

When I think of education, I recall back to the many efforts my parents put in to get me to this stage. Senior year of high school is the most stressful time when it comes to finding the best college that suits you and your career. From hundreds of universities around the world, how can we make the best choice for ourselves? I put in tons of efforts on findings ranking for each university in many different aspects to finally make a decision for my future, whether it was career based, university ranking, student population and much more. This lead me to become more curious in knowing the thought process behind all the rankings that we see.

Thus, I did my analysis on World University Rankings to truly find out what country has the most diverse education and the ethics behind rating a university. There are hundreds of universities across the world that are available for all the students, yet what makes us choose the universities that we do?

World University Rankings project is about analyzing a dataset to understand what the rankings are based from. The objective of this project is to be able understand and clean the dataset, create an analysis and work on different findings of the results with graphs and regressions. Based off the research the data I was able to get results to my conclusions.

LITERATURE

Within the dataset, there are many top statistic groups that have been able to collect data about universities to make a well-organized and sufficient table. It's important to understand the reliable sources that use accurate information to determine ranks.

One popular data collector being Times Higher Education (THE) World University Ranking is widely regarded as one of the most influential and widely observed university measures (O'Neill). As of 2019, THE World University Ranking, is definitive list of the top universities globally, including more than 1,250 institutions and 86 countries (Bothwell). It also is the only global university league table to judge research

intensive within the rankings. Times Higher Education works hand in hand with universities to create a reliable source of information. More than 500 universities from around the world submit data to THE in order to be placed on the rankings (Bothwell). This ranking does not only provide a new mechanism for assessing university performance that is vastly different from the methodologies behind research-focused, but also creates a new way to verify the data (Bothwell). Although the universities are in charge of submitting the data, THE is also looking for evidence to back up the information that is submitted (Bothwell). This ensures that the information is valid and reliable to back up the yearly rankings.

Academic Ranking of World Universities (ARWU), also known as the Shanghai Ranking, is an equally influential ranking (O'Neill). The initial purpose of the ARWU was to find the global standing of the top Chinese universities, but ended up attracting a great deal of attention from other universities, governments and media around the world (About). With the help about the major rankings based on research, patents and publications, AWRU help increase the rate for higher education across the countries. This led to more considerations on research and publications opportunities as student. The data was collected with evidence in order to create accurate rankings of the universities.

Center for World University Rankings (CWUR), is a less well known listing (O'Neill). CWUR is a leading consultant organization in providing policy advice and strategic insight (World). The main purpose was to just rank the top 100 universities in Saudi Arabia, but slowly became a worldwide demand. As the years went on the report increased and in 2019 the ranking was from 1-2000 top universities worldwide. Since this is a collaboration with THE and ARWU it is entitled to be a reliable resource when collecting all the university data.

It's important to understand that the top university rankings are determined by various

methodologies and not the basic ones like population, tuition and education. The various organizations were able to take a lot of other credentials to consider the ranking. THE, ARWU and CWUR are the most popular organizations in collecting data to create rankings by country and also worldwide.

HYPOTHSIS

Since education is really important in the 21st century. It's crucial to look at values that rank universities as a student. I found it interesting to learn about rankings since I am currently a student at a university. Initially I started off with making only one hypothesis.

1. What measures lead to higher ranking of the universities?

This is probably the typical hypotheses one should make in terms of just looking at the data. Although this is an interesting analyzation to see what the main aspects are in ranking universities. However, I did decide to add a second hypothesis because I was curious to learn what country or countries had the most distributed universities. So I added a second hypotheses to my research.

2. What country is most distributed in university rankings?

This would be interesting to see since if can help make a judgment on the education system. This was not to see which country has a better education because each country has top universities, but more to see which country gives students an opportunity at education under any circumstance. With the vast growing demand in education there is more demand in education and opens up doors. With the help of two hypotheses it can help come up with some conclusion or result based on the dataset that is being used.

DATASET

For the analysis on World University Rankings, the dataset was attained from a site called Kaggle. Kaggle is an online community of data scientists and machine learners. Kaggle allows users to find and publish data sets, explore and build in a web-based data science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. This is a resourceful site since it did collaborate with three well-known global university ranking data collectors. That includes Times Higher Education of World Universities, Academic Ranking of World Universities and Center of World University Rankings (Refer to the Literature section for more information). There is also supplementary data that was collected and inputted into the dataset from The World Data Bank and National Center for Education Statistics to create a dataset this extensive. There is so much information that is helping in ranking the Universities at a national and world level.

This dataset has 14 columns and over 2000 rows, covering a wide time range from 2012-2015. Each column in the dataset is specified to a specific category and is determined with the collection of data from each year. All the categories that are implemented in the dataset include: quality of education, influence, alumni employment, quality of faculty, patents, publications, broad impact, citations (population of students at the university). The rows are based on all university rankings from the years of 2012 – 2015. I went through to see if there were any null values and I was able to find one column broad impact to have null information for years 2012 and 2012. Broad impact was a new column in the dataset that was added later on and there was no previous data that was collected, thus no values in the table. For just looking at the table this was the information I was able to understand and interpret.

CLEANING/UNDERSTANDING DATA

The dataset from Kaggle was already pretty clean and did not need much cleaning. There was not really an unnecessary columns because I took into account most categories in order to make final conclusions and implications.

There is one exception for the dataset that I did come across. The dataset had some missing data from 2012-2013 for broad impact. I am assuming that the data was only started to collect from the years of 2014-2015. This made the analyzing a bit difficult since it was not as reliable as the other categories that covered the entire time period. However, I still did take into consideration when I was making my final conclusions.

Adding on, the coding aspect I did year by year for a comparing and contrast result. This was the most efficient way since it made sense when viewing the visuals. When I attempted to graph everything the visuals where mess and chaotic to interpret. There was no straight conclusions that could be made and the visuals would not be understandable to a random person without any knowledge about this project.
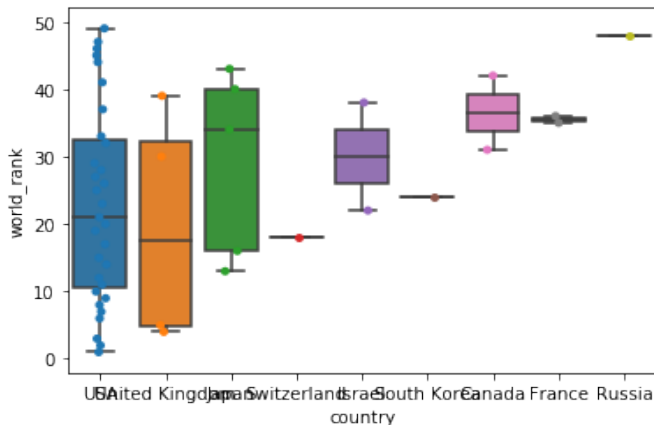
GRAPHS

I did most of the coding in Python to retrieve my graphs and the results. Python is an interpreted, high-level, general-purpose programming language. philosophy emphasizes code readability with its notable use of significant whitespace. I used all the topics that were used in class in order to create and assess the graphs. The topics included basic cleaning, making graphs, getting linear or logistic regression. For the dataset that I was using it made more sense to use linear regression. Linear regression is an approach that is used to model the relationship between a scalar response. Since I was looking at independent variables it was most sensible to use linear regression. I was able to create two box plots to describe a variety of distributions of

World University Rankings

universities to countries and 3 linear regression graphs to explain the most popular measures for ranking. (Each graph contains a description)

The first two visuals are based on distributions of countries and the top 50 universities. This will help me conclude my second hypothesis.

1. What country is most distributed in university rankings?

This box plot is in terms of top eight countries and distribution of top 50 universities. This graph is from the latest year (2015) in the dataset, since it was the most recent information. From this distribution, United States is the most distributed country with university rankings in relationship to the other countries. For example, we can see that Russia is fairly low in distribution and ranking.
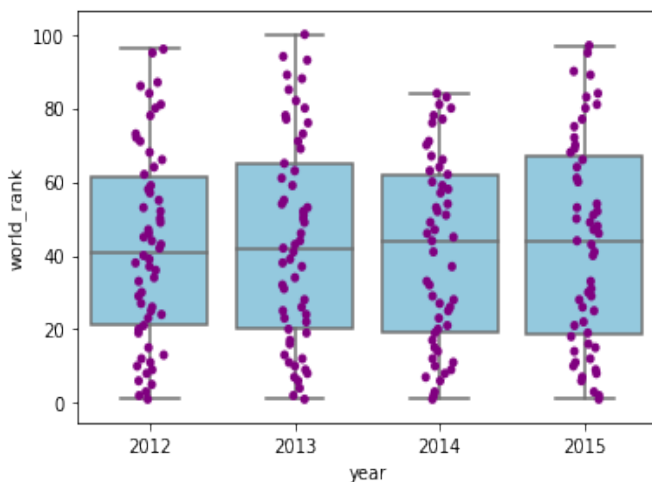
This box plot is in terms of United States distribution of the top 100 universities from 2012 – 2015. The distribution for the most part is consistent for the all four years and median was round 40 - 50, except for the year 2014. The year 2014 was more condensed in distribution but one can conclude the distribution median was still around 40 to 50 for rankings around the world. Another observation that can be made from the graph is that the median increasingly got worse throughout the years. This visual was interesting to see since it was based on United States Rankings.

For the next three visuals they are based on the relation between a measure and world ranking. This was to help determine the my first hypothesis that I made. I found the top three measures from the data to see the most impactful in ranking data.
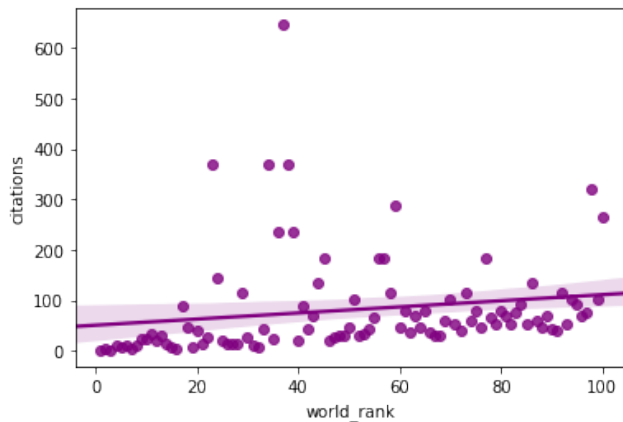
1. What measures lead to higher ranking of the universities?
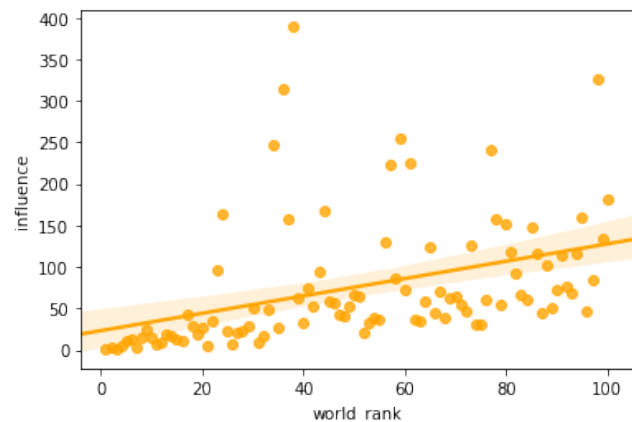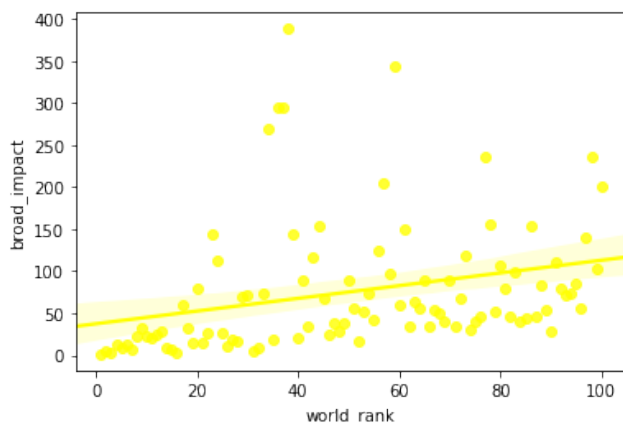

This graph is from the latest year (2015) in the dataset, since it was the most recent information. From this plot, the relationship is between world rank and the measure influence. We are able to see that the points are fairly in a straight line, thus I was able to do a linear regression for the values of influence. Compared to all the other measure influence happened to be the most impactful when it came to ranking a university. The linear regression was in bound with the points with a

few outliers in the graph. The outliers could be caused from misinformation collected.

Again just like the last graph this graph is from the latest year (2015) in the dataset, since it was the most recent information. From this plot, the relationship is between world rank and the measure citations. For this graph the points seemed to be in a fairly straight line so, I was able to make a linear regression. Citations was the second most impactful in ranking universities after influence. From the graph you are able to see a few outliers in the graph. The outliers could be caused from misinformation collected which we can see in the data.

Lastly, this visual is from the latest year (2015) in the dataset and was the third most impactful in the ranking universities. Since for this visual the lines seem to follow a similar slope pattern, it

made it easier to make a linear regression plot. From the graph you are able to see a few outliers in the graph. The outliers could be caused from misinformation collected which we can see in the data.

These were the main graphs that I found most important for my project and predictions. I did plot many more but did not find them important to include. The results of the graphs are discussed in the results and discussion section of this paper.

RESULTS

Like in the previous section with the graph each relationship was addressed. Besides just looking at the visuals it's also important to understand the results. For the analyzations are of the measures that were calculated to be the most impactful we can see the table down below. In the table each measure is displaying the slope, intercept, R-value, p-value, and standard deviation.

	Influence	Broad Impact	Citations
Slope	1.0445	.75501	0.6031
Intercept	23.3624	37.3818	50.7987
R-value	0.9999	0.999	1.0
P-value	0	0	0
Standard Deviation	3.1444	2.5412	0

To start off with, influence was the most impactful for the ranking, so it's crucial to understand the important values that were used to determine the results. Looking at the slope for out of all the measures it was the closet to one. This made it the best option for ranking universities. It's understandable that the other two measures were the second and third most closest to have a slope of one. Adding on, since the p-value is

supposed to be between zero and one and the smaller the p-values is the stronger the evidence and in this case all three cases were zero. Continuing on the with the results from the linear regression performed for the three measures, the R-value should also be considered when concluding the results. The R-value us always between positive one and negative one, which is used to describe the strength and direction of the linear relationship. In this case influence, broad impact and citations are in the positive direction and either really close to one or equal to one. Based on the linear regression, the conclusion can be made that these top three methodologies were heavily considered in ranking the world universities. In order to make these conclusions I did conduct the linear regression for all the different measures. After analyzing the data from each measure these with the results I concluded.

Additionally, Like compared earlier in the graph section , it was evident that United Stated had the best distribution for university rankings (referring to the boxplot visual). Looking at the results in terms of opportunity, I considered United States to be a country with more opportunity. One being that the most intellect students can attend the top ranked universities, but also there is a place for other student to improve their education and have an opportunity. Second, United States is more considerate of giving education to everyone where not all countries this like that. Like let's consider a country like India where girls are not given the opportunity (considering the fact that the country is not in the visual). This significant data help conclude my second hypothesis.

Lastly, I did find some concerns to address when making the final conclusion. Number one being that broad impact was only looking at two years' worth of data rather than the fours like the other measures. It's imperative that we know that because there is discrepancy in the data since all the statics aren't fairly contributed. For this is possible that since broad impact is relatively a new column added the dataset, it must have been critical to add since it must have been important.

Since the data was essential to consider the data must have been relatively higher in value then the other columns, causing the broad impact to come in the top three.

Overall, after conducting the linear regression for the different methodologies these were the top results and United States give education opportunity all students compared to other countries. All the results make sense logically based on the values.

DISCUSSION

Since the results have been concluded, it is also critical to think of how to make future predictions better and what other predications can be made. The analyze of this dataset was unitized to the extent that is could have. As we are aware that education and the system that is use is constantly changing and getting better, it would be interesting to take other factors into consideration. One may be aware that recently many colleges and universities are opting for going test optional. Many colleges are considering that making ACT and SAT optional for application it could increase minority admissions for education. This has been reported gains in minority applications and enrollments -- without a decrease in graduation rates (Inside). In a few years from now it will be interesting to take this factor into account and see how the results very in world rankings for universities.

CONCLUSION

In conclusion, with the World University Ranking data set it was easy to understand cleaning and analyzing a dataset. World University Ranking is a popular concept that may organizations take part in to determine which methodologies make a difference. As a student we should be aware of the standing of universities to ensure that we can get the best education that we deserve for the university that best fits our career path.

REFERENCES

[1] "About Academic Ranking of World Universities: About ARWU." *About Academic Ranking of World Universities | About ARWU*, http://www.shanghairanking.com/aboutarwu.html.

[2] Bothwell, Ellie, et al. "World University Rankings." *Times Higher Education (THE)*, 19 Nov. 2019, https://www.timeshighereducation.com/content/world-university-rankings#survey-answer.

[3] "College and University Rankings." *Wikipedia*, Wikimedia Foundation, 27 Nov. 2019, https://en.wikipedia.org/wiki/College_and_university_rankings.

[4] Doumbia, Kafoumba. "Why Is Education So Important in Our Life? - Kafoumba Doumbia - EdLab." *EdLab, Teachers College Columbia University*, EdLab, 19 Aug. 2013, https://edlab.tc.columbia.edu/blog/9886-Why-is-Education-So-Important-in-Our-Life.

[5] Inside Higher Ed. (n.d.). Retrieved from https://www.insidehighered.com/admissions/article/2019/04/01/more-colleges-go-test-optional-admissions.

[6] O'Neill, Myles. "World University Rankings." *Kaggle*, 13 Nov. 2019, https://www.kaggle.com/mylesoneill/world-university-rankings.

[7] Doumbia, Kafoumba. "Why Is Education So Important in Our Life? - Kafoumba Doumbia - EdLab." *EdLab, Teachers College Columbia University*, EdLab, 19 Aug. 2013, https://edlab.tc.columbia.edu/blog/9886-Why-is-Education-So-Important-in-Our-Life.

[8] "World University Rankings 2019-2020." *CWUR*, https://cwur.org/.