IMT ML | final Data Science project

AmirHosein Mousavian #8

Data set:

This data set was about 1,000 top universities globally. It was a (1000 , 12) dataset. This dataset divides columns by these factors:

- **World Rank**
- **Institution**
- **Location**
- **National Rank**
- **Quality of Education**
- **Alumni Employment**
- **Quality of Faculty**
- **Research Output**
- **Quality Publications**
- **Influence**
- **Citations**
- **Score**

At first it seems like we have no missing values but as we see in CSV file and Notepad++, this data is filled with "–" and "> 1000" values. Because these values aren't numeric, they will be a problem if we don't handle them. We always handle the Noise values first but because this is a ranking dataset, although we checked, we can't do anything because those universities could actually scored those points. So we don't go for Noise values here but we will handle Missing values.

Two columns (Quality of Education & Quality of Faculty) have (597 & 731) "-" values. We replace them with Nan values and handle them later. And now we have 1,328 missing values.

5 columns (Alumni Employment, Research Output, Quality Publications, Influence, Citations) have (509, 77, 51, 171, 110) "> 1000" values. Because the top 1000 universities are ranked here, we just replace them with "1001" value.

*Joupyter file is attached, you can red the code and interact with plots there.

When we start to do data science, we realize that some column names are not the way there shown. So we ‘repr’ to call the names and see that 2 columns have wrong characters, so first we had to fix those.

Now we check for duplicated values and we found nothing.

From earlier we know that column types are mostly ‘object’ but we want numeric type instead so we convert them to ‘float’.

Missing values

Earlier we talked about missing values (for comfort we call them MV from now on). We can’t just delete those indexes from data because we just have 1000 rows and we can’t lose 1,328 of them 😊 so we have to fill them with a value. Because this data is all about ranking, we can’t just fill them with median or mean value of columns.

As we spoke with the dataset owner, we come to this conclusion to fill these values with (max+1) of each column.

After we fill those, we check for missing values in the new DataFrame and we see that now we don’t have MVs. Now we saved the data and cleaned it.

Analyzing & ploting

*One thing to remember, here if you score less, you win! 😊

What we get from the top 10?

1. 8 if the first 10 universities are based in USA others are based in UK

What we get from the bottom 10?

1. 2 of the least scored universities are based in USA.
2. 7 of the least scored universities are based in Asia.
3. 1 of the least scored universities is based in Africa.

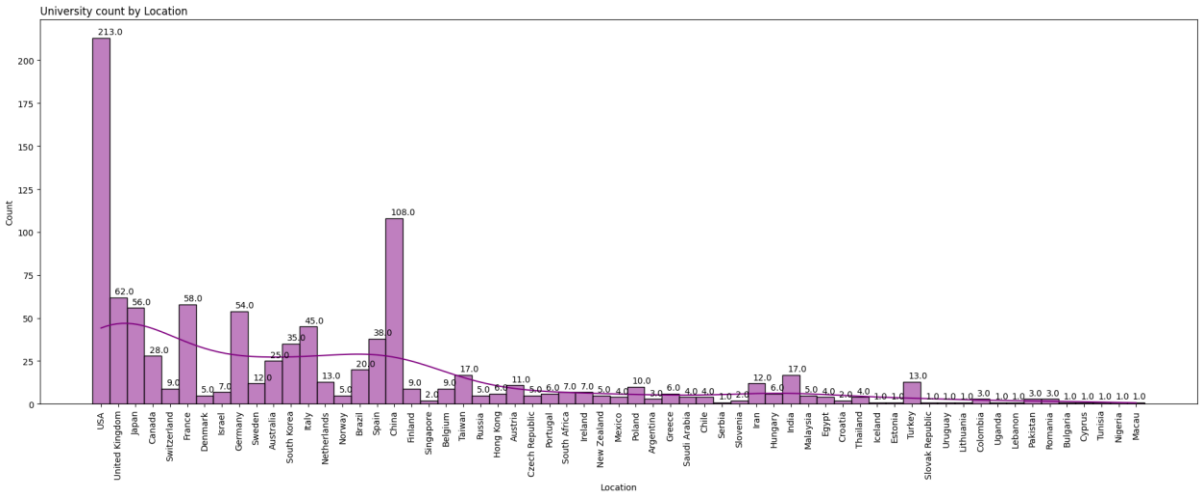

Fig 1.1

1. USA with 231 is the top country in the list.
2. second place is china with 108.
3. 3rd country is UK with 62 universities.
4. least top countries have just one university in the list. like Macau, Nigeria, Tunisia and more.
5. in this data countries have the mean of 16.39 universities.

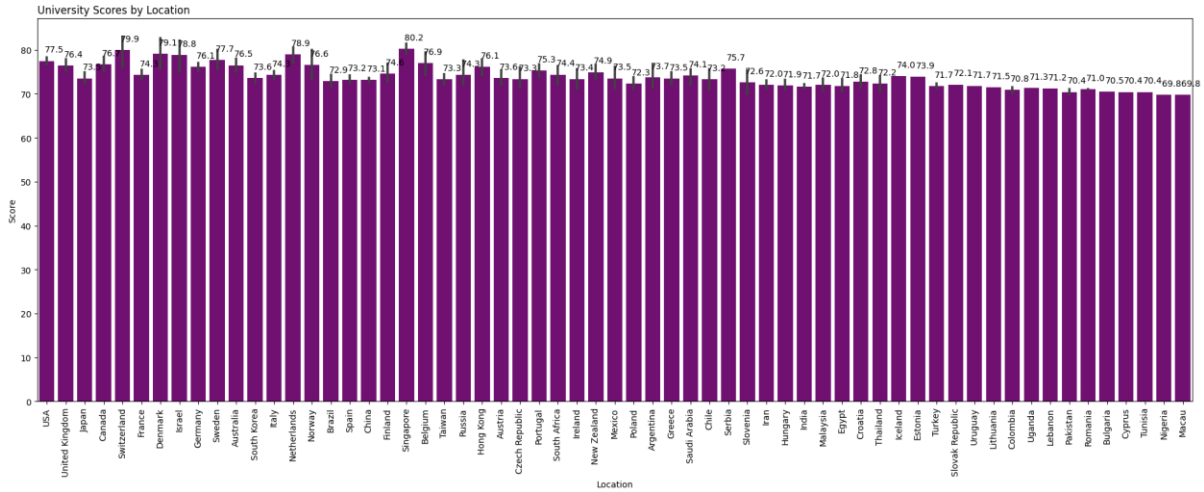

Fig 1.2

1. Best scored country is Singapore with 80.2 .
2. Second country is Switzerland with 79.9 .
3. 3rd country is Netherlands with score of 78.9 .
4. least scored country is Macau with 69.8 points.
5. mean of score is 75.03 .

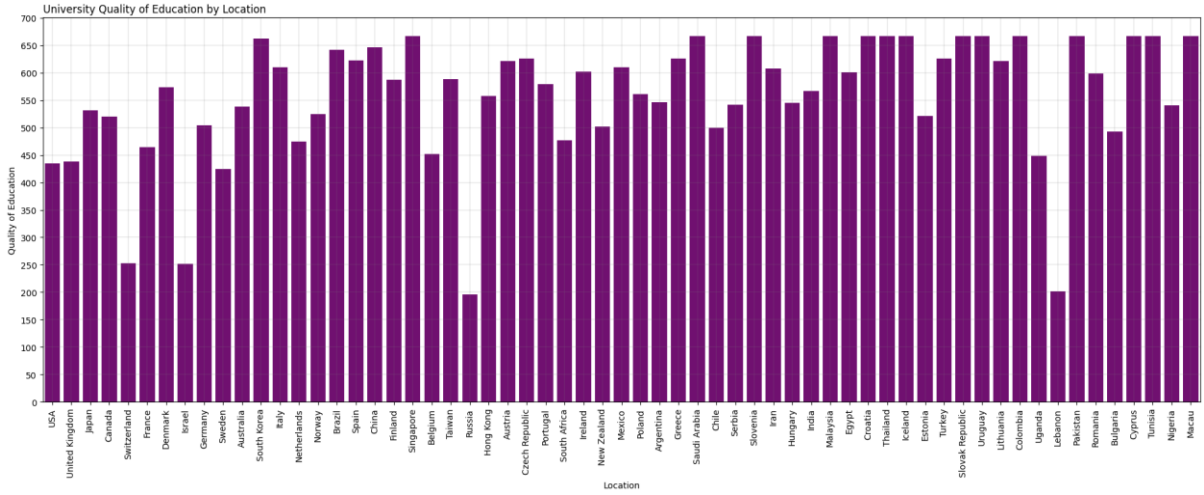

Fig 1.3

1. Singapore, South Korea, Saudi Arabia, Slovenia, Malaysia, Croatia, Thailand, Iceland, Slovak Republic, Lithuania, Pakistan, Cyprus, Tunisia and Macau have the best Quality of Education.
2. Russia and Lebanon have the least score in the Quality of Education.
3. Mean of Quality of Education score is 323.4 .

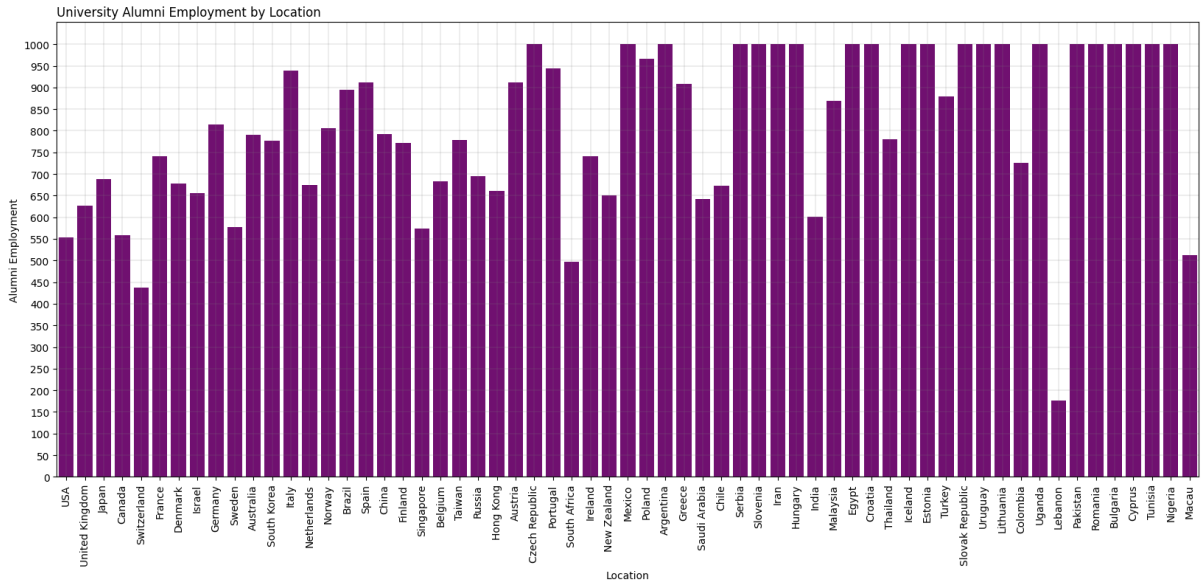

Fig 1.4

1. Czech Republic, Mexico, Argantina, Serbia, Slovenia, Iran, Hungry,... scored more in the Alumni Employment. with 1000
2. Top Scored countries like USA, UK ,... are not top.
3. Least scored country is Lebanon with 176.0 and next to it are Macau and South Africa with 512.0 .
4. Average scorevin this area is 726.52 .

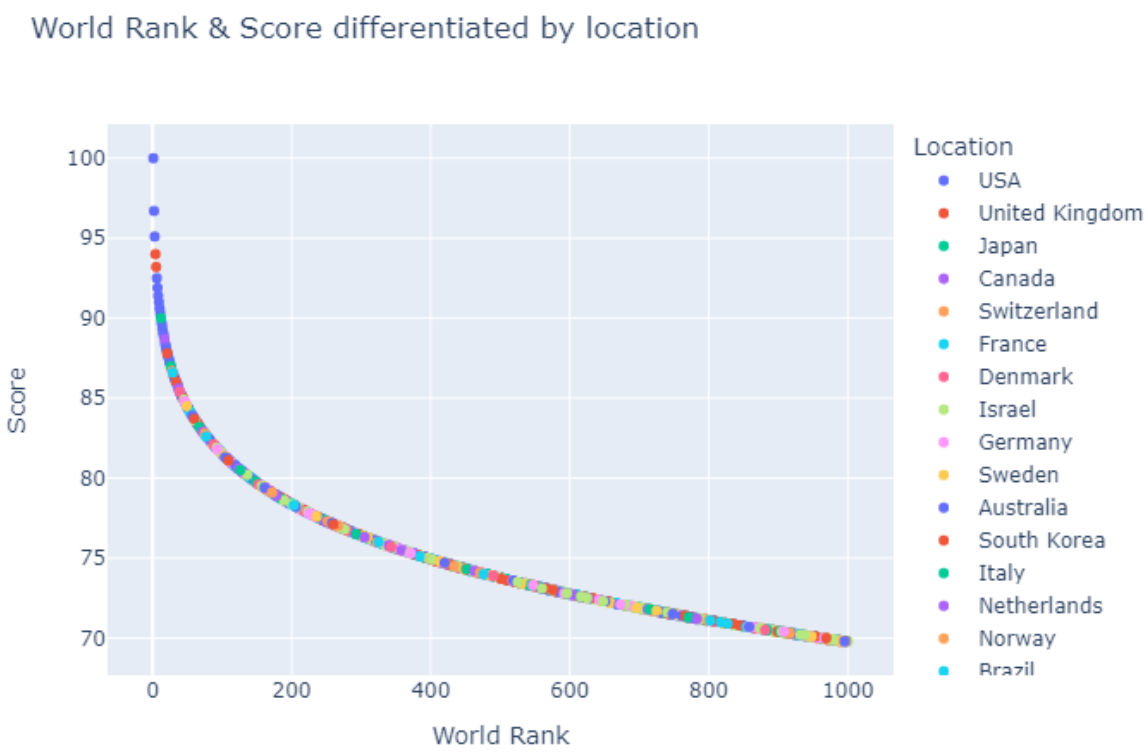

Fig 2.1

1. when Score reduces, World Rank reduces too.
2. Highest Score is 100 and the Lowest is 69.8.
3. The difference between Rand 1 and 2 is more than other ranks.
4. Difference between first 200 (around 28.6) is more than rest of the values(around 8.6).

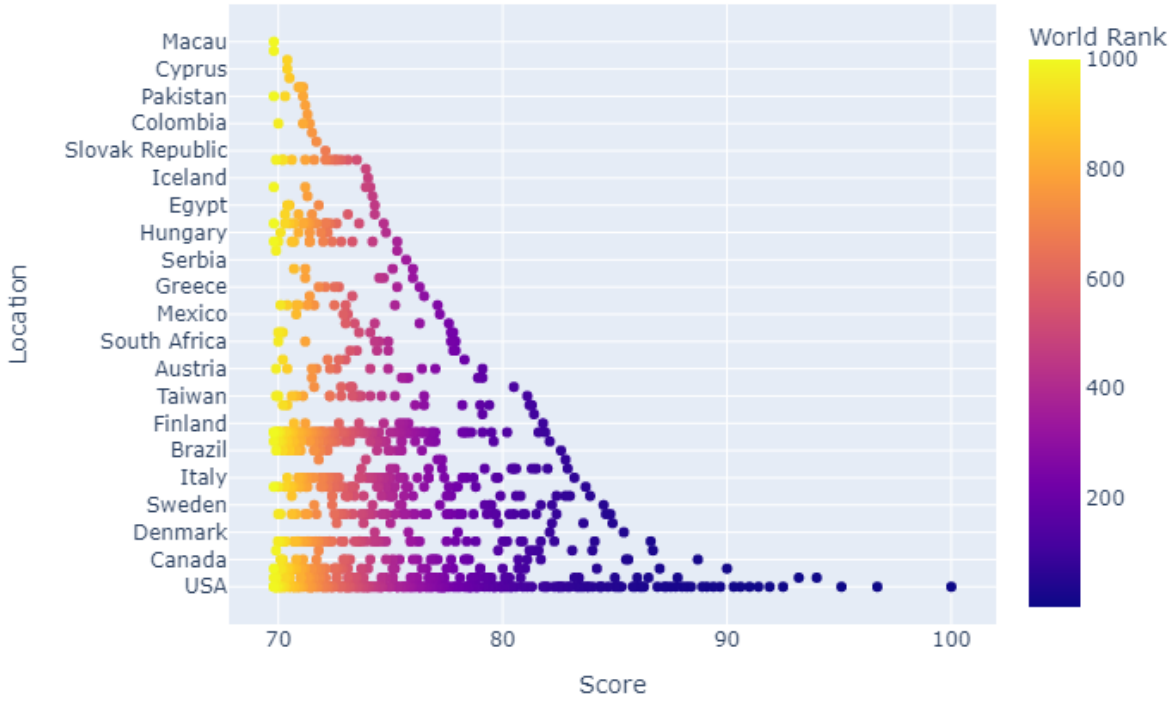

Fig 2.2

1. USA with 231 universities has the most universities in the data.
2. USA has the most verity in the data scores.
3. North America and Western Europe have the most top university in data.
4. Best university is Harvard with world rank of 1 and the Worst one is Capital Normal University with rank of 1000.

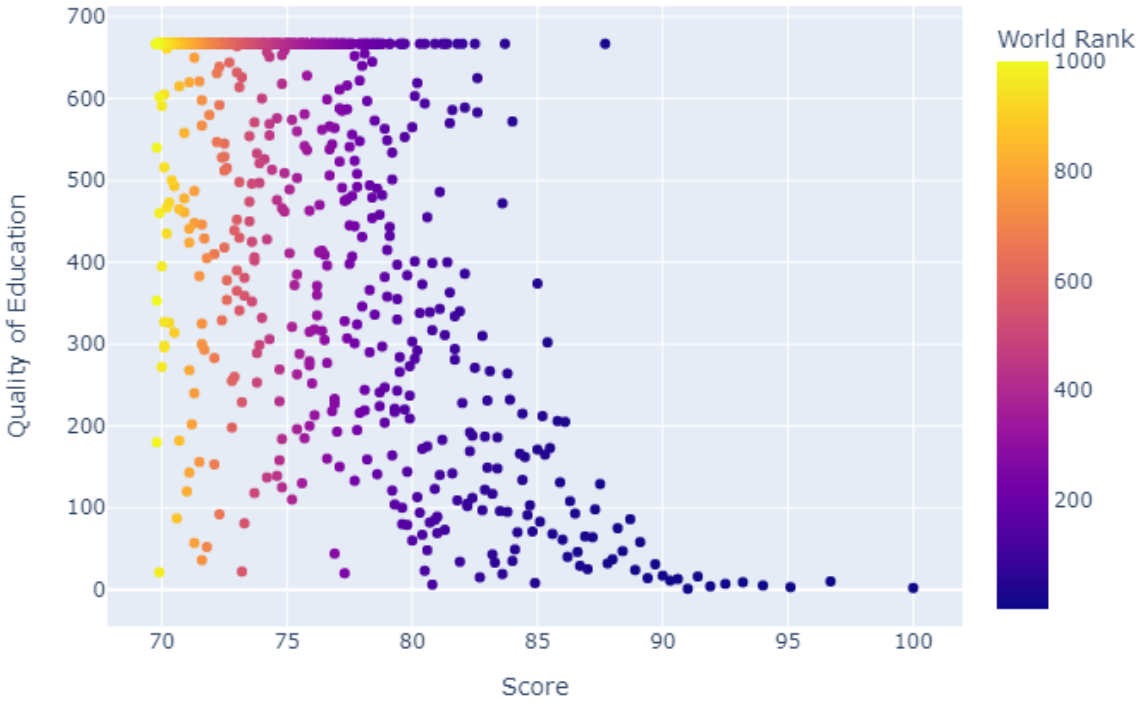

Fig 2.3

1. USA has the most verity in data
2. Most of the top universities have a better Quality of Education.
3. values with Quality of Education of 667 are the filled missin values.
4. worst Quality of Education university in top 100 universities is Texas A&M University, College Station with score of 625.0 .
5. Best university is Harvard with score of 1 and the worst one is University of Zaragoza with score of 666.

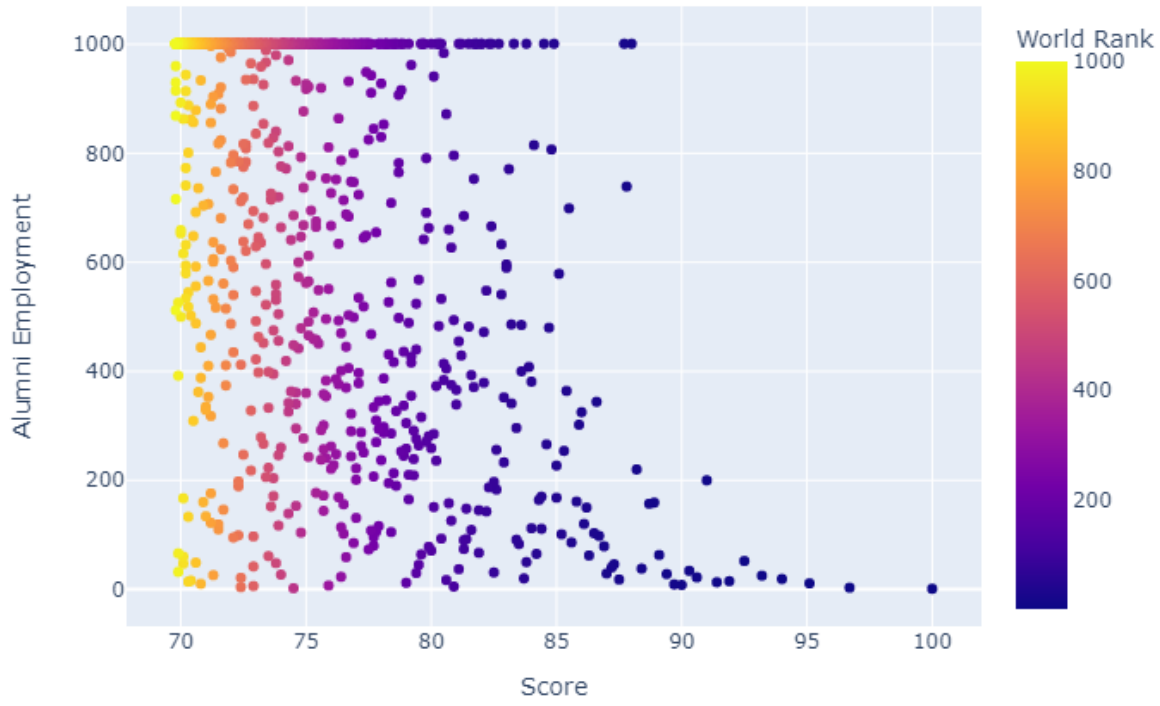

Fig 2.4

1. USA has the most verity.
2. Some of the universities with less score have a better Alumni Employment.
3. Top universities have a great Alumni Employment.
4. universities with value of 1001 in Alumni Employment are the ones that scored < 1000 and we replaced them with 1001.
5. Best university in this plot is Harvard with score of 1 and the worst one is University of Kaiserslautern with score of 997.

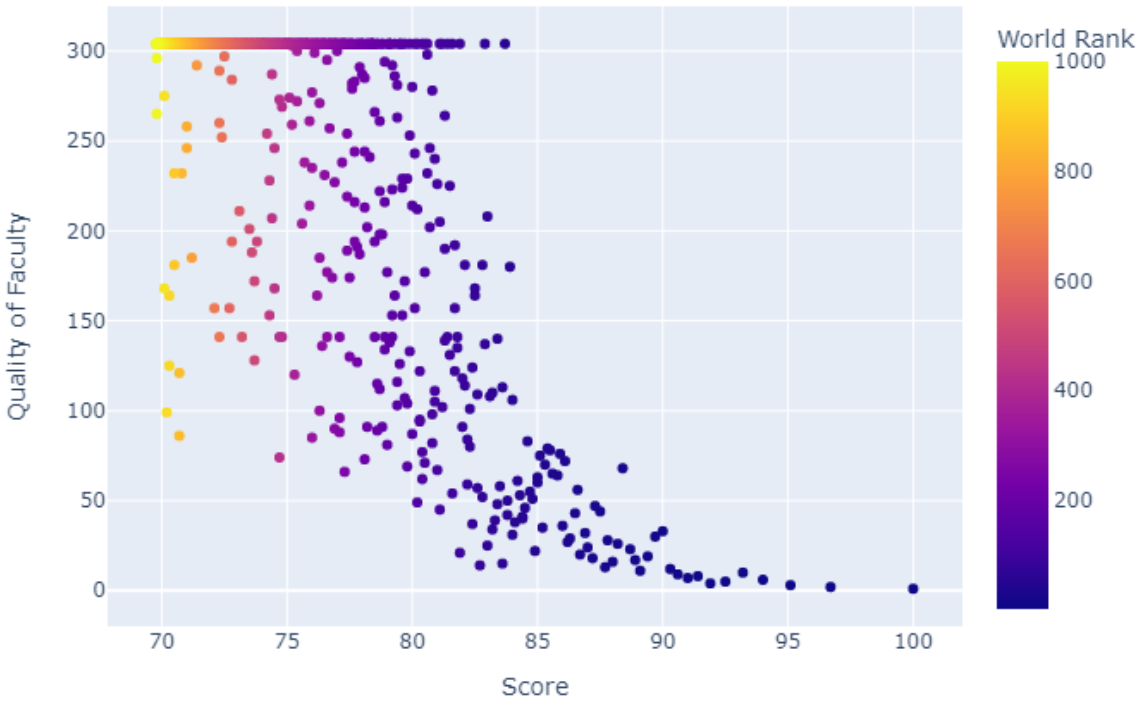

Fig 2.5

1. Universities with a better Faculty have a better score.
2. Universities with a better Faculty have a world rank.
3. Universities with Quality of Faculty of 304 are the ones we filled.
4. Best faculty belongs to university of Harvard and the worst with score of 303 is University of Marburg.

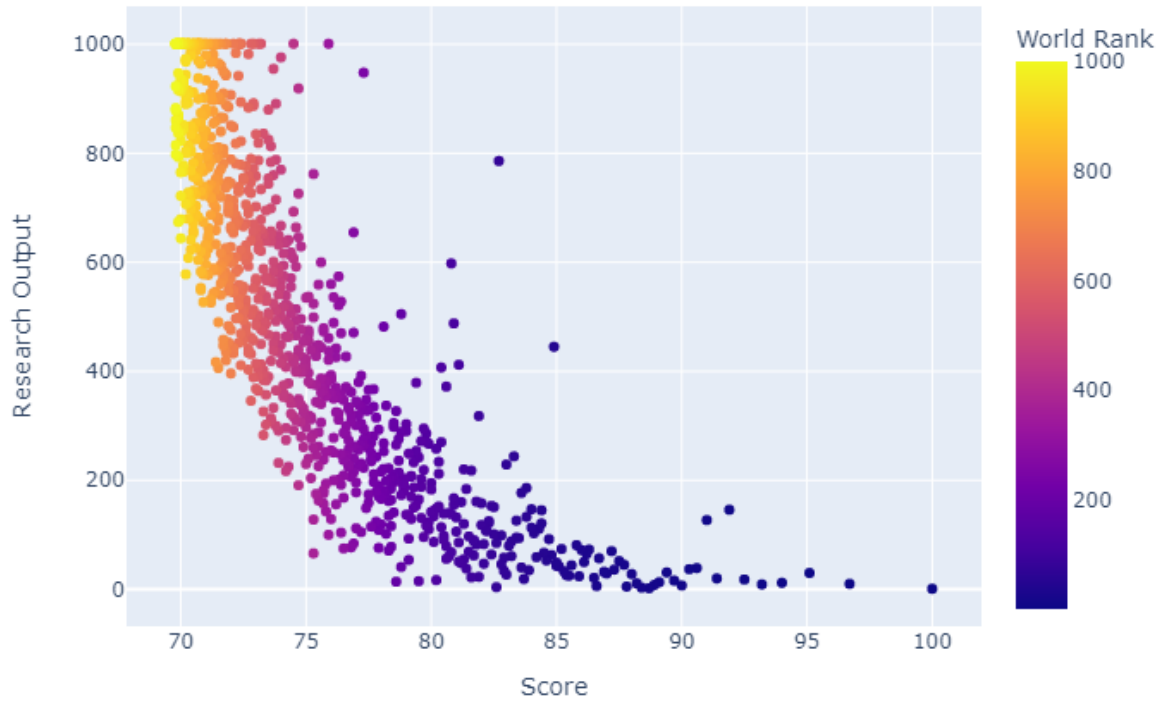

Fig 2.6

1. universities with a better score have a better research output.
2. universities with a better world rank have a better research output.
3. most of the universities are in the left side of the plot with score of less than 80 and research output of between 15 to 1000.
4. values of 1001 are the replaced ones.
5. best university is Harvard and the worst one is University of Jaén.

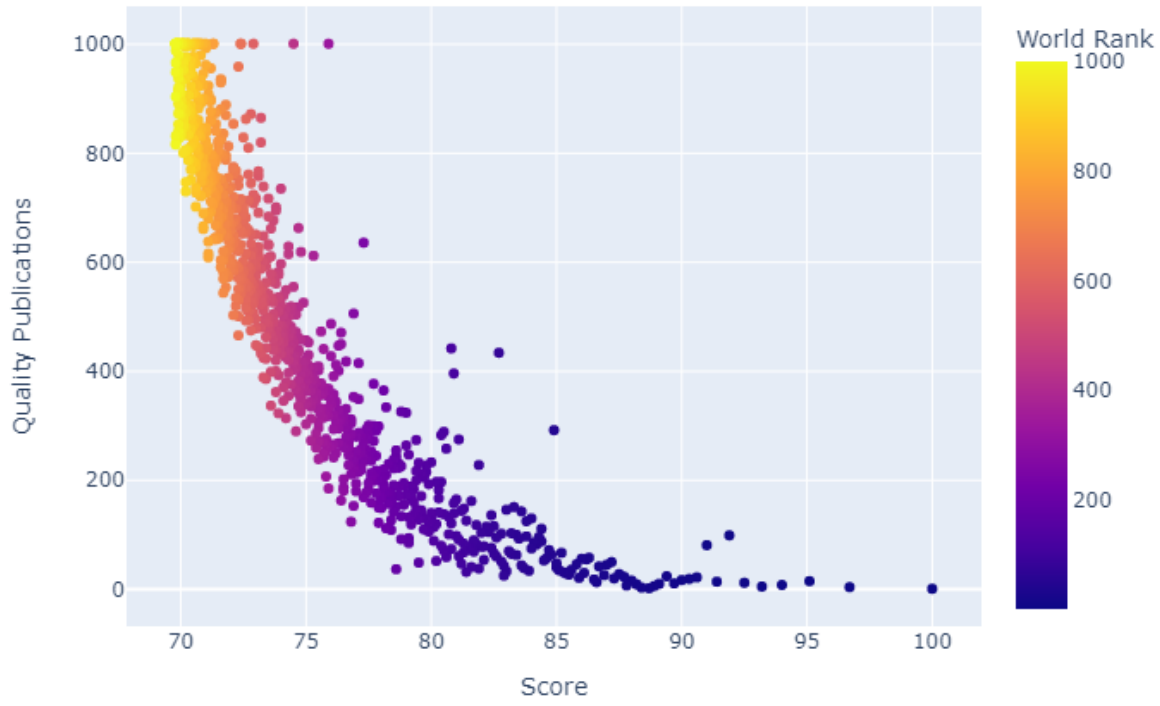

Fig 2.7

1. universities with a better score have a better Quality Publication.
2. universities with a better world rank have a better Quality Publication.
3. most of the universities are in the left side of the plot with score of less than 80 and research output of between 37 to 1000.
4. values of 1001 are the replaced ones.
5. best university is Harvard and the worst one is Federal University of Goiás.

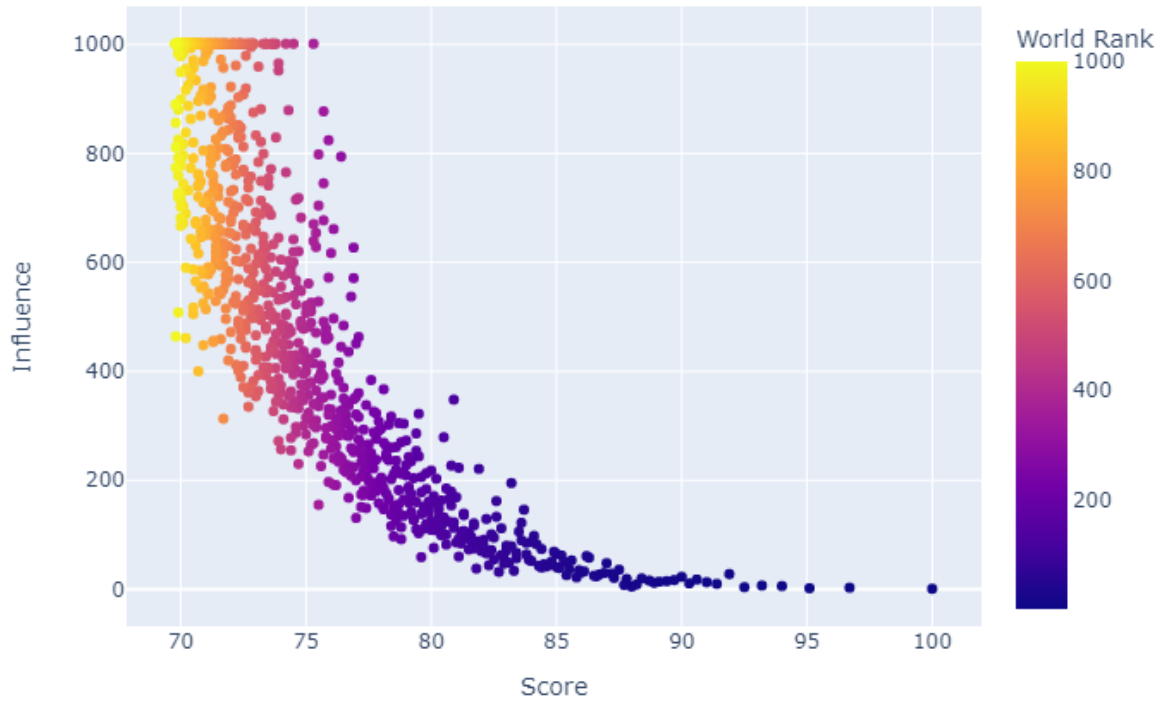

Fig 2.8

1. universities with a better score have a better Influence.
2. universities with a better world rank have a better Influence.
3. most of the universities are in the left side of the plot with score of less than 80 and research output of between 59 to 1000.
4. values of 1001 are the replaced ones.
5. best university is Harvard and the worst one is South China Agricultural University.

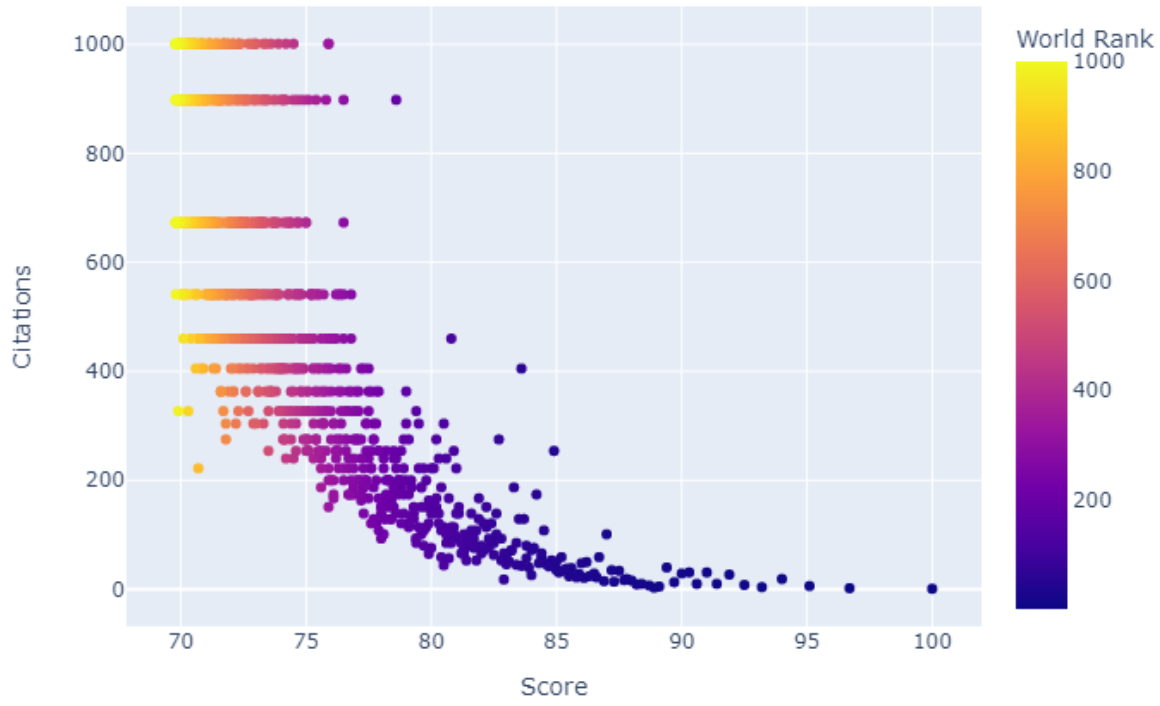

Fig 2.9

1. universities with a better score have a better Citations.
2. universities with a better world rank have a better Citations.
3. most of the universities are in the left side of the plot with score of less than 80 and research output of between 65 to 1000.
4. values of 1001 are the replaced ones.
5. best university is Harvard and the worst ones are with score of 898.

That's it for this data set, I hope you liked it. I'll appreciate any notes regarding this analysis.

E-mail: 4amirhm@gmail.com

Github: github.com/4amirhm

Kaggle: kaggle.com/amirhoseinmousavian