Data Science Project final

Mohammad Bagher Soroush

Project Master:

Mohammad Reza Momeni

Summer 2022

Introduction: the dataset that we have is about universities in the world. Our goal is to study the factors that

affecting on scores of universities and their world ranking. Here's a quick look of our dataset (the first five and the last five lines):

	World Rank	Institution	Location	National Rank	Quality of Education	Alumni Employment	Quality of Faculty	Research Output	Quality Publications	Influence	Citations	Score
0	1	Harvard University	USA	1	2	1	1	1	1	1	1	100.0
1	2	Stanford University	USA	2	10	3	2	10	4	3	2	96.7
2	3	Massachusetts Institute of Technology	USA	3	3	11	3	30	15	2	6	95.1
3	4	University of Cambridge	United Kingdom	1	5	19	6	12	8	6	19	94.0
4	5	University of Oxford	United Kingdom	2	9	25	10	9	5	7	4	93.2
995	996	Aga Khan University	Pakistan	3	-	> 1000	-	> 1000	> 1000	464	673	69.8
996	997	University of Calcutta	India	17	353	716	296	798	966	> 1000	> 1000	69.8
997	998	K?chi University	Japan	56	-	> 1000	-	> 1000	> 1000	811	673	69.8
998	999	Soonchunhyang University	South Korea	35	-	> 1000	-	881	> 1000	> 1000	898	69.8
999	1000	Capital Normal University	China	108	-	869	-	923	904	889	> 1000	69.8

As you can see, our dataset has 1000 rows and 12 columns. The columns include: world rank, institution, location (country), national rank, quality of education, alumni employment, quality of faculty, research output, quality of publications, influence, citations and score.

First of all, we check the null values before we analyze our plots:

```
World Rank              0
Institution             0
Location                0
National Rank           0
Quality of Education    0
Alumni Employment       0
Quality of Faculty      0
Research Output         0
Quality Publications    0
Influence               0
Citations               0
Score                   0
dtype: int64
```

As you can see the numbe of null values in all columns is zero, so there's no null values, but still we got a problem because of all the "-" in the "Quality of Education" and "Qulity of Faculty" column.

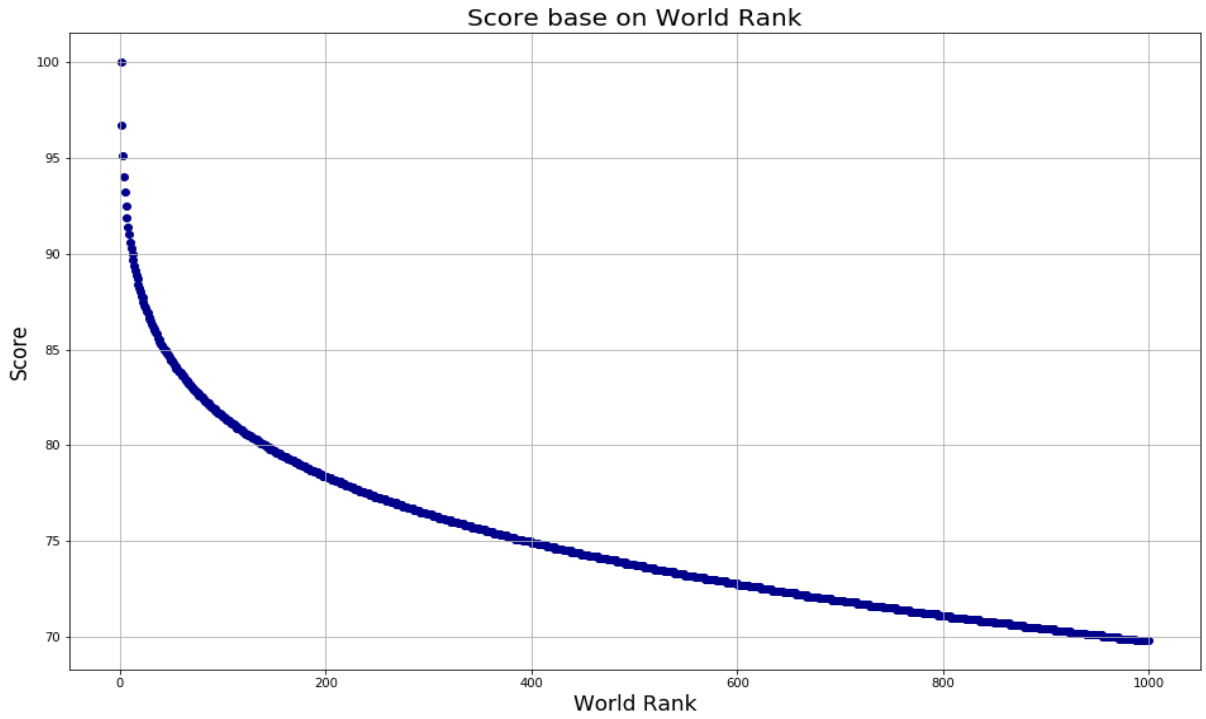
Because we have no information about what are they and we have to drop them or replace them with a logic value So we put them away and continue our analyze.

We also have to make sure that the values of our columns beside "Institution" and "Location" columns be of numerical type.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 10 columns):
World Rank              1000 non-null int64
Institution             1000 non-null object
Location                1000 non-null object
National Rank           1000 non-null int64
Alumni Employment       1000 non-null object
Research Output         1000 non-null object
Quality Publications    1000 non-null object
Influence               1000 non-null object
Citations               1000 non-null object
Score                   1000 non-null float64
dtypes: float64(1), int64(2), object(7)
memory usage: 78.2+ KB
```

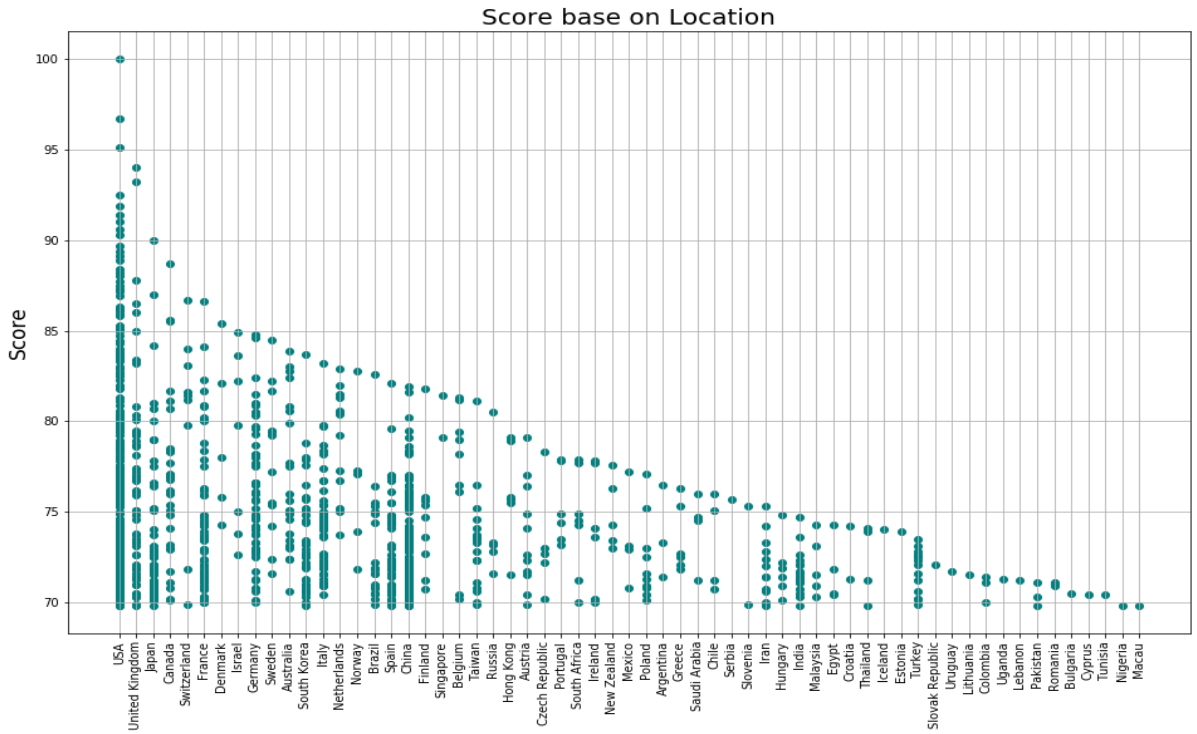
As it's obvious, we have to change the "object" types to "int64" from the desire columns.

At the end, we replace the "> 1000" values to "1001" and now we can get to the plots.



Score base on World Rank:

The plot behavior is exponential and as the world rank is decreasing, the score is extremely increases, so they have an inverse relationship.



Score base on Location:

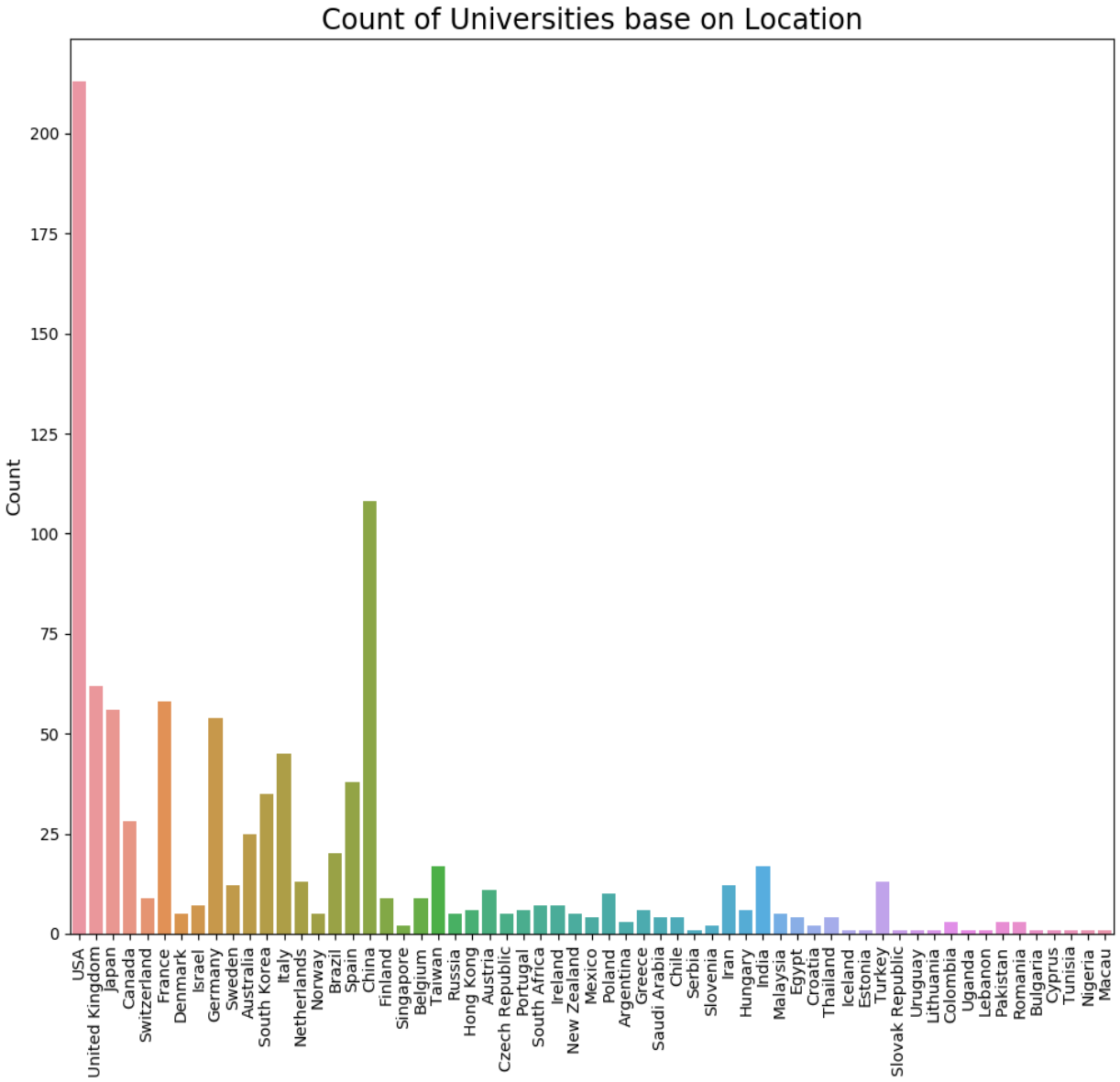
What attracts our attention at the first glance, is the density of samples in "USA" location from the lowest score to the highest which is approve the fact that America not only has the best universities, but also have a lot of good and average of them. "USA" is number one in having the highest score and the number of universities and "Macao" is the last one regarding to this dataset.

But having more universities than other countries doesn't mean being the best.

For example, "Switzerland" is number 5, but France clearly has more density of sample especially between 70 and 75

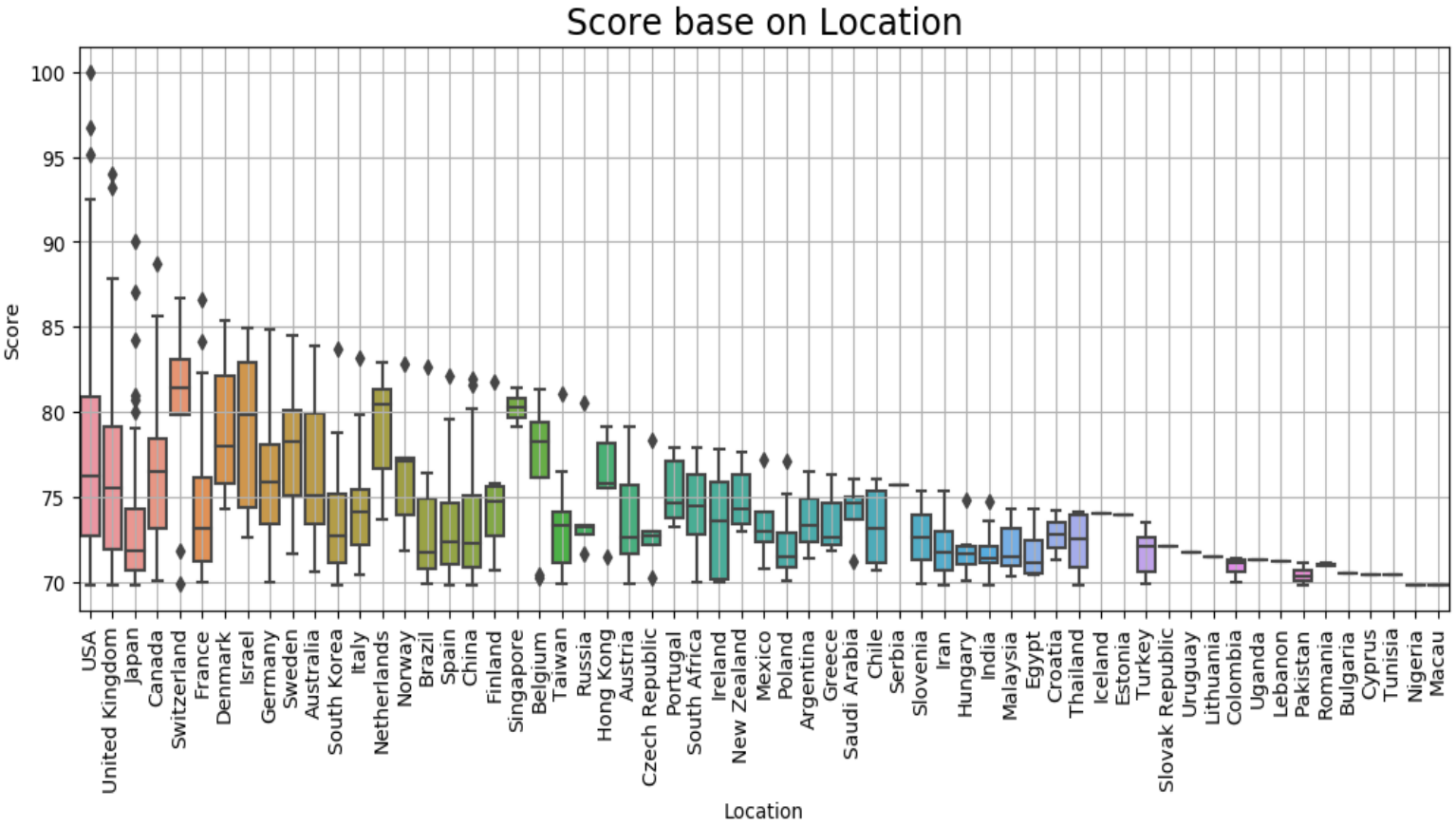
range of score which mean having the universities with high scores is more effective than having plenty of them with an average or even good scores.

Let's take a look to other type of plots so we could go deeper:



China has 125 universities in the world rank (the maximum amount after USA) and 17th country in the world who have the best universities base on score.

But we can see there are 8 countries that the number of their universities is less than 25 in the world ranking but still they have a better rank which approves the fact that we said before.



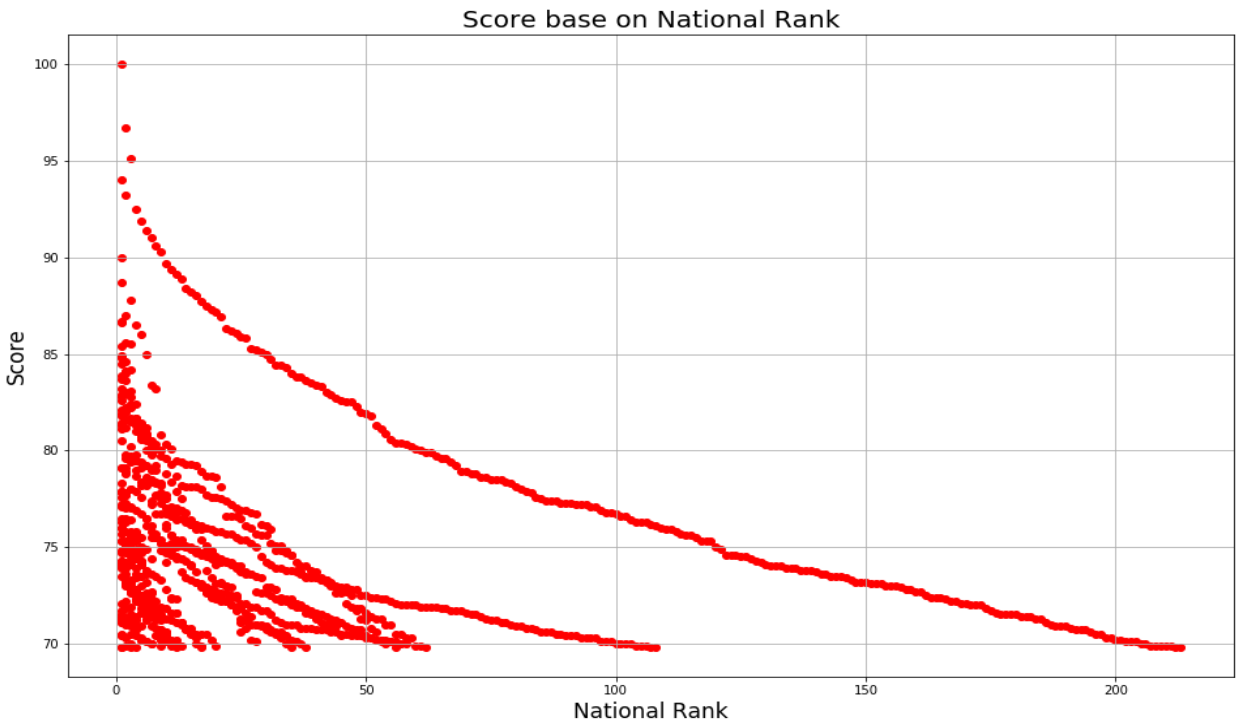
Regarding to this boxplot, although "USA" is number 1 at having the most and the best universities in the world, but it's not best at everything. The average score of the

universities in the world rank from USA is 77 and there are 9 countries that have a petter average.

Switzerland has the highest average of score which is 82 with less than 25 universities in the world rank.

But that doesn't mean as the number of universities is lower, the average is higher. We can see "United Kingdom" as an example.

United Kingdom has much less universities compare to America but still have a lower average value.

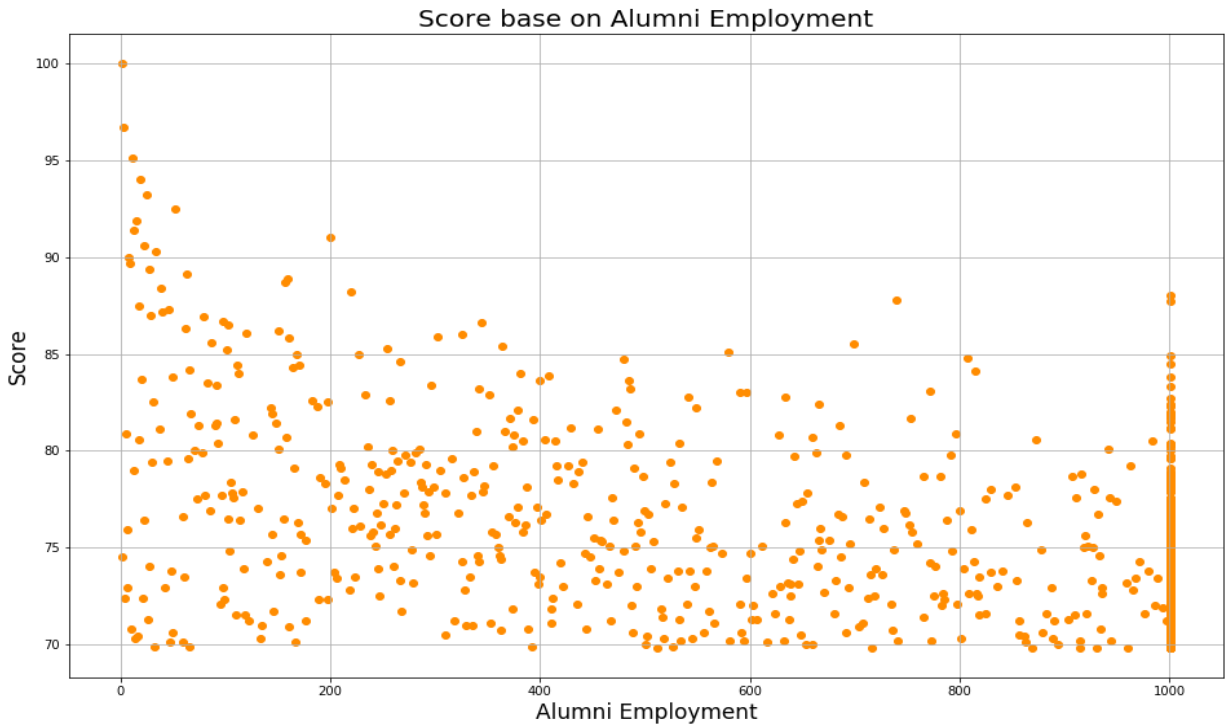


Score base on National Rank:

The plot behavior is also exponential here but the extreme of that is different for each country.

Some of them may not have a high score, but they have a very low value of national rank and that's the reason of very high density between 0 and 25 range of national rank.

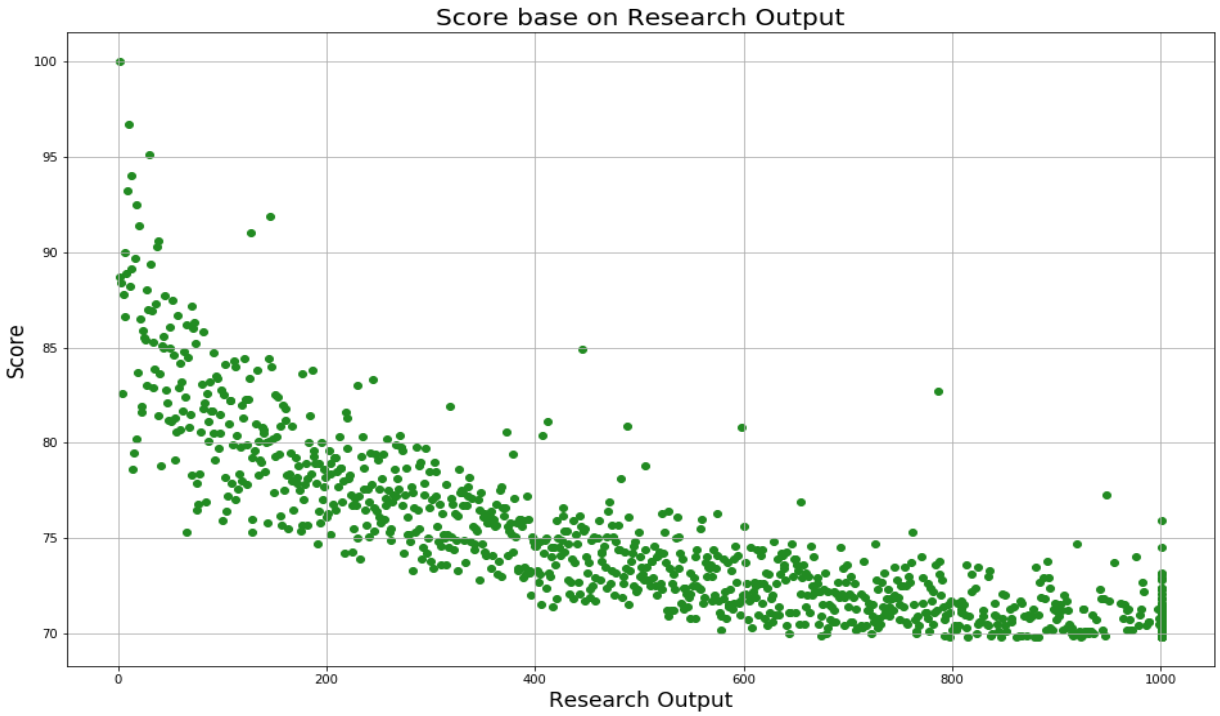
The highest chart in the plot is probably belong to "USA" because having more universities in general and in all scores from average to best, but still as long as we go forward in national rank, the density goes higher and that tells us not every university with high score is In the America, but there's a lot.

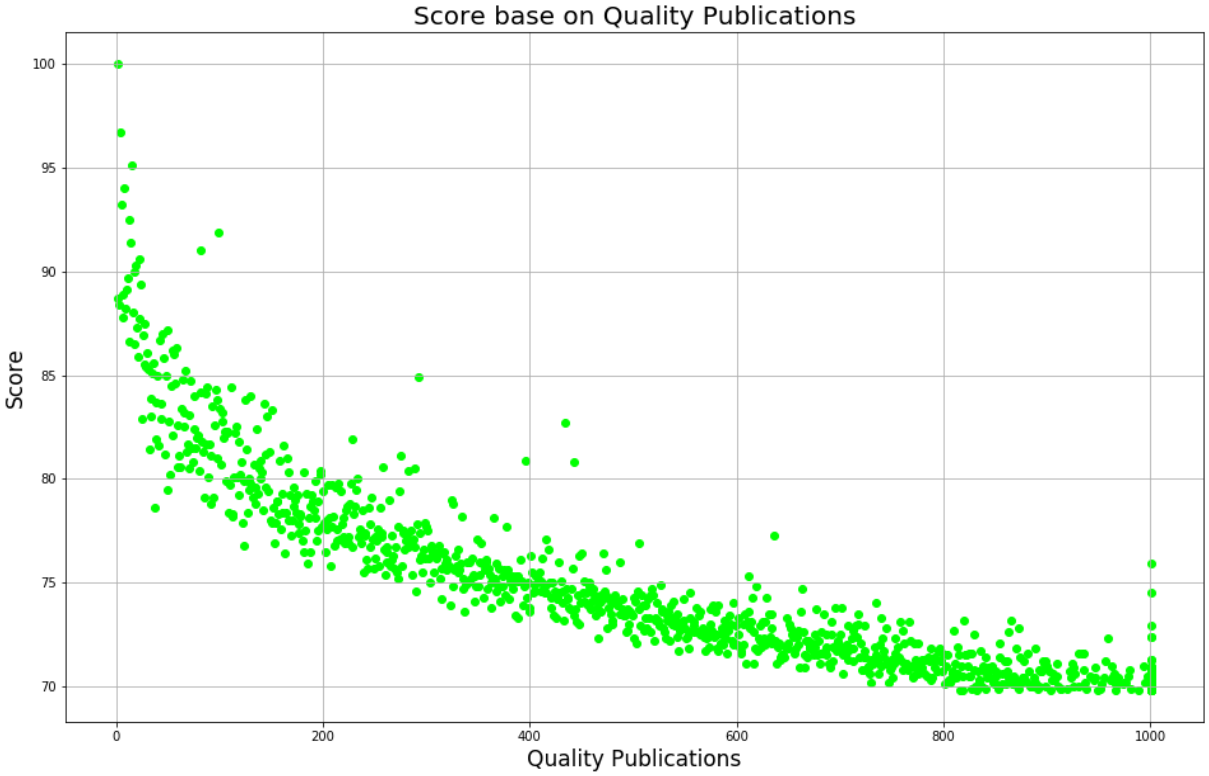


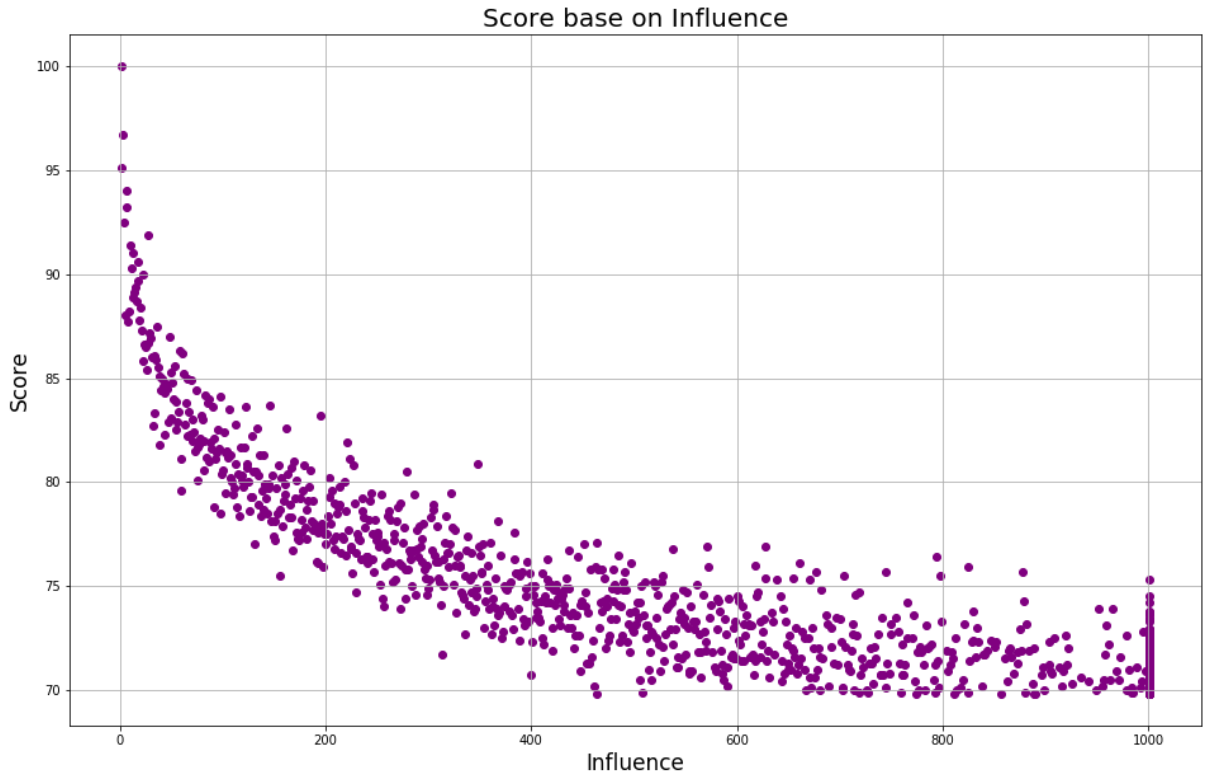
Score base on Alumni Employment:

The value of alumni employment is the same between 0 and 990 and we can say it's more between 70 and 85 range of score and the universities with a high score, doesn't have much of value of alumni employment.

But all of the sudden, there's very high density of samples in 1000 and more (because we replace the "> 1000" to "1001") of alumni employment that also between 70 and 85 but as the score decreasing, the alumni employment increases.


Score base on Research Output
Score
Research Output
100
95
90
85
80
75
70
0
200
400
600
800
1000

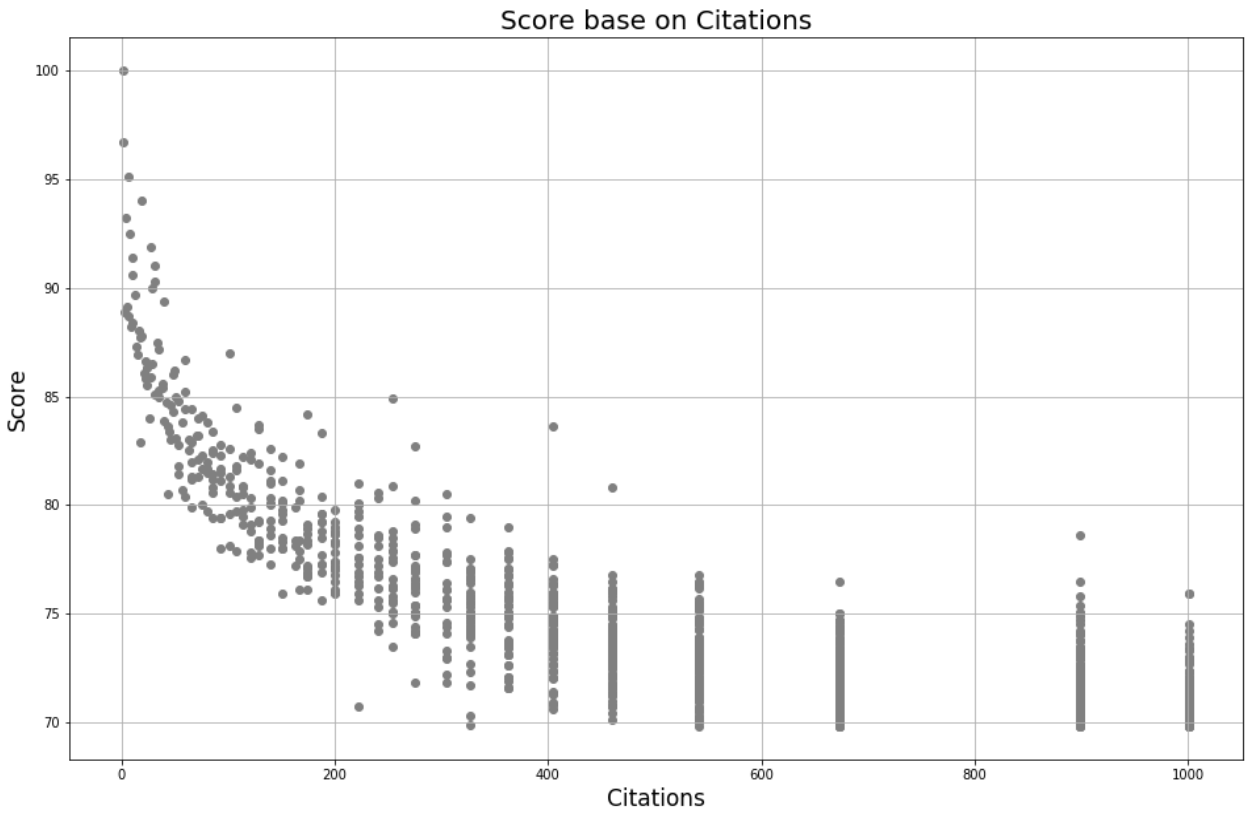

Score base on Quality Publications
Score
Quality Publications
100
95
90
85
80
75
70
0
200
400
600
800
1000



Score base on Research Output, Quality Publications and Influence:

We can say these plots have a same exponential behavior but second plot is denser than others. Here also the universities with high score have a low value of their features and the density is very high at 1000 value and more which is between 70 and 75 range of score for "Research Output" and "influence" plots.

But for the "Quality Publication" plot, the high density started from 250 and continues to 1000 in the range of 78 to 70 of score.



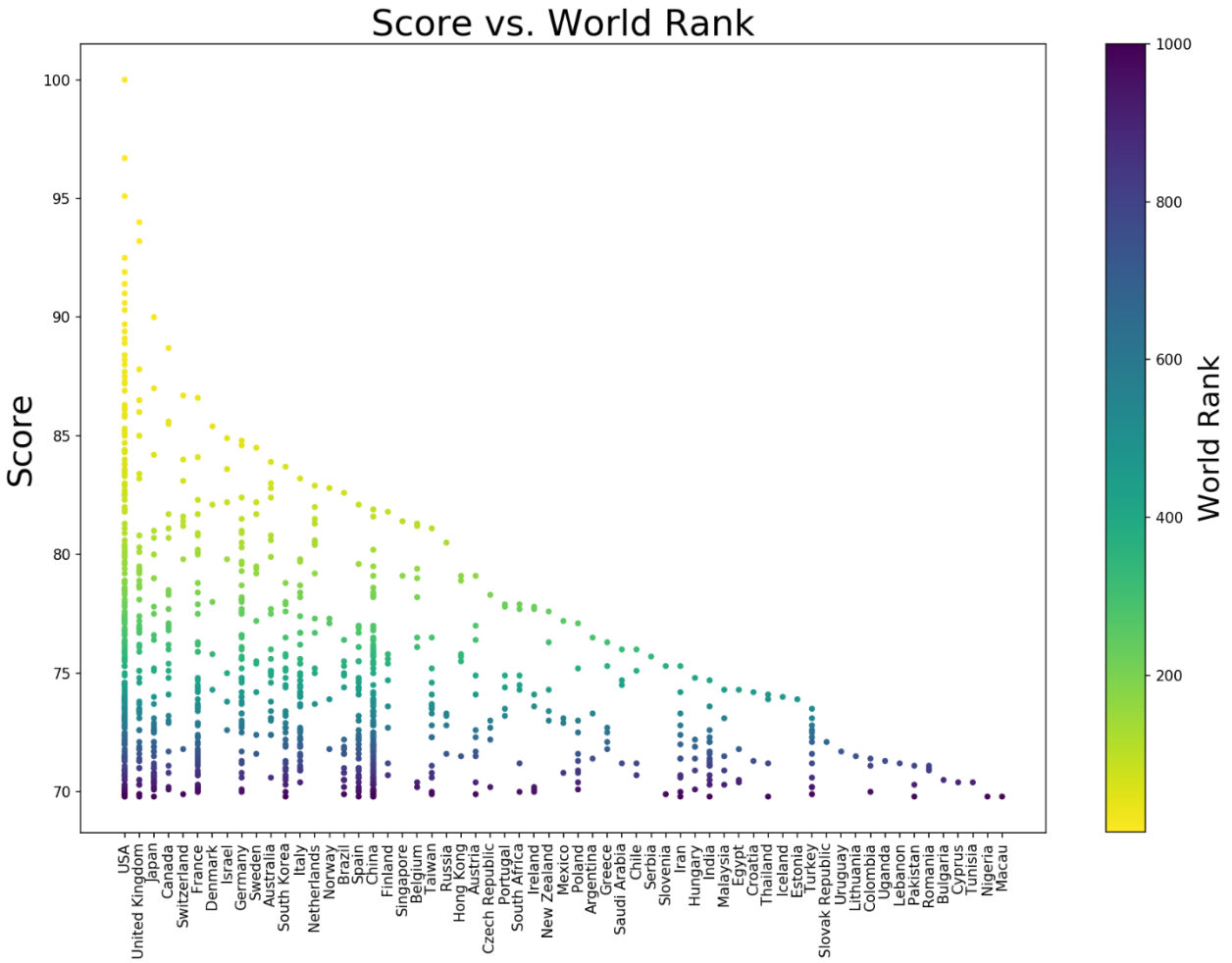
Score base on Citations:

This plot is a little different than the three plots before and that's high density of samples at specifics values of citations between 170 to 1000 range.

As the citations increases, the density and also rang of sore at that value, is increases until 550 and that's the biggest range (70 to 78).

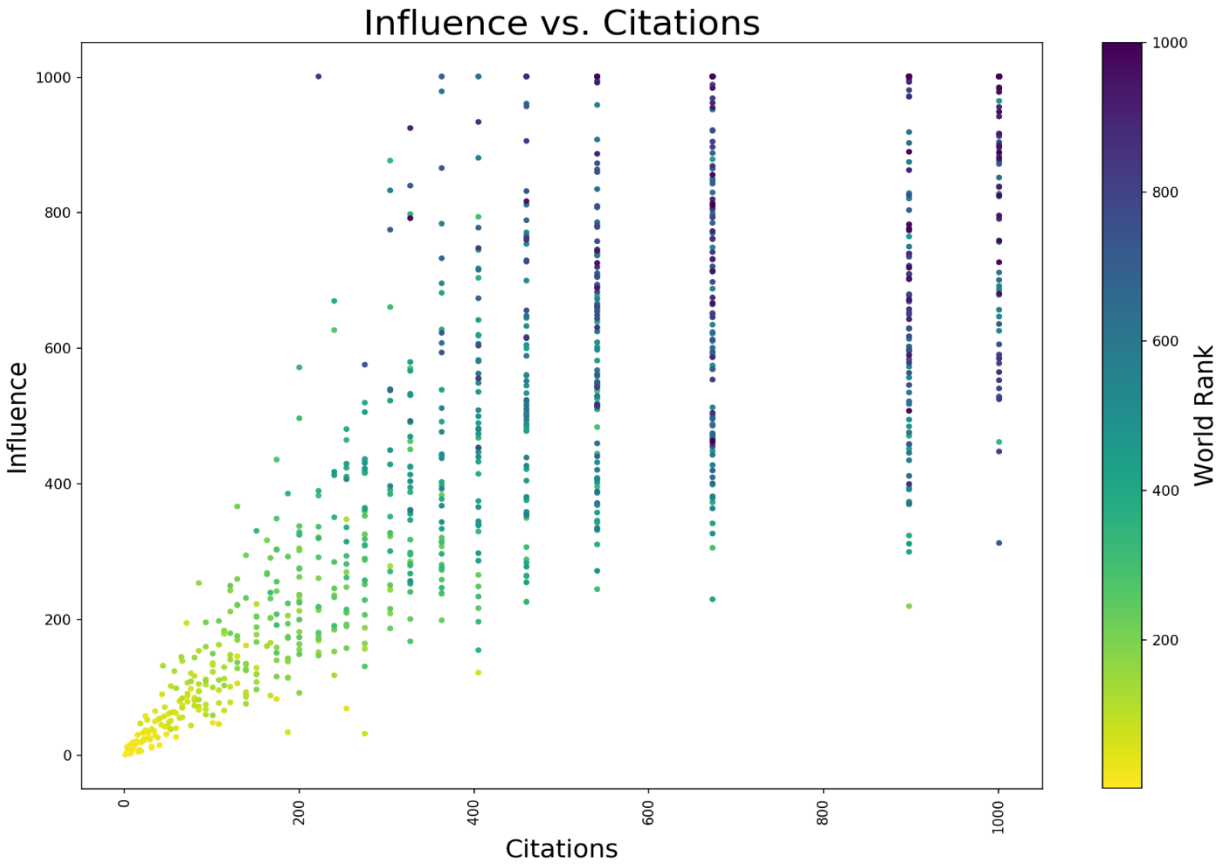
****

Now that we analyzed every feature with the target, it's time to compare some of features together.



The invers relationship is show it self in this plot and we can see the density of yellow color (world rank < 100) is pretty good in America and it's starts from 83.

We can see from "Russia" till the end, we just see green color (160 to 360 range of world rank) for the best ranks and after "Iran", it's just light blue color for the best rank.



We still see the density at specific numbers and as long the value of citations increases, the value of influence also increases and sems they have a direct relationship.

At this moment, higher two features become, the world rank is increasing as well.

But while the citations have an average value, influence get to the max value and also the world rank, so we can say "Influence" is a more affective feature than Citations on the "World Rank" and the target.

Regarding to the density of the most of the samples, we can say that "Research Output" and "Quality Publications" have a linear relationship with each other and the world rank.

The high density of the 1000 value is also because of the replacement that we did earlier.

Result: "USA" is the only country that has the most universities in the world ranking (regarding to our dataset) and the best of them in the same time, but it's not best at everything (like average score). In the top 10 countries that have the best universities base on score, Japan is the only country from Asia and Europe has the most representative.

We also learn two facts which is:

1. having more universities than other countries doesn't mean the best.
2. having less universities than other countries doesn't increase the average.

Project code and description of its steps: