DATA MINING IN WORLD UNIVERSITIES RANKING

BACHCHAR ISMAIL

CONTENTS

LIST OF FIGURES

LIST OF TABLES

* *Machine Learning and Data Mining master's program, Unevisité Jean Monnet, Saint-Étienne, France*
** *Code on GitHub*

1 INTRODUCTION AND PROBLEM UNDERSTANDING

University ranking is the rank of world universities based on several factors. Basically put, this ranking tries to answer the question of which university in best in the world in order to assist students to widely choose where to build their careers. But ranking universities is not a straightforward task, there are thousands of national and international ranking systems with conflict between each others. The three globally known ranking organizations are:

1. **The Times Higher Education World University Ranking (Times Ranking)** is widely regarded as one of the most influential and widely observed university measures. Founded in the United Kingdom in 2010, it has been criticized for its commercialization and for undermining non-English-instructing institutions.
2. **The Academic Ranking of World Universities (Shangai)**, also known as the Shanghai Ranking, is an equally influential ranking. It was founded in China in 2003 and has been criticized for focusing on raw research power and for undermining humanities and quality of instruction.
3. **The Center for World University Rankings (CWUR)**, is a less well-known listing that comes from Saudi Arabia, it was founded in 2012.

In this data mining project, the goal is to investigate three datasets provided by each ranking system, and find some interesting insights if they exist, and also compare the results obtained from each source of data to see if the three systems are ranking the same universities even they use different attributes to determine the ranking score.

2 DATA UNDERSTANDING

In this study three datasets are used, the first one is provided by The Times Higher Education World University Ranking which contains a total of 2603 rows represented by 14 attributes such as number of citations the university has received in each year. The second one is provided by The Academic Ranking of World Universities which contains the same data as the previous one, but the attributes are more representative of the quality of studies at each university such quality of education feature. The last one is provided by The Center for World University Rankings which has double the number of rows of previous datasets (about 4000 rows) and same number of attributes describing the quality of the university in the way of the number of winning Nobel Prizes by the institution's stuff.

2.1 Data Preprocessing

In this part the goal is to see in any pre-processing of the data is needed, such as handling missing values, data integration when having several sources of data and lastly but not the least data reduction techniques such as PCA if needed to reduce the dimensionality of the dataset.
In this work the data come in three separated CSVs files each one for a different source (from the three ranking systems), meaning they don't have an important number of common attributes which make the task of merging the three of them impossible. Also, each system (data source) uses their own way of calculating the score of universities thus ranking them. For all these reasons, we were restricted to only handle the missing values in each dataset separately and not merging them at any step of the whole project, but we took advantage of this and make it as comparison metric of this work, as we believe that even these three systems use different ranking strategies they should rank the same (at least know) top universities.

Missing values in CWUR Dataset
AS shown in Figure 1 there is only the 'broad impact' attribute with missing values

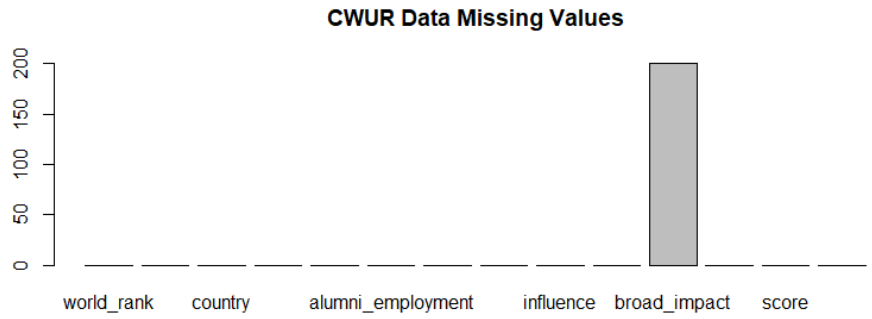

Figure 1: CWUR Dataset Missing Values

of years before 2014, but this will not cause any problem for the work of this project as the attribute is only available for this dataset.

Missing values in Times Ranking Dataset
AS depicted in Figure 2, the Times Ranking dataset has a lot of missing values for

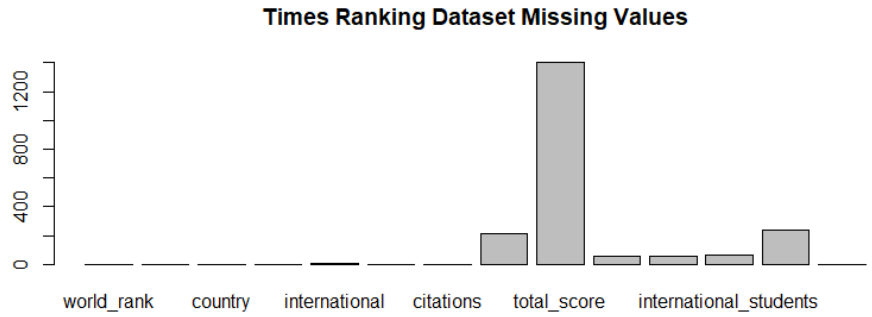

Figure 2: Times Ranking Dataset Missing Values

'total score' attribute which is a very important feature as used the one used to rank universities, with all this amount of missing values we might not rely on the use of this dataset.

Missing values in Shangai Dataset
Also, as previous datasets, the Shangai dataset has a lot of missing values, as shown in Figure 3 on the following page, for the most important feature, 'score'. Fortunately it has around 4 thousands rows; meaning we can delete these missing values, and we will still have an important number of rows to work with.

2.2 Data Exploration

In this part of the project we focus only on one dataset as they are all somehow the same. The one used for data exploration is The Academic Ranking of World Universities (CWUR). This dataset contains 2200 rows described by 14 attributes as follows:

1. **world rank** world rank for university
2. **institution** name of university
3. **country** country of each university

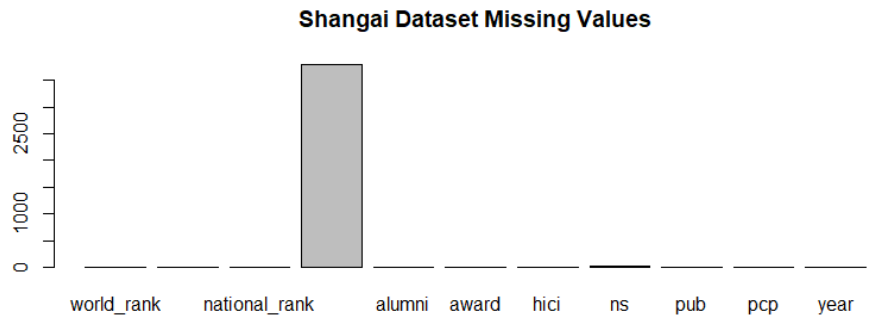

Figure 3: Shangai Dataset Missing Values

4. **national rank** rank of university within its country
5. **quality of education** rank for quality of education
6. **alumni employment** rank for alumni employment
7. **quality of faculty** rank for quality of faculty
8. **publications** rank for publications
9. **influence** rank for influence
10. **citations** number of students at the university
11. **broad impact** rank for broad impact (only available for 2014 and 2015)
12. **patents** rank for patents
13. **score** total score used for determining world rank
14. **year** year of ranking (2012 to 2015)

Top ranked countries all the time

We can see in this dataset that the top ranked countries, in Figure 4, in terms of number of their ranked universities are expected within the USA, China and France are presented.

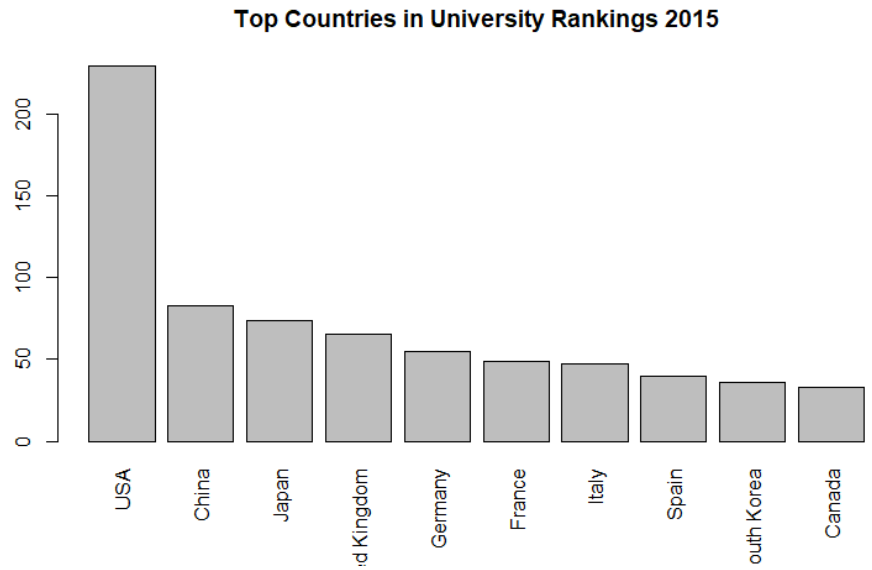

Figure 4: Top Countries in University Ranking 2015

Top 10 ranked universities from 2012 to 2015

In Figure 5 on the following page we see, as expected, the well known universities such as Harvard and MIT are always at the top.

Correlation analysis

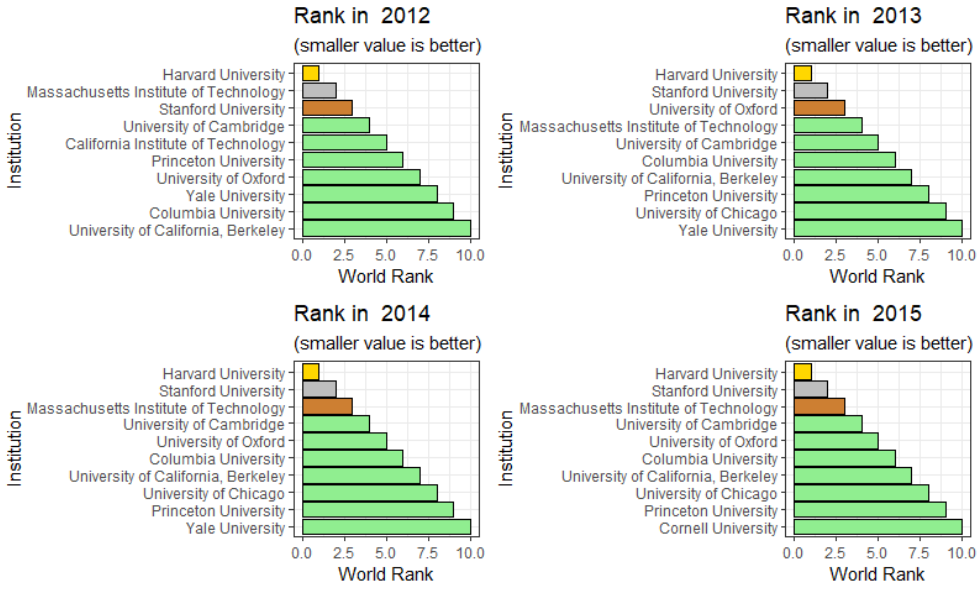

Figure 5: Top 10 ranked universities from 2012 to 2015

In all datasets used in this project the rank attribute is very important. To know how it is influenced by other variables we do a correlation analysis.

We choose France as a showcase for this analysis, and we want to find the variables that are correlated with national rank attribute that is used to rank French schools and universities in France.

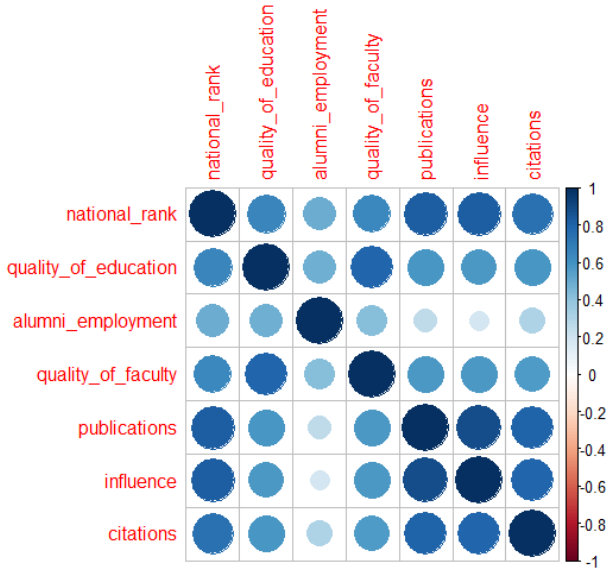

Figure 6: French Schools Features Correlation

From the Figure 6 we see that the national rank is correlated with a lot of variables namely with publications and citations.

Now we see the regression of national rank and publications and then with citations:

As shown in Figure 7 on the following page, the national rank is highly correlated with the number of publications the school is making. Not surprising as we know the rank is calculated based on that.

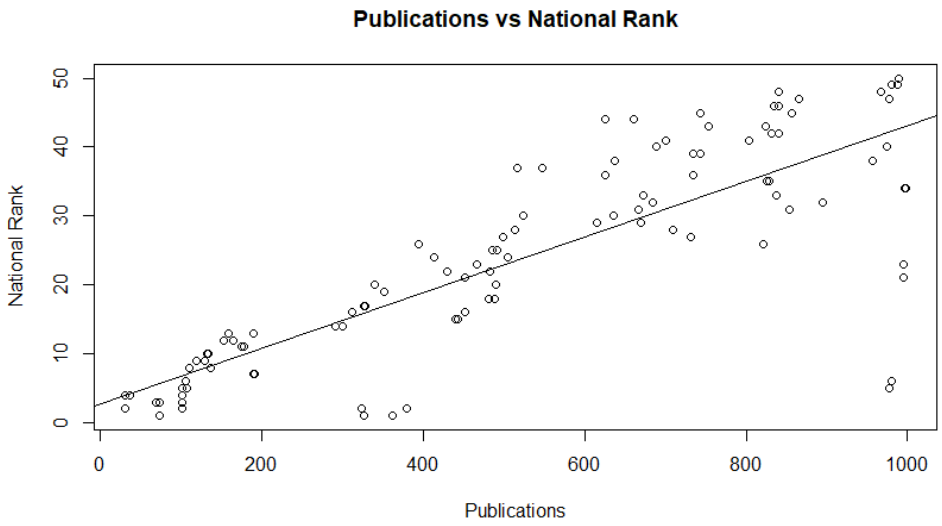

Figure 7: Regression of national rank and publications

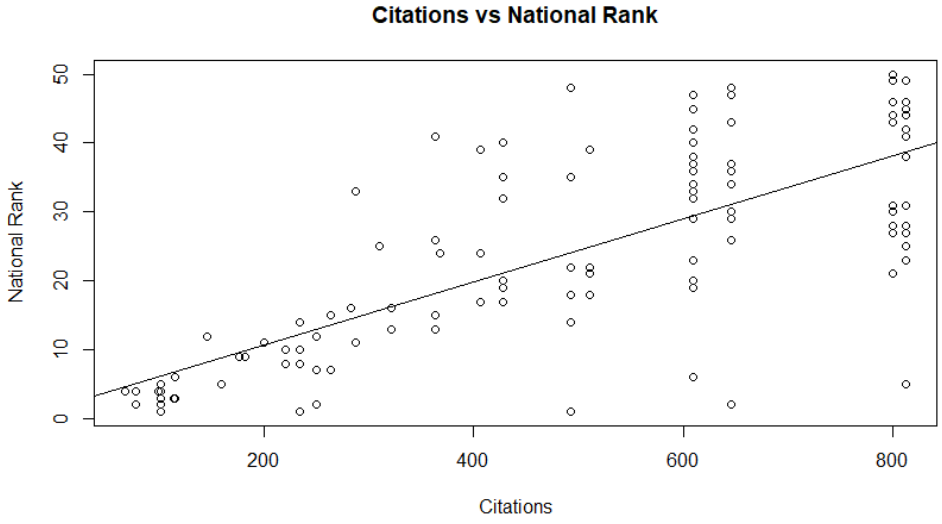

Figure 8: Regression of national rank and citations

AS depicted in Figure 8, the national rank is regressed by the number of citations that the school receives, which is again not surprising as the publications' are weighted by number of citations.

3 MODELING AND EVALUATION

In this part we go through the algorithms used in this project and for each one we provide its settings, used variables and parameters, and the obtained results using different datasets (provided by different ranking systems) in order to compare and to ask the question : are the three different ranking systems rank deserved to be on top universities without any biases?

3.1 Clustering

For the clustering task, the objective is to cluster the universities in a reasonable number of classes (clusters) for two datasets, the CWURD and Shangai datasets.
The clustering method used is Kmeans with different k (number of clusters) values, and we only restricted this modeling to two important attributes university 'score' and 'world rank' which. The goal is to find out if the two ranking systems agree on the clustering of universities.

In order for Kmeans to perform well we have to decide on the number of clusters to be used. For this latter the Elbow method is used to determine the parameter k. Also, to verify the choice of k, Silhouette method is used.
The first method uses within-cluster sum of squares to the cluster centroid to tell the optimal number of clusters. The number of clusters k can be found where the line drastically falls forming an elbow. The second method determines how well a data point fits into a particular cluster, with high average silhouette width values indicating a cluster's number solution.

3.1.1 Kmeans on CWURD Dataset

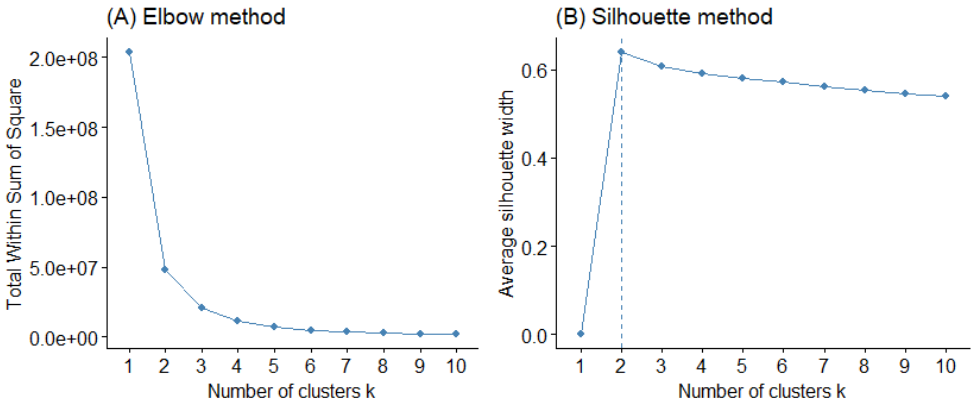

Figure 9: Elbow and Silhouette methods to determine k on CWURD Dataset

As shown in Figure 9, the value of k, number of clusters, could be 2 which is supported by the two methods. However, to my not really skillful eye, there is a potential of three clusters hinted by the elbow plot. Therefore, two values of k are used, k=2 and k=3.

2 Clusters

In the following Figure 10 on the next page the data is perfectly clustered into two groups, the first contains universities with high score and top world rank and second group or cluster contains low-ranked universities.

3 Clusters

To further understand the previous clustering, another clustering was obtained with 3 clusters, as depicted in Figure 11 on the following page where universities are grouped into 3 clusters. This makes sense as we only further divided the second cluster obtained from previous clustering. Precisely, we get top-ranked (in green) universities, medium-ranked universities (in black) and low-ranked-universities (in red).

3.1.2 Kmeans on Shangai Dataset

As in the previous case, the values of k are chosen to be k=2 and k=3 which they are supported by the two methods shown in Figure 12 on page 10.

2 Clusters

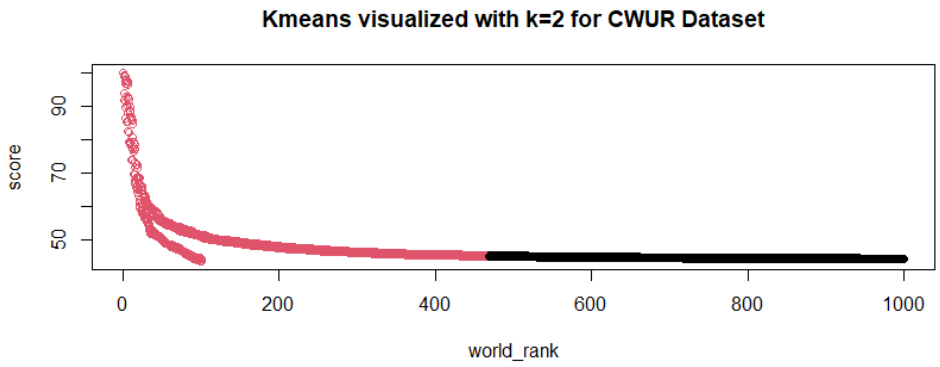

Figure 10: Kmeans visualized with k=2 for CWUR Dataset

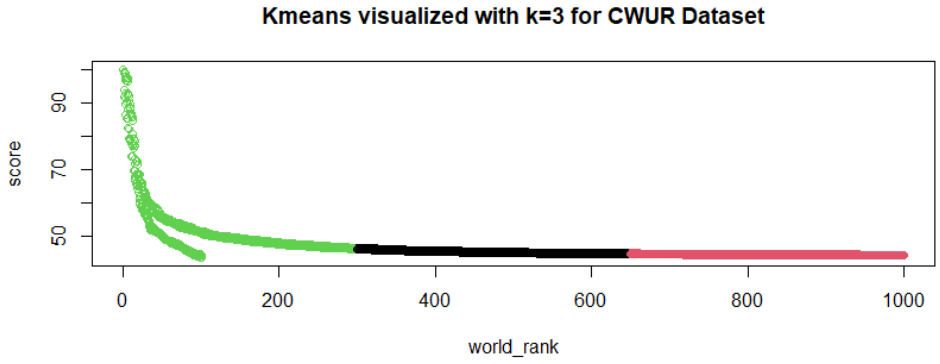

Figure 11: Kmeans visualized with k=3 for CWUR Dataset

In the Figure 13 on the following page the data is perfectly clustered into two groups, as in previous dataset. The first cluster contains universities with high score and top world rank and the second cluster contains low-ranked universities where they are not ranked among the top 40 universities, and they have a score lower than 30.

3 Clusters

Again, based on result shown in Figure 14 on page 11, the 3-clustering is exactly similar to the one obtained from CWUR dataset. This means that the two sources are reliable in ranking the universities within we found the same universities are grouped in the same clusters, following the top-medium-low ranked universities clusters.

To evaluate the clustering results the BSS- Between Sum of Squares, CH - Calinski and Harabasz, and BH - Ball and Hall clustering quality indices are used. The measures of these indices are presented in the Table 1 on the following page. But it is hard to say if the clustering is good or bad based on these measures. But the good thing is that the measures are close enough to each other in every dataset which supports our previous insight about the reliable similarity between the two ranking systems.

3.2 Regression

In this regression part the focus put on The Times Higher Education World University Ranking dataset, where the goal is to predict either Total Score or World Rank from the other available variables such as university income and teaching quality.

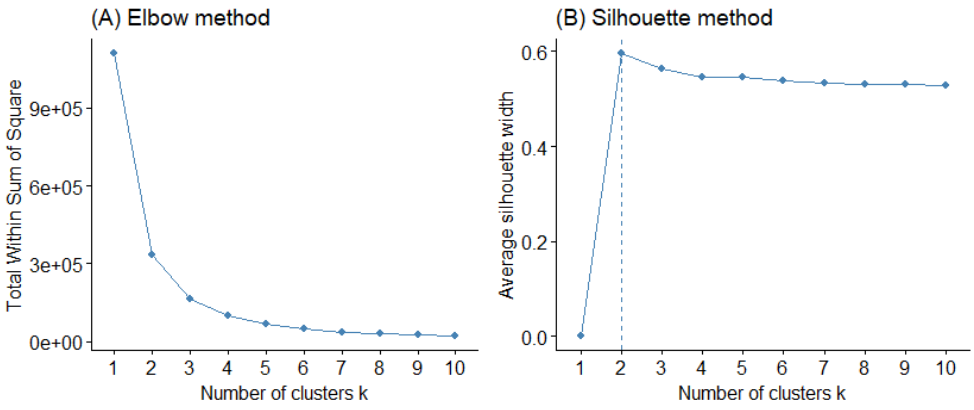

Figure 12: Elbow and Silhouette methods to determine k on Shangai Dataset

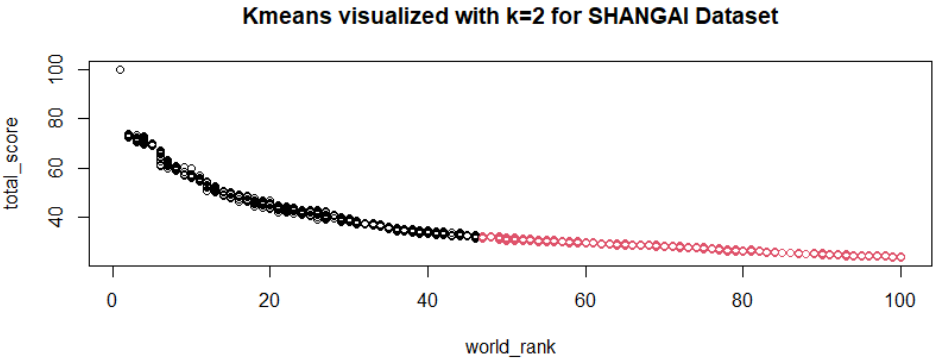

Figure 13: Kmeans visualized with k=2 for Shangai Dataset

The need of doing regression in this dataset required us to re-do some preprocessing on the dataset to make it perfectly ready for numerical analysis.

To start with, we charged some variables such as world rank, which is for some universities is given as an interval, and number of students and international students to numerical values. This latter, introduces some missing values which we had to deal with. To solve this problem, we followed a data imputation approach where we filled only the dataset missing values (not the ones in intervals). The result of this method is displayed in Figure 15 on the next page and in Figure 16 on page 12 where we can see a map of these missing values in all features of the dataset. In Figure 16 on page 12 we see that one of the most important variables which is Total Score has a lot of missing values.

After the data imputation we ended up with 13,015 number of samples in the dataset. Then we divided this into a training set of 9110 samples and a testing set of 3905 samples.

Before doing the regression, in Figure 17 on page 12 is shown the correlation matrix between the variables in the dataset. We see that the World Rank is only

Table 1: Quality Measure of Clustering for Different k Values on CWUR and Shangai Datasets

	CUWR Dataset				Shangai Dataset			
k	WSS(%)	BSS(%)	BH	CH	WSS(%)	BSS(%)	BH	CH
2	23.57	76.43	11.79	7126.95	30.15	69.85	15.07	2546.61
3	10.9	89.91	3.36	9792.21	14.77	85.23	4.92	3166.1

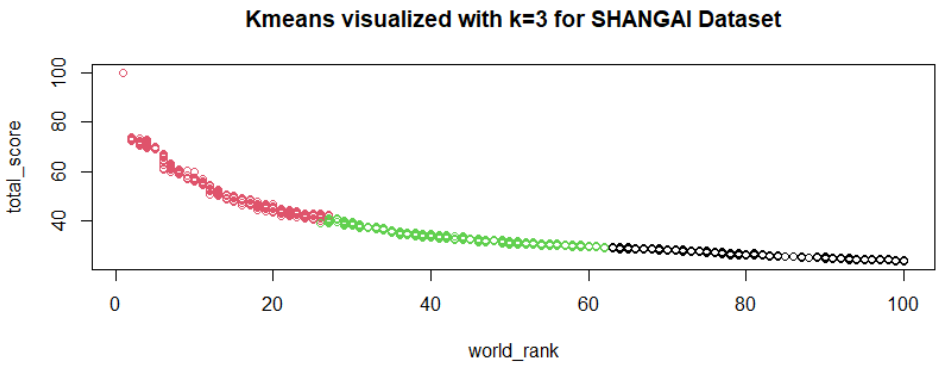

Figure 14: Kmeans visualized with k=3 for Shangai Dataset

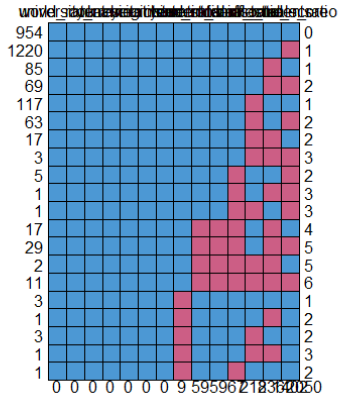

Figure 15: Map of Missing Values in Times Ranking Dataset

correlated with research activities of the university. And, the Total Score variable is correlated with the teaching quality and research activities of the university.

3.2.1 *Regression of World Rank Variable*

In this part multiples regressors are used in combination and alone to see if they are able to predict the world rank of a university. The Table 2 on page 13 summarizes all the experiments.

In all the models we see how they perform well on the training set but very badly on the testing set. This could be caused by the pre-processing done earlier, or simply the ranking system uses a non-linear relationship between the variables to determine the 'total score' and 'world rank' variables.

Regression of 'world wank' with 'international' variables

As depicted in Figure 18 on page 13 and Figure 19 on page 14, the world rank variable is hard to be predicted using other variables, especially the international variable even if it is the most correlated with variable. But as with in other models, the model performed very well on the training dataset.

3.2.2 *Regression of Total Score Variable*

In the other hand, the Total Score variable which is correlated with Teaching and Research variables, seems at first an easy target to predict, but it turned out it is as hard as the World Rank. All the models performed well on the training set, and again they did very bad on the testing set. As shown in Figure 20 on page 14 and Figure 21 on page 14 the most correlated variable, International, with Total Score could not generalize to unseen data, testing set.

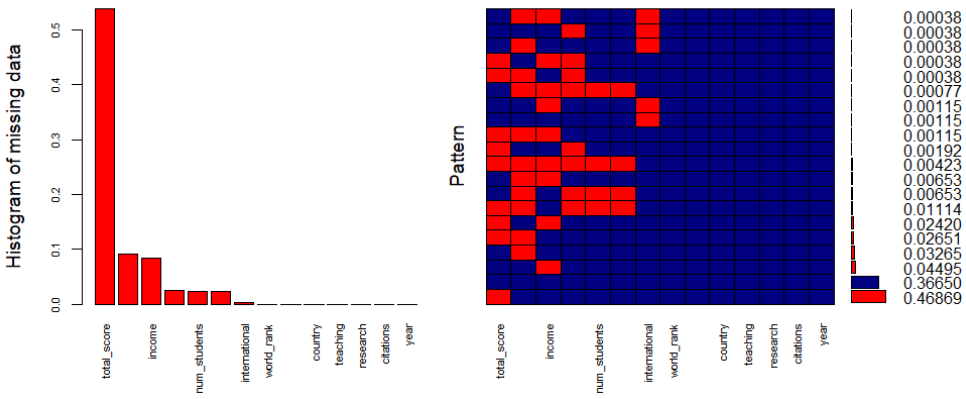

Figure 16: Pattern of Missing Values in Times Ranking Dataset

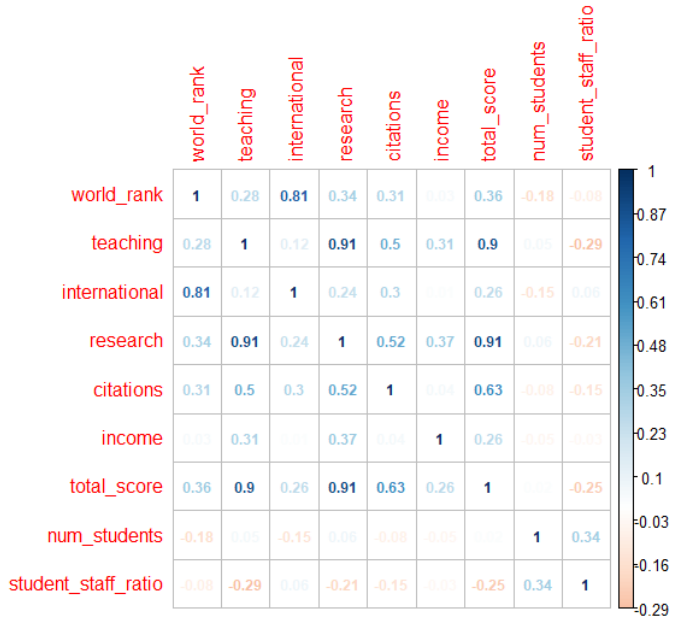

Figure 17: Correlation matrix of variables in Times Ranking Dataset

4 CONCLUSION AND FEATURE WORK

In this project the goal was to investigate the data behind the three known world ranking of universities systems. In doing so, multiple data mining approaches were used. Firstly, is exploratory data analysis, where just from the visualization and basic statistics of the data we were able to extract some insights such as if the most known university are ranked in top, we mean by known in terms of their available educational materials in the web and number of publications such an example here is Harvard University. Secondly, a clustering task were made to classify the similar universities into one group. In this task the Kmeans method was used with several number clusters that were defined earlier by the Elbow method. The goal in this part was to compare two ranking systems, Shangai and CWUR, in terms if they rank universities without any biases towards any part. As a result, in fact we found the two systems provided data clustered roughly the same universities in the same groups. Lastly, we wanted to know if we can infer the ranking method used by the Times Ranking organization. In doing so, a regression analysis were done in order to infer of exist any linear relationship between the World Rank variables and the rest of variables and also between the Total Variable and the rest. The result of this analysis, showed that it was hard to predict neither the World Rank nor the Total

Table 2: Evaluation of different Regression Models for World Rank and Total Score Variables

Regression model	Train MSE	Train R^2	Test MSE	Test R^2
world rank~research,citations,international,teaching, total-score,student-staff-ratio,num-students,income	33.15	0.7	32.68	0.71
world rank~research,citations,international,income	35.32	0.68	34.65	0.69
world rank~international	37.89	0.66	36.85	0.67
total score~research,international,student-staff-ratio	24.91	0.83	1384.16	-11.22
total score~research,citations,international,income	20.31	0.86	1386.9	-11.24
total score~research,citations,international,income, student-staff-ratio	19.87	0.86	1385.44	-11.23
total score~research,teaching	21.88	0.85	1395.64	-11.32
total score~research	25.93	0.82	1393.91	-11.3

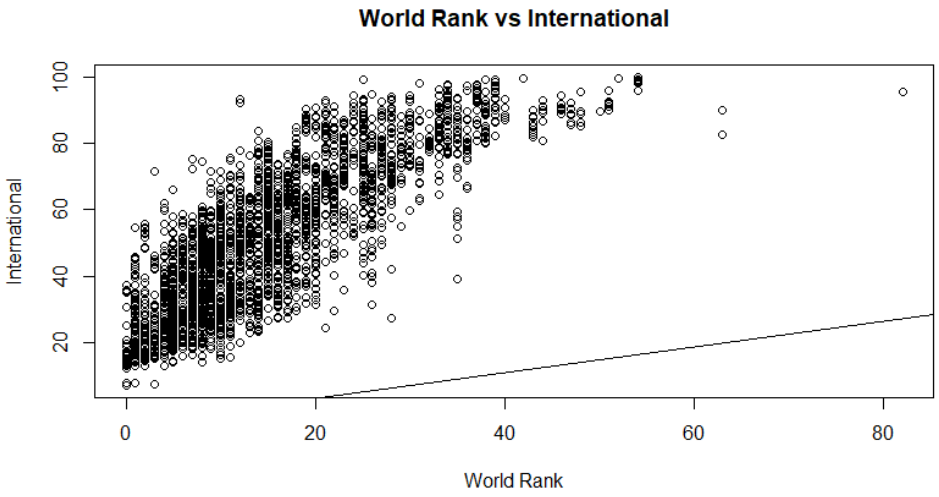

Figure 18: Regression of World Rank by International Variable

Score variables from the other variables such as income and number of students even though they are correlated with some of them.

To sum up, the most important part of data mining in general is the availability of dataset which was the main problem in this project. The datasets from the different sources were not that good as they were all scraped from the official websites of the three ranking systems. The scrapped data need a lot of cleaning which also requires having a huge number of instances so that cleaning will not reduce so much the dataset. In this project the data were only around some thousands of samples which is again not an important number if more data cleaning is needed.

The future work of this project is to scrap the data by myself this way I will be free to choose what variables to include and what not. Also, I could scrap as much data as needed and this way the data pre-processing part will not hugely reduce the dataset. Furthermore, it could be a good idea to put in production a model that is able to re-rank universities and schools based on the data provided by the three known ranking systems. This way students struggling to choose where to continue their studies will be to see a comparison of the three systems in one place without having to go to every source individually.

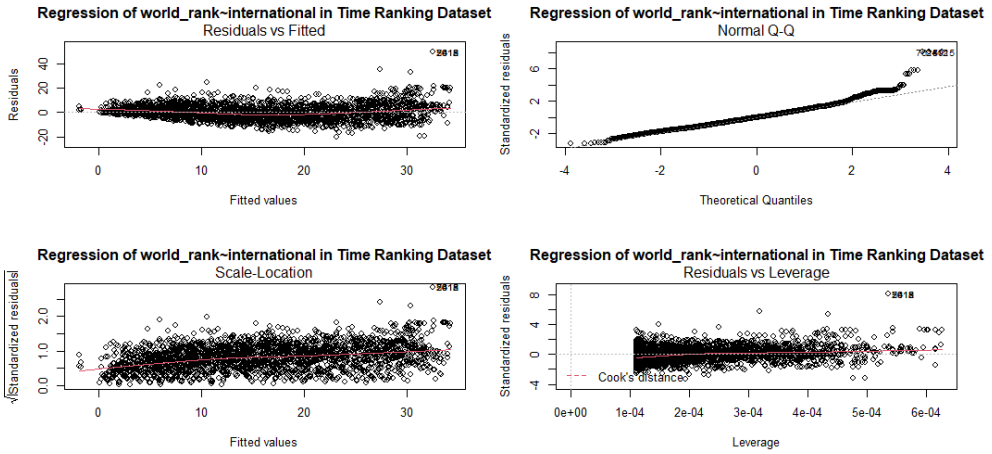

Figure 19: Regression Result of World Rank by International Variable

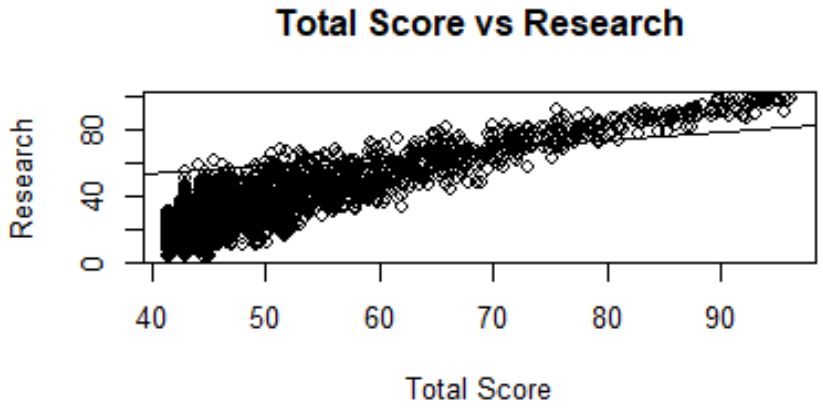

Figure 20: Regression of Total Score by Research Variable

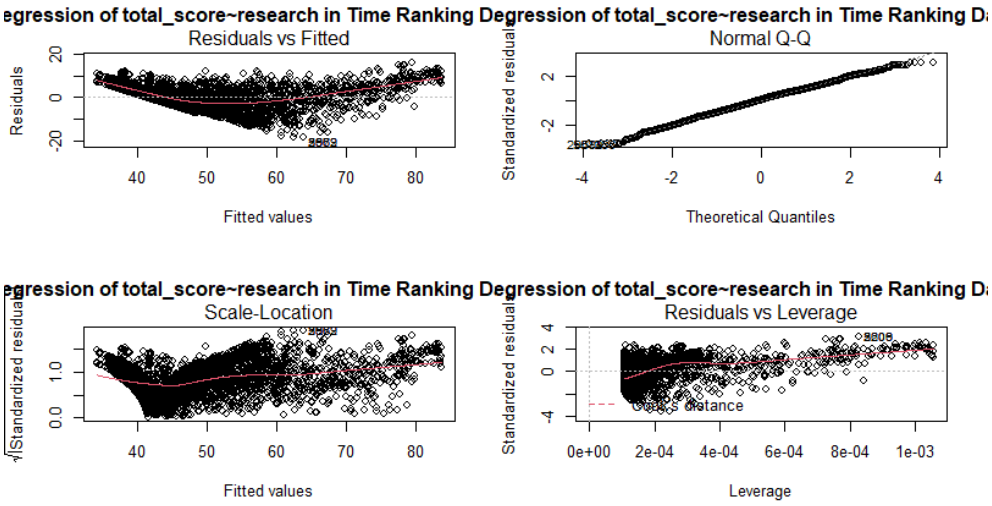

Figure 21: Regression Result of Total Score by Research Variable