**Assignment 2**

**Big-Data Systems and Intelligence Analytics**

**CSYE 7245 - SPRING 2018**

PROFESSOR:

**Nik Brown**

Report By:

**Harshad Pardeshi**

Table of Contents

[Problem Statement 3](#_Toc505359548)

[Data Description: 3](#_Toc505359549)

[Part A: Get some data 4](#_Toc505359550)

[Data cleaning 4](#_Toc505359551)

[Part B: Exploratory data analysis in Python 7](#_Toc505359552)

[Data pre-processing 9](#_Toc505359553)

[Implementing Machine Learning Algorithms 9](#_Toc505359554)

# **Problem Statement**

Find a public dataset or machine learning competition and use machine learning techniques to analyse the data. This should be REVIEW of what you learned in INFO 7390 - Advances in Data Sciences and Architecture.

I have selected the following data set:

**World University Rankings** - Investigate the best universities in the world

Of all the universities in the world, which are the best?

Ranking universities is a difficult, political, and controversial practice. There are hundreds of different national and international university ranking systems, many of which disagree with each other. This dataset contains three global university rankings from very different places.

# Data Description:

**University Ranking Data**

The Times Higher Education World University 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.

The Academic Ranking of World Universities, 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.

The Center for World University Rankings, is a less well-known listing that comes from Saudi Arabia, it was founded in 2012.

# **Part A: Get some data**

In this section, we will perform the data the following operations:

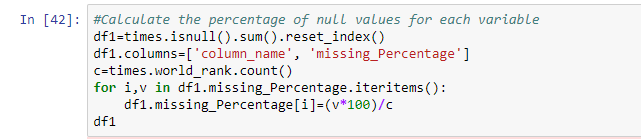
* Data cleaning
* Exploratory data analysis in Python
* Storing clean data on Amazon S3 bucket

Also, we will answer professor’s Questions with relevant code snippets in this section

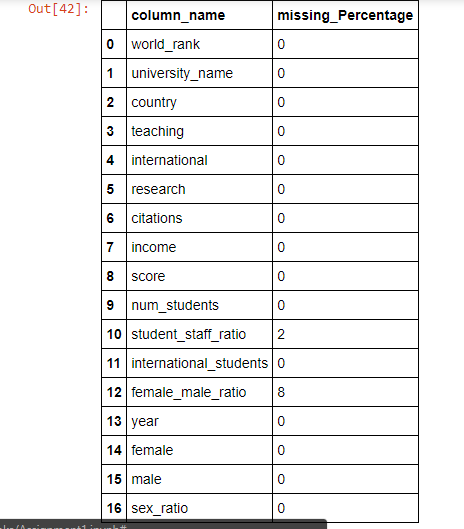
## **Data cleaning**

This section comprises of handling some special and important missing data. We get the data in form of csv files from Kaggle, on which we perform various alteration and changes on the blank data without changing the meaning of the data available and has some logical significance.

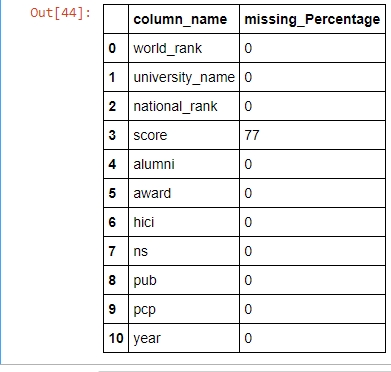
We first analyzed all the columns having empty values as shown below:

Code:

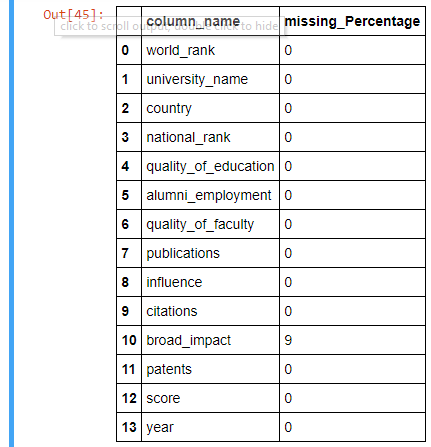
Output for times data



Output for sangha



Output for CWUR



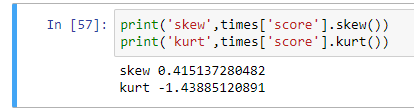
To filter data, I started off with eliminating data prior to 2012 as Center for World University Rankings was started from 2012 and we need to combine data from all three sources. Further, we find all unique universities in all three data sets and combine all universities. Based on this set of universities we filter universities that are ranked by all three institutes.

We fill the missing values of student to staff ratio and broad impact with the mode and mean of that column respectively. Screen Clipping

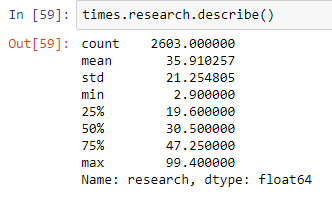
* Are there inappropriate values?
  + No. There are no inappropriate values
* Remove or impute any bad data.
  + Functions are written to convert data into appropriate formats



* How is the data distributed?
  + Data is right skewed.



* What are the summary statistics?



* Are there anomalies/outliers?
  + No, there are no anomalies in data.
* Plot each column as appropriate for the data type:
* Write a summary of what the plot tells you
  + Explanation of each plot is done in Exploratory Data Analysis.
* Are any of the columns correlated?
  + The columns are correlated and this is shown in exploratory data analysis

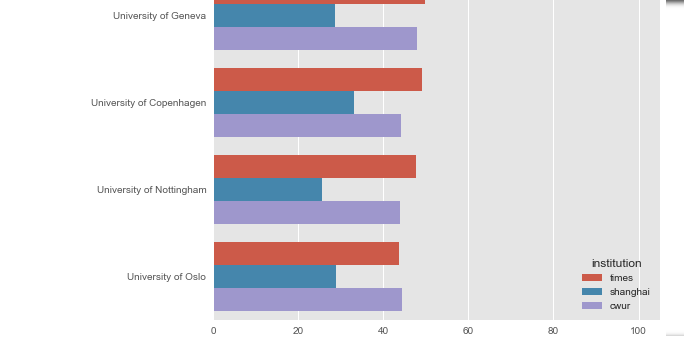
# **Part B: Exploratory data analysis in Python**

**Analysis 1:** Compare scores allocated by all three institutes to every university

**Approach:**

Join the above dataframes and keep only universities present in all three datasets. Group this dataframe for every university and count the score of each university.

A graph is drawn which represent this flow.



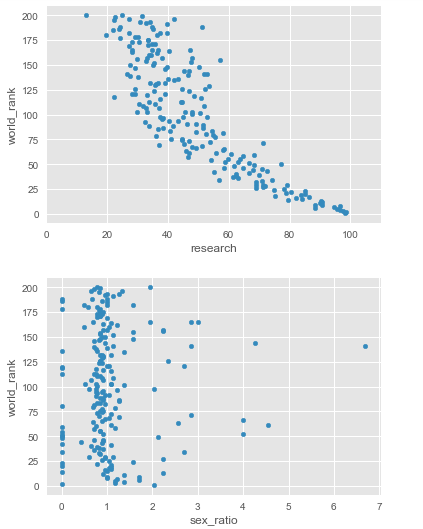
**Conclusion:** This graph shows us the comparison between scores assigned by each institutes in year 2012.

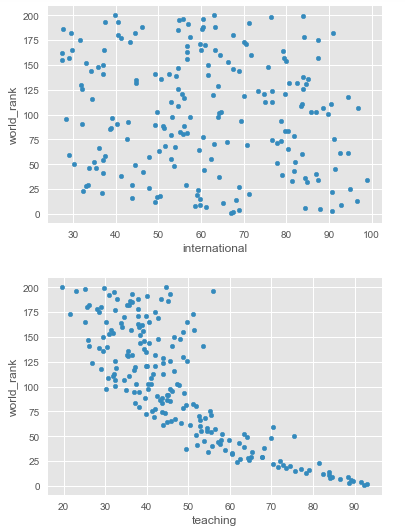
**Analysis 2:** Compare research, teaching, citations, sex ratio and scores

**Approach:**

From the above dataframe use all combinations to compare these factors affecting the ranking of university.

A series of graphs are drawn which represent this flow.





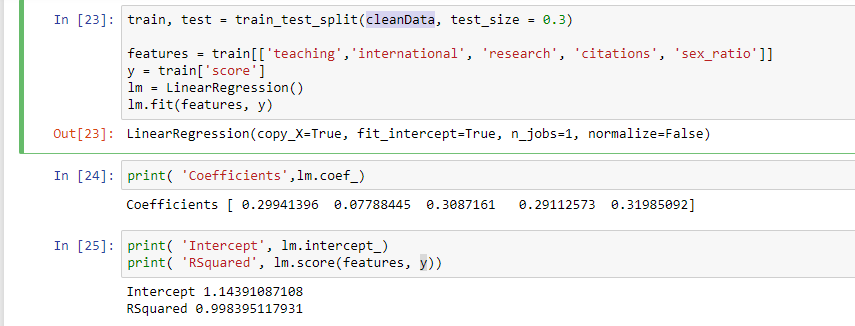
**Conclusion:** This graphs shows us that ranking largely depends on teaching, research and citations.

## **Data pre-processing**

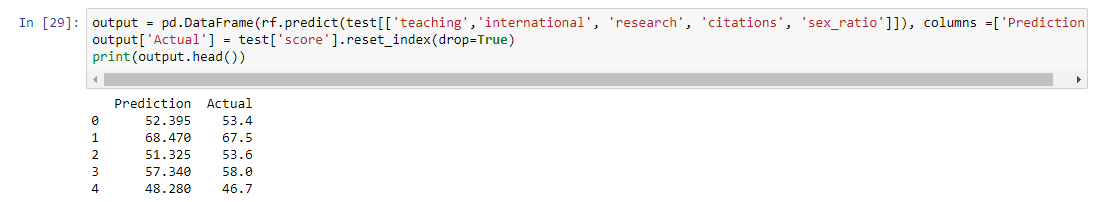
Converting all the columns to appropriate format so that we can use only numeric data ignore all junk characters. This is done using same functions mentioned above. Also we fill missing value with ‘0’ for columns like citations and availability of facilities.

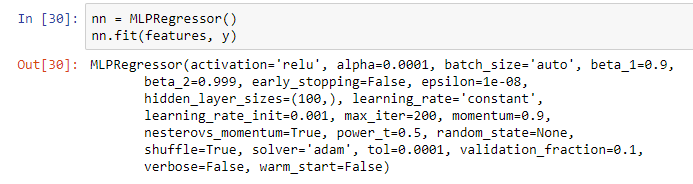
## **Implementing Machine Learning Algorithms**

* + - The first step in implementing machine learning algorithms is to Create a training dataset I have done that by including variables like ‘teaching’, ‘international’, ‘research’, ‘citations’, ‘sex\_ratio’ from the ‘cleanData’ dataset. Sample output i.e. ‘score’ in the form of a series of values form the ‘train\_y’ dataset.
    - Next step is to split data into training data and testing data. We randomly select 80% from data to train models and remining 20% to test the models.
    - We then implement linear regression and observe the predicted scores of universities. The predicted values are saved in .csv file named ‘TimesLinear.csv’



* + - We then implement Random Forest Regressor for 5,10,15 and 20 variables and check the prediction of scores. Top 1000 rows of the output are saved in .csv file named LogErrorPrediction with a time stamp. Few of the predicted values are as follows



* We then implement Neural networks to check if can give us better results than the above two models implemented

But this model gives us R squared value of 0.8899. Thus, we can conclude that random forest with an estimator of 20 gives us a better model of all.

# **Conclusion:**

It was a very interesting dataset to work with. High percentage of variations in data, but even after picking normal data and ignoring junk values, my machine learning model had proper data to train and get the desired results.

# **References:**

* https://www.kaggle.com/mylesoneill/world-university-rankings
* https://www.kaggle.com/shabaaaa/comparing-university-rankings