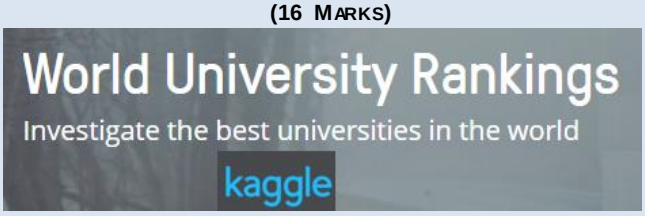
CS24210/CS54111 Introduction to Data Science

Dr. Sabah Mohammed
Department of Computer Science

The Class Project: Investigate World University Rankings

(16 Marks)



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 Kaggle dataset contains three global university rankings from very different places.

Download the Kaggle Dataset from:

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

The Kaggle University Ranking Dataset

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.

To further extend your analyses, Kaggle have also included two sets of supplementary data.

The first of these is a set of data on **educational attainment around the world**. It comes from The World Data Bank and comprises information from the UNESCO Institute for Statistics and the Barro-Lee Dataset. How does national educational attainment relate to the quality of each nation's universities?

The second supplementary dataset contains information about **public and private direct expenditure on education across nations.** This data comes from the National Center for Education Statistics. It represents expenditure as a percentage of gross domestic products. Does spending more on education lead to better international university rankings?

Kaggle have placed these datasets as a competition with lots of people providing their contributions and programming scripts (in R or Python) for addressing variety of issues and questions such as:

Ranking my University?
Ranking Canadian Universities?
Ranking Research of My University?
Ranking Teaching of My University?
Ranking my University over Time?

You are requested to raise some similar or different questions and answer them using the various data analytics that you have learned during this course. Marking this project will depend on your role approaching this data as a data scientist having exposed to variety of data analytics tools including visualization, knowledge extraction and machine learning. One thing you may note that Kaggle website includes variety of scripts as well as other scripts available at many other places on the internet or published somewhere. In this direction, you are **not allowed** to copy and use their scripts and solutions to the questions that they have raised. I am expecting that you work individually on formulating certain questions (after studying the different set of data) and start using your data science skills to answer them. One last thing, **you are not allowed to publish your findings on Kaggle or anywhere else without the proper written permission of our academic system starting with my approval. Your submission is only for the purpose of this course project on D2L**.

Requirement:

1. You need to use RStudio and the knowledge that you understood from Lectures to work on this project.
2. You need to save each observation solution in a file named **observation_number.r** and change the word number according to the observation number that you are making when trying to analyze universities rankings. Zip all these R files into one file that need to be named as follows **Project_STDID.zip** **(any other Zipping like RAR will not be accepted)** and change the STDID with your student number. Submit the Zip file to D2L before the due date/time. You must include **README.pdf** **(MS Word or WordPad will not be accepted)** file to describe your solutions with screenshots (Code + Run Output) of your outputs on the RStudio.
3. In addition to the ZIP file, you need to record a video for max five minutes explaining your project. There will be three marks on the video. Your video must play on **Windows Media Player** otherwise will not be accepted.

Remember that the best way to learn Data Science is to do Data Science. There is no substitute to it.