Spreadsheet Data Analytics

CONTINUOUS ASSESSMENT 2 – DATA EXPLORATION

ASSESSMENT

This assessment is due on Sunday 7th November 2021 and is worth 10%. It should be submitted through Moodle in the form of a single excel file which includes sheets as detailed below. The Marking scheme is as follows:

Step 1: Variable Categorisation: 5%

Step 2: Data Cleaning: 10%

Step 3: Univariate Data Exploration: 30%

Step 4: Addition of Continent Variable: 10%

Step 5: Outliers: 10%

Step 6: Bivariate Data Exploration for Response Variable: 25%

Spreadsheet Structure including Notes Sheet: 10%

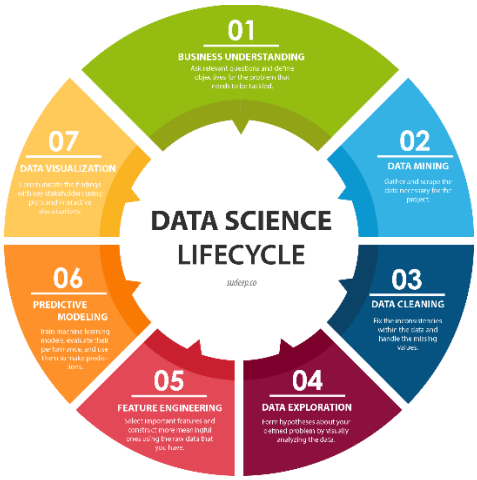
Spreadsheet structure including notes and details of checks and cleaning: 10%

Analysis of features 1-5 & Prediction 1: 90% (15% each).

Background

In this assignment you will focus on the **Data Exploration** step (Step 4) in the Data Analytics Lifecycle shown below through a real data set with details of the scoring for the top 800 universities in 2016.

Data Analytics Lifecycle



Background

The data set in the file **UniversityRankings2016.csv** file gives the name, country of the 800 top rated universities in 2016. It also gives the total score (0-100), the scores for the categories of Teaching, Research and Income (all 0-100), the number of students and the percentage of students that are classified as international.

Step 1: Variable Categorisation

Categorise each of the variables in the data set as Numerical (Discrete or continuous) or Categorical (Nominal or Ordinal). You should also identify the key response or dependent variable and the variables which are predictors or this variable. Include this information on a sheet named **Step 1- VariableCategorisation**.

Step 2: Data Cleaning

Perform whatever checks and cleaning on the data as you see fit. Note in a separate sheet named **Notes** the checks you have made and any changes to the data when cleaning it and also the details of the cleaning checks you have performed. You should also use this sheet to detail the *version history* of the document as well as any other important notes.

Step 3: Univariate Data Exploration

Explore the distribution of each variable by producing suitable plots and measures for centrality and spread. Comment on the distributions and the implications for the measures that you would use when analysing the data. You should include a sheet for each variable. Name these sheets **Step3-Country**, **Step3- ResearchScore** etc

Step 4: Addition of Continent Variable

Given that there are approximately 70 different countries in the data set, it may be useful to include a continent field in the data set. Using the **continents.xlsx** file provided, and with the help of the **Match()** and **Index()** functions in Excel, add a new column named Continent which gives the continent for each university. You should perform appropriate univariate exploration for this variable on a new sheet named **Step4-Continent**.

Step 5: Outliers

Based on your univariate exploration identify the outliers for each variable. Discuss the options as to how to treat these outliers. Include this information in a sheet named **Step5-Outliers**.

Step 6: Bivariate Data Exploration for Response Variable

Perform appropriate bivariate analysis to understand the relationship of each of the variables with the response variable you have identified. Include suitable measures and plots with a short summary of the relationships.

You should pay particular attention to the relationship between the Continent variable and the response variable.

Include a short statement in each sheet on the relationship between each of the variables and the response variable.

Name the sheets **Step6 – ResearchScore,** etc

Step 7: Optional – Bivariate Data Exploration for Relationship between non-response variables.

If you want you can perform appropriate bivariate analysis to understand the relationship of each of the non-response variables with each other.