# Project – PDA Course

**Title:** Data analysis on a merged dataset after cleaning.

**Submission Details**: The submission has two parts:

1) **Project Report (20 marks) – due by 21-Jan-2025 10pm**: A well formatted Python notebook that has the code for answering each task and includes your insights, visualizations, and comments. Make sure to have a nicely designed cover page with Project title, author names and submission date. Include an Executive Summary at the beginning of your notebook. The notebook should be named ***firstname_project_report.ipynb***.

2) **Presentation (10 marks) – due 22-Jan-2025 9am – 2pm:** A nicely designed Powerpoint/Google Slides/Canva presentation that highlights key findings, insights & recommendations. Maximum 15 slides. You will have 10 minutes to present it and 3 min for Q & A. Post the class, you would e-mail your presentations to me. There are 6 marks for presentation content and 4 marks for presentation delivery. Please be professionally dressed and stick to allocated time limits. Ensure that your presentation is optimized for Zoom in terms of font-sizes, colors and design. The presentation should be named ***firstname _project_presentation.pptx (pdf would be fine too)***.

Note: Marks Indicated here are marks for the report. I shall share a separate Rubric for presentation.

## Objectives:

Q1) [Advance] **[TOTAL – 9 MARKS]**

a) Read the energy data from the file **Energy Indicators.xls**, which is a list of indicators of [energy supply and renewable electricity production] from the [United Nations] (http://unstats.un.org/unsd/environment/excel_file_tables/2013/Energy%20Indicators.xls) for the year 2013, and should be put into a Data Frame with the variable name of **energy**. Keep in mind that this is an Excel file, and not a comma separated values file. Also, make sure to exclude the footer and header information from the data file. The first two columns are unnecessary, so you should get rid of them, and you should change the column labels so that the columns are: **[1 Mark]**

['Country', 'Energy Supply', 'Energy Supply per Capita', '% Renewable']

b) Convert variable Energy Supply to gigajoules (there are 1,000,000 gigajoules in a petajoule). For all countries which have missing data (e.g. data with "...") make sure this is reflected as NA values. **[1 Mark]**

c) Rename the following list of countries. **[2 Marks]**

"Republic of Korea" to "South Korea",
"United States of America" to "United States",
"United Kingdom of Great Britain and Northern Ireland" to "United Kingdom",
"China, Hong Kong Special Administrative Region" to "Hong Kong"

There are also several countries with parenthesis in their name. Be sure to remove these, e.g. `'Bolivia (Plurinational State of)'` should be `'Bolivia'`.

d) Next, load the GDP data from the file **world_bank.csv**, which is a csv containing countries' GDP from 1960 to 2015 from [World Bank] (http://data.worldbank.org/indicator/NY.GDP.MKTP.CD). Call this Data Frame **GDP**. Make sure to skip the header, and rename the following list of countries: **[2 Marks]**
"Korea, Rep." to "South Korea",
"Iran, Islamic Rep." to "Iran",
"Hong Kong SAR, China" to "Hong Kong"

e) Finally, load the [Sciamgo Journal and Country Rank data for Energy Engineering and Power Technology] (http://www.scimagojr.com/countryrank.php?category=2102) from the file **scimagojr-3.xlsx**, which ranks countries based on their journal contributions in the aforementioned area. Call this Data Frame **ScimEn**. **[1 Mark]**

f) Join the three datasets: **GDP, Energy, and ScimEn** into a new dataset (using the intersection of country names). Use only the last 10 years (2006-2015) of GDP data and only the top 15 countries by Scimagojr 'Rank' (Rank 1 through 15). The index of this Data Frame should be the name of the country, and the columns should be ['Rank', 'Documents', 'Citable documents', 'Citations', 'Self-citations', 'Citations per document', 'H index', 'Energy Supply', 'Energy Supply per Capita', '% Renewable', '2006', '2007', '2008', '2009', '2010', '2011', '2012', '2013', '2014', '2015']. You should finally get a Data Frame with 20 columns and 15 entries. **[2 Marks]**

[**For Q2 to Q12, use the joined dataset you created in task 1(f)** ]

Q2) [Moderate] What are the top 15 countries for average GDP over the last 10 years? [NB: This function should return a Series named ‘avgGDP’ with 15 countries and their average GDP sorted in descending order.] **[1 Mark]**

Q3) [Moderate] By how much had the GDP changed over the 10 year span for the country with the 6th largest average GDP? [NB: This function should return a single number.] **[1 Mark]**

Q4) [Trivial] What is the mean energy supply per capita? [NB: This function should return a single number.] **[1 Mark]**

Q5) [Trivial] Which country has the maximum % Renewable and what is the percentage? [NB: This function should return a tuple with the name of the country and the percentage.] **[1 Mark]**

Q6) [Trivial] Create a new column that is the ratio of Self-Citations to Total Citations. What is the maximum value for this new column, and which country has the highest ratio? [NB: This function should return a tuple with the name of the country and the ratio.] **[1 Mark]**

Q7) [Moderate] Create a column that estimates the population using Energy Supply and Energy Supply per capita. What is the third most populous country according to this estimate? [NB: This function should return a single string value.] **[1 Mark]**

Q8) [Moderate] Create a column that estimates the number of citable documents per person. What is the correlation between the number of citable documents per capita and the energy supply per capita? Use the “.corr()” method, (Pearson's correlation). [NB: This function should return a single number.] Plot to visualize the relationship between Energy Supply per Capita vs. Citable docs per Capita. **[1 Mark]**

Q9) [Moderate] Create a new column with a 1 if the country's % Renewable value is at or above the median for all countries in the top 15, and a 0 if the country's % Renewable value is below the median. [NB: This function should return a series named “HighRenew” whose index is the country name sorted in ascending order of rank.] **[1 Mark]**

Q10) [Advanced] Use the following dictionary to group the Countries by Continent, then create a dataframe that displays the sample size (the number of countries in each continent bin), and the sum, mean, and std deviation for the estimated population of each continent. **[1 Mark]**

```
 ContinentDict  = {'China':'Asia',
                  'United States':'North America',
                  'Japan':'Asia',
                  'United Kingdom':'Europe',
                  'Russian Federation':'Europe',
                  'Canada':'North America',
                  'Germany':'Europe',
                  'India':'Asia',
                  'France':'Europe',
                  'South Korea':'Asia',
                  'Italy':'Europe',
                  'Spain':'Europe',
                  'Iran':'Asia',
                  'Australia':'Australia',
                  'Brazil':'South America'}
```

[NB: This function should return a DataFrame with index named Continent ['Asia', 'Australia', 'Europe', 'North America', 'South America'] and with columns ['size', 'sum', 'mean', 'std'].]

Q11) [Advanced] Cut % Renewable into 5 bins. Group Top15 by the Continent, as well as these new % Renewable bins. How many countries are in each of these groups? [NB: This function should return a Series with a MultiIndex of 'Continent', then the bins for '% Renewable'. Do not include groups with no countries.] **[1 Mark]**

Q12) [Moderate] Convert the Population Estimate series to a string with thousands separator (using commas). Do not round the results. **[1 Mark]**

e.g. 317615384.61538464 -> 317,615,384.61538464

[NB: This function should return a Series PopEst whose index is the country name and whose values are the population estimate string.]