## \*\*Data Question 1: An Exploration of UN data\*\*

In this project, you'll be doing some exploratory analysis on two country-level metrics, gross domestic product (GDP) per capita and overall life expectancy. After completing the guided practice section, you will have a chance to find some additional data and do some more exploring of your own.

### Guided Practice:

1. Download the Gross Domestic Product (GDP) per capita dataset from [http://data.un.org/Data.aspx?d=WDI&f=Indicator\_Code%3aNY.GDP.PCAP.PP.KD](http://data.un.org/Data.aspx?d=WDI&f=Indicator\_Code%3aNY.GDP.PCAP.PP.KD). Rename it to gdp\_per\_capita.csv and place it in the `data` folder of your project repository.

2. Create a Jupyter Notebook in the `notebooks` folder and name it `UN\_Data\_Exploration`.

\* You are likely to get errors along the way. When you do, read the errors to try to understand what is happening and how to correct it.

\* Use markdown cells to record your answers to any questions asked in this exercise. On the menu bar, you can toggle the cell type from 'Code' to 'Markdown'. [Here](https://www.markdownguide.org/cheat-sheet/) is a link to a cheat sheet showing the basics of styling text using Markdown.

3. In the first cell of your notebook, import the required packages with their customary aliases as follows:

`import pandas as pd`

`import numpy as np`

`import matplotlib.pyplot as plt`

`import seaborn as sns`

Keep all imports in this cell at the top of your notebook.

4. Using the pandas `read\_csv()` function, read the GDP dataset into your notebook as a DataFrame called `gdp\_df`.

\* Take a look at the first 10 rows.

\* Look at the last 5 rows. Do you see a problem?

\* Redo the read\_csv() call to correct this issue - \*\*do not\*\* modify the original csv file.

5. Drop the 'Value Footnotes' column, and rename the remaining columns to 'Country', 'Year', and 'GDP\_Per\_Capita'.

6. How many rows and columns does gdp\_df have? What are the data types of its columns? If any of the columns are not the expected types, figure out why and fix it.

7. Which years are represented in this dataset? Take a look at the number of observations per year. What do you notice?

It increased every year until 2013 stated contact till 2017and dropped after that

8. How many countries are represented in this dataset? Which countries are least represented in the dataset? Why do you think these countries have so few observations?

9. Create a new dataframe by subsetting `gdp\_df` to just the year 2014. Call this new dataframe `gdp\_2014`.

10. Use `.describe()` to find the summary statistics for GDP per capita in 2014.

11. Create a histogram of GDP Per Capita numbers for 2014 (you may wish to adjust the number of bins for your histogram). How would you describe the shape of the distribution?

12. Find the top 5 counties and bottom 5 countries by GDP per capita in 2014.

13. Now, return to the full dataset, `gdp\_df`. Pivot the data for 1990 and 2017 (using the pandas `.pivot\_table()` method or another method) so that each row corresponds to a country, each column corresponds to a year, and the values in the table give the GDP\_Per\_Capita amount. Drop any rows that are missing values for either 1990 or 2017. Save the result to a dataframe named `gdp\_pivoted`.

14. Create a new column in `gdp\_pivoted` named `Percent\_Change`. This column should contain the percent change in GDP\_Per\_Capita from 1990 to 2017. Hint: Percent change is calculated as 100\*(New Value - Old Value) / Old Value.

15. How many countries experienced a negative percent change in GDP per capita from 1990 to 2017?

16. Which country had the highest % change in GDP per capita? Create a line plot showing this country's GDP per capita for all years from 1990 to 2017. Create another showing the country with the second highest % change in GDP. How do the trends in these countries compare?

\*\*Bonus:\*\* Put both line charts on the same plot.

17. Read in continents.csv contained in the `data` folder into a new dataframe called `continents`. We will be using this dataframe to add a new column to our dataset.

18. Merge gdp\_df and continents. Keep only the countries that appear in both data frames. Save the result back to gdp\_df.

19. Determine the number of countries per continent. Create a bar chart showing this.

20. Create a seaborn boxplot showing GDP per capita in 2014 split out by continent. What do you notice?

21. Download the full csv containing Life expectancy at birth, total (years) from [https://data.worldbank.org/indicator/SP.DYN.LE00.IN?name\_desc=false](https://data.worldbank.org/indicator/SP.DYN.LE00.IN?name\_desc=false). Read this data into a DataFrame named `life\_expectancy`.

22. Drop the Country Code, Indicator Name, and Indicator Code columns. Then use `.melt()` to convert your data from wide to long. That is, instead of having one row per country and multiple colums per year, we want to have multiple rows per country and a single column for year. After melting, rename the columns to `Country`, `Year`, and `Life\_Expectancy`.

23. What was the first country with a life expectancy to exceed 80?

24. Merge `gdp\_per\_capita` and `life\_expectancy`, keeping all countries and years that appear in both DataFrames. Save the result to a new DataFrame named `gdp\_le`. If you get any errors in doing this, read them carefully and correct them. Look at the first five rows of your new data frame to confirm it merged correctly. Also, check the last five rows to make sure the data is clean and as expected.

25. Create a new DataFrame, named `gdp\_le\_2019` by extracting data for the year 2019 from `gdp\_le`. How many countries have a life expectancy of at least 80 in 2019?

26. Find the countries that had the top 3 largest GDP per capita figures for 2019. Create a seaborn FacetGrid showing the change in life expectancy over time for these three countries. Each individual figure in the facet grid will represent a single country.

27. Create a scatter plot of Life Expectancy vs GDP per Capita for the year 2019. What do you notice?

28. Find the correlation between Life Expectancy and GDP per Capita for the year 2019. What is the meaning of this number?

29. Add a column to `gdp\_le\_2019` and calculate the logarithm of GDP per capita. Find the correlation between the log of GDP per capita and life expectancy. How does this compare to the calculation in the previous part? Look at a scatter plot to see if the result of this calculation makes sense.

### Solo Exploration and Presentation:

1. Choose and download another data set from the UN data [http://data.un.org/Explorer.aspx](http://data.un.org/Explorer.aspx) to explore. You may want to combine your new dataset with one or both of the datasets that you already worked with. Prepare a short (< 5 minute) presentation of your findings. Report any interesting correlations or trends that you find.

2. If time allows, check out the plotly library to add additional interativity to your plots. [https://plotly.com/python/plotly-express/](https://plotly.com/python/plotly-express/).