https://github.com/NewForce-Cohort-11-Data/gdp-and-internet-usage-Shambala-ThreeDogNight

**A Guided Exploration of UN Data (Gross Domestic Product and Internet Usage)**

Hint: Go to the documentation <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html> look at the nrows argument.

1. Create a data folder in your local project repository.
2. Download these two CSV files and place them in the data folder:

a. Gross Domestic Product (GDP) per capita <http://data.un.org/Data.aspx?d=WDI&f=Indicator_Code%3aNY.GDP.PCAP.PP.KD> **DO NOT APPLY ANY FILTERS**

* + rename the file to gdp\_percapita.csv
  + open it with a text editor (**not excel**) and take a look at the top and bottom, do you notice anything?

b. Percentage of Individuals using the Internet <http://data.un.org/Data.aspx?d=ITU&f=ind1Code%3aI99H> **DO NOT APPLY ANY FILTERS**

* + rename the file to internet\_use.csv
  + open it with a text editor (**not excel**) and take a look at the top and bottom, do you notice anything?

1. Create a notebooks folder and launch a Jupyter Notebook in this folder. Give it a meaningful name.

* **IMPORTANT:** 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.

1. 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

1. Using the pandas read\_csv() method, read the GDP dataset into your notebook as a DataFrame called gdp\_df. Take a look at the first few and last few rows to familiarize yourself with what is contained in this dataset.
2. How many rows and columns are in gdp\_df? What are the data types of each column?
3. Drop the Value Footnotes column and rename the remaining three to 'Country', 'Year', and 'GDP\_Per\_Capita'.
4. How many countries have data for all years? Which countries are missing many years of data? Look at the number of observations per year. What do you notice?
5. In this question, you're going to create some plots to show the distribution of GDP per capita for the year 2020. Go to the Python Graph Gallery (<https://www.python-graph-gallery.com/>) and look at the different types of plots under the Distribution section. Create a histogram, a density plot, a boxplot, and a violin plot. What do you notice when you look at these plots? How do the plots compare and what information can you get out of one type that you can't necessarily get out of the others?
6. What was the median GDP per capita value in 2020?
7. For this question, you're going to create some visualizations to compare GDP per capita values for the years 1990, 2000, 2010, and 2020. Start by subsetting your data to just these 4 years into a new DataFrame named gdp\_decades. Using this, create the following 4 plots:
   * A boxplot
   * A barplot (check out the Barplot with Seaborn section: <https://www.python-graph-gallery.com/barplot/#Seaborn>)
   * A scatterplot
   * A scatterplot with a trend line overlaid (see this regplot example: <https://www.python-graph-gallery.com/42-custom-linear-regression-fit-seaborn>)  
     Comment on what you observe has happened to GDP values over time and the relative strengths and weaknesses of each type of plot.
8. Which country was the first to have a GDP per capita greater than $100,000?
9. Which country had the highest GDP per capita in 2020? Create a plot showing how this country's GDP per capita has changed over the timespan of the dataset.
10. Which country had the lowest GDP per capita in 2020? Create a plot showing how this country's GDP per capita has changed over the timespan of the dataset. **Bonus question:** Is it true in general that coutries had a higher GDP per capita in 2020 than in 1990? Which countries had lower GDP per capita in 2020 than in 1990?
11. Read in the internet use dataset into a DataFrame named internet\_df. You will likely get errors when doing this. Check the arguments for the read\_csv function to find ones that can help correct the errors (<https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html>) Once you are able to read it in, take per a look at the top and bottom few rows to make sure that it has been read in correctly. Also, check the datatypes of the columns.
12. Drop the Value Footnotes column and rename the remaining three to 'Country', 'Year', and 'Internet\_Users\_Pct'.
13. Look at the number of observations in this dataset per year. What do you notice?
14. What is the first year to have a non-zero internet users percentage value?
15. How does the distribution of internet users percent differ for 2000 and 2014?
16. For how many countries was the percentage of internet users below 5% in 2014?
17. Merge the two DataFrames to one. Do this in a way that keeps **all rows** from each of the two DataFrames. Call the new DataFrame gdp\_and\_internet\_use. Look at the first and last few rows to confirm that it merged correctly.
18. Find the three countries with the highest internet users percentage in 2014. Use a seaborn FacetGrid (<https://seaborn.pydata.org/generated/seaborn.FacetGrid.html>) to compare how the GDP per capita has changed over time for these three countries. What do you notice?
19. Subset gdp\_and\_internet\_use to just the year 2014. Save this as a new dataframe named gdp\_and\_internet\_use\_2014.
20. Create a plot which compares Internet Users Percentage and GDP per Capita for the year 2014. What do you notice from this plot? If you see any unusual points, investigate them.
21. **Stretch Question:** Use the qcut function from pandas (<https://pandas.pydata.org/docs/reference/api/pandas.qcut.html>) to divide countries in gdp\_per\_capita\_2014 into three groups based on their GDP per capita values. Label these groups as "Low", "Medium", and "High". Put these labels in a new column, named "GDP\_group".
22. **Stretch Question:** How does the median internet users percentage compare for the three gdp groups?

**Bonus exercise:**

1. Download another data set from the UN data (<http://data.un.org/Explorer.aspx>) to merge with your data and explore.