Data Question 1: An Exploration of UN data

Guided Practice:

1. Download two CSV files and place them in the `data` folder of your local Data Question 1 repository:

    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_per_capita.csv
    - open it with a text editor (not Excel) and take a look

    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

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 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

    import scipy.stats as stats

    import statsmodels.api as sm
    ```

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

4. At the bottom of your imports cell, use the `%matplotlib inline` magic command so that your plots show in the notebook *without* having to call `plt.show()` every time.

5. 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.

6. How many rows and columns does gdp_df have? What are the data types of its columns?

7. Drop the ‘Value Footnotes’ column, and rename the remaining columns to ‘Country’, ‘Year’, and ‘GDP_Per_Capita’.

8. 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. However, we must first do a little preparation of the country names in `gdp_df` so that we can merge the two correctly. Use the `.replace` method (or another method if you prefer) to make the following changes:

    - Change “CÃ´te d’Ivoire” to “Ivory Coast”
    - Change “CuraÃ§ao” to “Curaçao”
    - Change “SÃ£o TomÃ© and Principe” to “Sao Tome and Principe”

- Change “Sint Maarten (Dutch part)” to “Sint Maarten”

9. Merge gdp_df and continents. Keep only the countries that appear in both data frames. Save the result back to gdp_df.

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

11. How many countries are represented in this dataset? What range of years are represented? Take a look at the number of observations per year. What do you notice? Which countries are least represented in the dataset?

12. Create a new dataframe by subsetting `gdp_df` to just the year 2014. Call this new dataframe `gdp_2014`.

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

14. Which country had the highest GDP per capita in 2014? Which had the lowest? Find the top 5 counties by GDP per capita in 2014.

15. Add a column to your 2014 dataframe called ‘GDP_Group’. In this column, assign each country to one of three categories: “Low” if the country’s GDP Per Capita is below the 25th percentile of GDP Per Capita, “Medium” if it is between the 25th and 75th percentile, and “High” if it is above the 75th percentile. How do the continents differ in terms of numbers of countries in each group?
    Bonus: To make your code reusable, you should avoid using hard-coded values. For this part, programatically extract the 25th and 75th percentiles.

16. Create a histogram of GDP Per Capita numbers for 2014 (you may wish to adjust the number of bins for your histogram). What can you say about the distribution of GDP per capita figures in 2014?

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

18. Pivot the data for 1990 and 2017 (using the pandas `.pivot()` method) and then use the pivoted data to calculate percentage change in GDP Per Capita from 1990 to 2017. Hint: you may want to reset the index on the data frame after the pivot. Drop any countries that are missing GDP numbers for at least one of these two years.

19. What percentage of countries or areas experienced a positive % change in GDP per capita? What percentage experienced a negative % change?

20. Which country had the highest % change in GDP per capita? Create a line plot showing this country’s GDP per capita 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.

21. Read in internet_use.csv into a DataFrame called `internet_df`. You will most likely get an error message when doing this - figure out what is going wrong and fix it. Take a look at the first and last five rows and make any corrections to your `read_csv()` call to fix this. Again, **do not** modify the original datasets.

22. How many rows and columns does this new dataset have? What are the types of its columns? Ensure that the `Value` column is a numeric data type.

23. Change the columns for the Internet Users data frame to ‘Country’, ‘Year’, and ‘Internet_Users_Pct’.

24. What is the first year that has a nonzero internet users percentage reported? What is the general trend in internet users percentage over the years. Pick a visualization that you think illustrates this trend well.

25. Merge `gdf_df` and `internet_df` (on Country and Year) into a single DataFrame named `gdp_and_internet_use`. Keep only countries and years that appear in both tables.
    Difficult Bonus: Do not attempt this part until you have completed all other portions of the data question. Some countries have slightly different names in the internet use and gdp dataframes. For example, the Central African Republic is “Central African Republic” in the gdp dataframe and “Central

African Rep.” in the internet use dataframe. Find as many instances like this as you can and resolve them so that when merging you keep the maximum number of countries possible.

26. 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.

27. Create a new DataFrame, named `gdp_and_internet_use_2014` by extracting data for the year 2014 from `gdp_and_internet_use`. What is the mean internet users percentage in 2014? How many countries have at least 90% internet users in 2014?

28. Find the countries that had the top 3 largest GDP per capita figures for 2014. Create a seaborn FacetGrid showing the change in internet user percentage over time for these five countries. Each individual figure in the facet grid will represent a single country. What trends do you notice?

29. Create a scatter plot of Internet Use vs GDP per Capita for the year 2014. What do you notice?

30. Find the correlation between GDP per Capita and Internet Use for the year 2014. What is the meaning of this number?

31. Using the statsmodels library, create an ordinary linear regression model with independent variable GDP per capita and dependent variable internet users percentage for the year 2014. Be sure to include an intercept term. Print the model summary. What Rˆ2 value do you get for this model? What is the meaning of this number?

32. Add a column to `gdp_and_internet_use_2014` and calculate the logarithm of GDP per capita. Find the correlation between the log of GDP per capita and internet users percentage. Run an ordinary linear regression with response variable internet users and explanatory variable the log of the GDP per capita, and view the model summary. What is the meaning of the coefficients you get? Which of the two models that you created appears to do a better job?

33. Do you have any concerns about precicting percentage of internet users using a linear regression model?

34. Filter the original dataset down to just the United States for all available years. Calculate correlation between internet use and gdp per capita. Is this meaningful or useful?

Solo Exploration and Presentation:

1. Choose and download another data set from the UN data (http://data.un.org/Explorer.aspx) to merge with your data and explore.
2. Prepare a short (< 5 minute) presentation of your findings. Report any interesting correlations you find. Include visualizations and consider adding interactivity with `ipywidgets`. This presentation can be done either in a Jupyter Notebook or using another presentation software, such as PowerPoint. (Check out Jupyter Slides if you have time. This allows you to turn your jupyter notebook into a slideshow: https://medium.com/learning-machine-learning/present-your-data-science-projects-with-jupyter-slides-75f20735eb0f)
3. If time allows, check out the plotly library to add additional interativity to your plots. https://plotly.com/python/plotly-express/).