Laboratory / assignment #6:
Exploratory Data Analysis with with Python

Programming in Science (420-SN1-RE)
Teacher: Tiago Bortoletto Vaz

Goals:

1. Learning and practicing basics of EDA in Python

Instructions:

This is lab must be done in pairs. Students will be evaluated partly during the lab session and partly based on the work they have submitted. During one of the next classes, students will be called up individually to do a demo about their solutions. Failure to do so will significantly impact your grade, so please be sure that you understand everything you write. Good luck!

Assessment:

Students should submit a single PDF file containing a header with the course name, date, teacher's name and student's name. This is your short report for this laboratory exercise and it should be organized and look professional. **One submission per group.** This document must include the following content:

1. The URL of your GitHub repository containing the dataset and your Python script. Your script must contain the code used to answer all the questions from all parts of this lab and be properly commented.
2. Your answers and insights for Part 4, including plots generated from your Python code, your findings and any difficulties you encountered, as well as anything else you found interesting during the process.

Part 1 - Reading and practicing (10 minutes, optional)

If you don't feel confident enough about the topic, spend 10 minutes reading the following tutorial before answering the questions:

https://seaborn.pydata.org/tutorial/relational

Part 2 - Setting up the environment

1. Create a new Git repository for this lab using GitHub Desktop, then push it to http://GitHub.com.

2. For this lab we'll be using the "World Demographic Indicators Extract". This dataset is an extract of 11 development indicators across 217 countries from World Bank. Add this dataset (provided along with the Lab instructions) to your Git repository.
3. Create a new Python script and save it to your Git repository. Add your name at the start of the file, within the triple quotation marks.

Part 3 – Understanding and preparing the data

1. Take some time to understand the "World Demographic Indicators Extract" dataset. For instance, be aware of numerical and non-numerical data available and start thinking about a few questions to be answered for this lab. Take note of at least 3 questions that you think you'd be able to answer using your computer programming skills.
2. Import the modules pandas (as pd) and seaborn (as sns) to your code. Load the "World Demographic Indicators Extract" using the read_csv() function from pandas. Run your code and verify if you have the dataset loaded through the Variable Explorer window.
3. Once loaded in a variable, your dataset is a **Python object**. It's an object of type **Pandas Dataframe**. This kind of object has many functions to be explored. For instance, use the function info() to list the size of the dataset and its data types. You can now know how many entries this dataset has, how many columns, and estimate if the dataset has too many empty (null) values. Answer the question: how many empty values for the column "Physicians" and "Population"?
4. Use the nunique() function to print the number of unique values for each column. This can give you an idea about how to better represent your data in your plots later in this lab.
5. Use the describe() function to print further information about your data. What exactly does the output of this function provide?
6. Gross National Income (GNI) is an absolute value. It's more appropriate to use GNI per capita to answer some questions. Add a new column called "GNI per capita" to your dataset with data containing GNI by population. Then use the function round() to round it to the nearest cent.
7. Using the value_counts() function, print the answer for the following questions:
   a) How many countries are there in each region?
   b) How many hight income economies are there?
8. Using the pd.crosstab() function, print the answer for the question "Where are the high income economies?". Per region, and including Yes and No.
9. You might need to filter some data from your dataset. Often it can be made without using a loop. For example, the syntax to get only rows with values > 10 in a given column would be:

```
filtered_data = data[data["Your Column"] > 10]
```

Can you tell me how many countries there are where women can expect to live for more than 80 years? And which countries those are?

Part 4 - Visualizing statistical relationships

1. Using the relplot() function from seaborn, generate a plot that helps you answer the following question: “Is there any association between GNI per capita and life expectancy?”
2. By adding a third “feature” to your plot using colors to represent it in order to answer the following question: “Does the association between GNI per capita and life expectancy vary by region?”
3. Generate a the plot from item 2, now using lines along with standard deviation.
4. Use the lmplot() function to generate a linear regression for the previous plot.
5. Use relplot() to explore relationships between female life expectancy and some of the other numerical features. Are these relationships similar for male life expectancy? Elaborate at least 5 more questions and generate one plot for each to help you answering them. Use “Faceting” feature from seaborn to visualize side by side results for male and female.
6. Using your new programming skills, answer the following questions:
   - **a)** Is there any association between Internet use and emissions per capita?
   - **b)** Which are the countries with high emissions? (> 0.03)
   - **c)** Is there much variation by region (with respect to high emissions vs Internet use)?
   - **d)** Do all high income economies have high emissions?