Assignment 1

- Answer **ALL** questions.
- Total marks: 30
- Data for Questions 1 & 2 are available from the course L@G website. Data for Question 3 can be loaded from the internet.
- Use Jupyter Notebook/Python to achieve the answers.
- Type up your answers in a Word file (you may copy & paste some of the Python outcomes to the file).
- Assignment submission: upload **two** files – [1] the Word file [2] the Jupyter Notebook/Python file (either pdf (preferred) or ipynb). 10 marks will be deducted if only one file is submitted.
- A title page is required for the Word file with the safe assignment check. 5 marks off without safe assignment check.
- Late submission without approval is subject to penalty (10% off per day).
- Due: 11:59pm, (Thursday) **20-August-2020**

Question 1 (9 marks)

Use dataset **stock_p2.csv** to answer this question. The dataset has three columns: date, coke, and pepsi, where coke and pepsi are daily stock prices (at close) of Coca-Cola Bottling Co. Consolidated and PepsiCo, Inc., respectively, from 1990-12-31 to 2020-06-30.

1. How many rows and columns are in the dataset? Set "Date" as the row index. Print out the last 5 observations in year 2019.
2. Find the dates that Coke & Pepsi stocks reach their highest price levels, respectively. Find the dates that the two companies have the same level of stock price.
3. Compute the “log-return” (as defined in Topic 4) of the two stocks. Make violinplots of the two returns (specify inner as quartile). Comment on the results.
4. Find mean, standard deviation, skewness, and kurtosis of Coke & Pepsi returns, respectively. Comment on the results.
5. Make a scatterplot of the two returns with a fitted line & compute the correlation coefficient of them. Comment on the results.
6. Consider three subperiods: [1] 1990.12.31 to 1999.12.31 [2] 2000.01.01 to 2009.12.31 [3] 2010.01.01 to 2020.06.30. Redo the previous question (part 5) for each of the three subperiods. Comment on the relationship of Coke & Pepsi stock returns over the three subperiods.

Question 2 (15 marks)

Use data in **w2000.csv** to answer this question. The dataset contains information of top 2000 wealthiest persons in the world (of 2018), including the following columns: position, name, age, country, gender, wealthSource, industry, and worth (in millions of USD).

1. How many countries have at least one wealthiest person in the dataset? How many have at least one wealthiest man? How many have at least one wealthiest woman?
2. Make a list of the top 10 wealthiest women and compute the total wealth of them.
3. What are the top 20 countries having the majority of wealthiest persons? Make a bar chart showing the number of wealthiest persons for each of these 20 countries.
4. Make two pie plots of total worth by each industry, one for Japan and another for South Korea. Compare the two plots.
5. Find the oldest wealthiest person of each industry in Australia. Find all wealthiest men in the world of the "finance & investments" industry who are over 90 years old.
6. Find two lists of wealthiest persons -- first list contains those associated with Google (in terms of wealth source) and second list contains those with Facebook. Compare the average age across the two lists.
7. Which industry (apart from Philanthropy/NGO) has the highest percentage of wealthiest women? Are there any countries in the dataset with all wealthiest persons being female?
8. Find the 10 most popular first names of wealthiest persons in the US. How many wealthiest men in the world are with first name 'Jack'?
9. Create a subset of the so-called BRICS countries (Brazil, Russia, India, China, South Africa). In total, how many wealthiest persons are from BRICS? What is the average age of them by each country?
10. Use the BRICS subset to make a boxplot of age by each industry, separated by gender with "hue" (in one graph). Comment on the plot.

Question 3 (6 marks)

Load the data from the following link

https://raw.githubusercontent.com/datasets/covid-19/master/data/countries-aggregated.csv

and use the data to answer this question. Note that data from this link have been applied in Part 1, Topic 3. **Use data up to the end of July (2020-07-31).**

1. Find countries with deaths exceeding 10000. Get the first date for each of countries with more than 500000 confirmed cases.
2. Compute the overall daily death rate (i.e. "Deaths" divided "Confirmed" across all countries in each day) and plot it in a line chart. Identify the period that the overall daily death rate is higher than 5%.
3. Make two rolling (moving-average) line plots over 7 days of new confirmed cases of Australia and Japan, respectively. Note: “new cases” is defined as the difference of confirmed cases over two consecutive days. Comment on the results.
4. Find countries in the world with at least one day having more than 10000 new cases.