FIT1043 Assignment 1: Description
Due date: Friday 4 Sept 2020 - 11:55 pm

Aim

This assignment aims to explore and visualise data using Python as a data science tool. It will test your ability to:

1. read data files **in Python** and extract related data from them
2. use various graphical and non-graphical tools for performing exploratory data analysis and visualisation
3. use basic tools for managing and processing data
4. communicate your findings in your report.

This is an individual assignment.

Data

COVID-19 is a respiratory illness caused by a new virus which has changed our lives significantly. We aim to explore two datasets which contain relevant information about the virus and see whether different decisions and features such as applying lockdown, or the GDP of a country had any effect on the spread of the Coronavirus or not.

To achieve the goal of this analysis, we need some information about the new/total confirmed cases and deaths due to coronavirus as well as GDP and the lockdown date of each of the countries to do our analysis.

We use the following two datasets in this assignment:

1. The Corona Virus dataset (Covid-data.csv) generally contains information about new cases, deaths and GDP of several countries. Although, there are many countries in the world, we filtered the information to look at only the following countries in order to keep the level of assignment as simple as possible for this assignment: **Australia, China, France, Iran, Italy, Spain, United Kingdom**, **United States.**
   Moreover, the data set has the following columns:
   - date
   - location
   - total_cases
   - new_cases
   - total_deaths
   - new_deaths
   - gdp_per_capita
   - population

2. The lockdown dataset (CountryLockdowndates.csv) which contains information about the lockdown date and the name of the country which applied the lockdown.

Hand-in Requirements

Please hand in three files including a **PDF file** containing your answer, a **CSV file** containing the cleansed data set and a **Jupyter notebook** file **(.ipynb)** containing your Python code to all the questions respectively. Please consider the following cases for your submission:

1. PDF file should contain:
   - Answers to the questions. Make sure to include screenshots/images of the graphs you generate in your report (You will need to use screen-capture functionality to create appropriate images). Moreover, please include your Python code, **not the screenshot of your codes**, to justify your answers to all the questions. The Turnitin would not be generated if you include a screenshot of your codes and you will lose **20% of the assignment mark** if you include a screenshot of the codes instead of writing/copying your codes.
     To generate a pdf report, you can use Word to write your report, but you need to convert it to PDF before your submission. Alternatively, an easier way is to generate a pdf version of your Juputer notebook by hitting Ctrl+P in the Jupyter notebook. This pdf file is a mandatory requirement to check the Turnitin by Monash University.

2. Ipynb file should contain:
   Your Python codes for this assignment. Please use the provided **template** under Assignment 1 resources on Moodle ('StudentID_FIT1043_Assignment1_Template.ipynb').

3. CSV file should contain:
   - The cleaned data that is exported at the end of Task 1, based on the specified requirements in Task1.

You will need to submit three **separate** files. **"Zip"**, **"rar"** or any other similar file compression format **is not acceptable** and will have a **penalty of 10%**.

You will be penalized by **<u>5%</u>** of the assignment mark (5% out of 10 marks) if you submit after the due date for every day that you are late. If you could not submit your assignment before the due date, please make sure to submit your files at most 7 days after the assignment due date, we do not mark assignments which will be submitted after 11th of September 11:55 pm.

Assignment Tasks:

There are two tasks that you need to complete for this assignment. You need to use Python to complete the tasks.

Task 1-Data wrangling

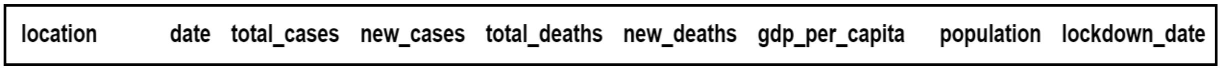
First, you need to extract the required information from two data sources, namely “Covid-data.csv”, and “CountryLockdowndates.csv” based on our analysis requirements mentioned in the previous section, ‘**Data**’. Then, you need to clean the data and integrate the data sets. We call this process as data wrangling! Please pay attention that you should not delete any row from dataset Covid-data.csv during the data wrangling process.

Regarding the cleansing of data set Covid-data.csv, you need to check all the columns one by one and make sure their values are correct. For example, we do not expect to see any value higher than 100% in a column which shows the percentage. Moreover, if there are some missing values, you would be able to find the correct values based on the value of other columns. This is an important part of data science and you need to make sure you check all the columns one by one, detect their errors and fix them.

Please pay attention that in lockdown information, you would see different dates for different states/provinces of a country. Consider the earliest(minimum) date as the lockdown date for a country.

You need to export the cleansed dataframe which is the result of this task, as a CSV file at the end of the task and submit it in Moodle with the other two files as required. Please name the dataframe as follows: <student_ID>_Task1DataSet.csv.

Following is a screenshot showing the columns of a dataframe which is the required output of this task (**Order of columns is important**).



location	date	total_cases	new_cases	total_deaths	new_deaths	gdp_per_capita	population	lockdown_date

Required column names and order for the CSV file which should be printed is as follows (location, date, total_cases, new_cases, total_deaths, new_deaths, gdp_per_capita, population, lockdown_date)

Task 2 Exploration

In this part, you need to explore the dataset which you generated in Task1. Please pay attention that exploration is not just a visualisation with a brief explanation. You can watch the assignment explanation recording provided in Moodle for further clarification about how a good exploration can be performed on a dataset.

1. Create a line chart to show the trend of the daily number of new cases for each country and **explore** the result of visualisation (**Create one line chart for each country**).
2. Add a vertical line for the lockdown date to the line chart of each country which you created in the previous question and **explore** if the lockdown affected the trend which is shown in the plot? Is the effect similar for all countries? Why do you think so?
   Following is an example of the expected plot for this question.

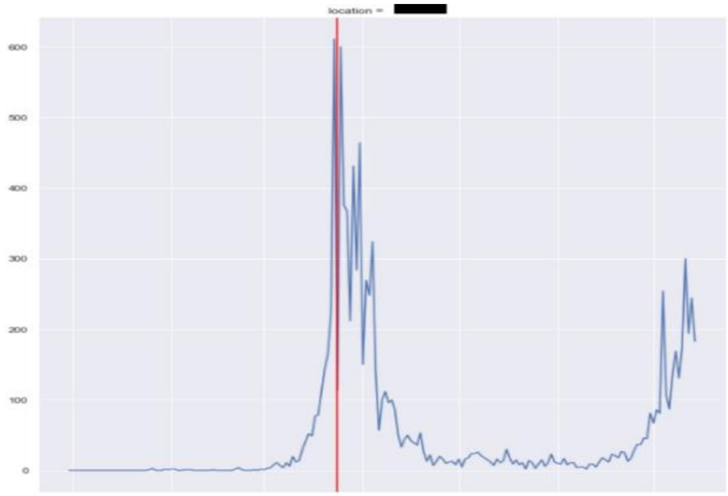


Figure 1. Example for the output of question 2 of Task 2

3. Explore whether there is a relation between daily new case/death rate and the GDP of a country. To this aim, you need to calculate:
   - The average of GDP of the countries, and then divide the countries into two groups, a group which its GDP is above the average GDP, and another group which its GDP is below the average GDP. We call the former group as "AboveGDP" and the later as "BelowGDP" from now onwards.
   - The daily new cases rate (new cases divided by population) for each country
   - The daily new death rate (new deaths divided by population) for each country

Then, you need to create two line charts, one which shows the new case rate of groups "AboveGDP" and "BelowGDP"; and, another line chart to show the death rate of the two groups ("AboveGDP" and "BelowGDP").

a) Which group ("AboveGDP" or "BelowGDP") usually had higher values of case rate?
b) Which group ("AboveGDP" or "BelowGDP”) usually had higher values of the death rate?
c) We would have expected that the case rate and death rate of group "AboveGDP" will be lower than group "BelowGDP". Is the result of your visualisation the same as the mentioned expectation? If no, why do you think the expectation is different from the reality?

GOOD LUCK!