FIT1043 Assignment 1 Specifications

Due date: Week 5, Friday 1st April 2022 - 6:00 pm (Malaysia Time)

Marks: 10%

Objective

The objective of this assignment is to investigate and visualise data using **Python** in the **Jupyter Notebook** environment. This assignment will test your ability to:

- Read data from files in Python,
- Manipulate the data,
- Describe the data using basic statistics,
- Produce non-graphical and graphical visualization to explore the data,
- Communicate your findings as insights, and
- Self-learn new techniques from other resources to complement what is taught in this unit.

Data

The data is presented in three comma-separated (CSV) files sourced from Kaggle, The World Bank, and The United Nations. The files should be obtained from Moodle and saved in the "data" folder (directory) where your Jupyter ipython notebook is. The data is:

- "`LifeExpectancyData-v2.csv`" contains information related to life expectancy, health factors for 193 countries have been collected from the same WHO data repository website, and its corresponding economic data was collected from the United Nation website (source: https://www.kaggle.com/kumarajarshi/life-expectancy-who). As part of the exercise, you can get the description of the fields (columns) on the Kaggle site.
- "`2019-GDP.csv`" is the recorded Gross Domestic Product (GDP) of almost all countries in the world for the year 2019. There are 4 columns in there but you will only need the last 2 columns which are the country name and the GDP stated in US Dollars. (source: https://datacatalog.worldbank.org/dataset/gdp-ranking).
- "`2020-Population.csv`" contains information about country and region population from 1950 to 2020. (source: https://population.un.org/wpp/Download/Standard/CSV/)

Most of the columns are self-explanatory but do participate in the Moodle forum to ask for clarifications or discussion on the data.

Note: For this assignment, **DO NOT** download the latest data from the sources. Because some of the columns have been removed, only utilize the provided data files.

Submission

This assignment has to be done using the **Jupyter Notebook only**. Your Jupyter Notebook has to use the Markdown language for proper formatting of the report and answers, with inline Python code and graphs.

You are to hand in **two** files:

1. The **Jupyter Notebook file (.ipynb)** that contains a working copy of your report (using Markdown) and Python code that answers the questions.
2. A **PDF** file that is generated from your Jupyter Notebook. Execute your Python code and then download it as a PDF document. To do so (in Windows), you can do a "Print Preview", then "Print" the document, and then select "Save as PDF". Note that there are other ways to do this, depending on the environment that you are in. Alternatively, you can download as HTML and then "Print" that to PDF. Again, participate in the Moodle forum if you need assistance on this.

Clarifications

This assignment is not intended to provide step-by-step directions, and I anticipate some clarification questions. The questions can range from "What is a PDF file?" to" to something relating to the possible meaning of a column in the CSV file.

I would like you to post these questions on the Moodle Forum and I strongly encourage interactions between all of you in the forum. Some of the questions probably don't have a single answer or a correct answer and is up to each individual's interpretation. Just make sure that you do not post answers in the forum.

Link to Moodle Forum (https://edstem.org/au/courses/8243/discussion/)

Assignment

This assignment is worth 20 marks, which makes up for 10% of this Unit's assessment. This assignment has to be done using the **Python programming** language in the **Jupyter Notebook environment**. It should also be formatted properly using the Markdown language. As an example, your Jupyter file should produce something like below (image taken from https://stackoverflow.com/questions/36288670/how-to-programmatically-generate-markdown-output-in-jupyter-notebooks). For each section, you are to write about your approach, then your code and the output (can be non-graphical or graphical).

Python Code in Markdown Cells

In [15]:
```
a = 2.123
```

The variable a is 2.123.

In [16]:
```
b=Latex(r'$b = \frac{\epsilon}{2}$')
```

You can also embed Latex: $b = \frac{\epsilon}{2}$ in here!

In [17]:
```
from IPython.display import Image, SVG
i = Image(filename='mouse-toy.jpg');
```

Even images can be embedded:

Tasks

You should start your assignment by providing the title of the assignment and unit code, your name and student ID, e.g.

FIT1043 Introduction to Data Science
Assignment 1

Sicily Ting
0123456789
1st April 2022

Introduction

You can write something about the assignment here and explain how you are going to approach the assignment. In some cases, there is a need to explain the flow of your submission. Note that wrong submission formats can result in totally 0 marks, or in some cases, a mark penalty. Also, if the writing flow is not easily understood by the grader, penalty may also be imposed. What is important here as a future Data Professional is communication skills.

Importing Libraries

The first step is to import the library **pandas**, which is an open source data analysis tool for the python programming language. The purpose of the importing this library is that we would like to use the data structure such as *DataFrame* and it's associated functions such as reading from CSV files and so on.

In [125]:
```
import pandas as pd
# You can import the matplotlib here as well but can also be done later.
```

The tasks will involve:

- Importing the necessary libraries,
    - ensure you explain each step (like the "hidden" example above)
- Read the files,
    - do not change the location of the intended files, i.e. they should be in a folder called "`data`"
    - make sure you show that you have read the data correctly
- Wrangle the data,
    - sub-setting the necessary data,
        - For the Life_expectancy related `DataFrame`, you are to keep the columns: `country`, `Status`, `max_life_expectancy`, `mean_BMI`, `mean_income_composition_of_resources`, `mean_schooling`, and `mean_life_expectancy` (aggregated from the respective columns).
    - proper renaming of the columns (and indexing),
    - this assignment only needs the South East Asian countries, including East Timor. A little bit of geography needed here and a bit of general knowledge as well. For this, you are expected to create a `list` or `tuple` or other data structure to store the names of the countries that is required (and explain why you selected the data structure).
- merge the files correctly,
- manage any data type issues or data issues,
- feature engineer (create) the column "perCapitaGDP", and
    - as a *guide*, your final DataFrame should have 11 rows (the countries) and 10 columns (`country`, `Status`, `max_life_expectancy`, `mean_BMI`, `mean_income_composition_of_resources`, `mean_schooling`, and `mean_life_expectancy`, `population`, `GDP`, `perCapitaGDP`). You may have extra fields, and that will be ok.
- provide some statistical description of the final data that you have.
    - Interpret the data that you have obtained using basic statistics.

You are then to select the appropriate plots (graphs) and provide some basic insights to the following **questions** (referred to **as Question 1, 2 & 3** in the rubrics):

1. Each country will be classified as developing or developed. With this in mind, how would you visualise the expected life expectancy for the South East Asian population for developed or developing countries? Give some kind of insight (although it may be straight forward and easily understood from the visualisation).

2. Create a bar graph for each country, with side-by-side bars for population, mean life expectancy, and adult mortality. There are two difficulties here: first, the default graph will be difficult to visualise due to big disparities in the numbers, and second, this information may not provide a decent visualisation. These two challenges need you to

figure out, create the necessary code adjustments for the visualisation, and explain why the data used for the graph may be misleading (some general knowledge / domain expertise required).

3. For the final question, you will probably need the non-aggregated data from "`LifeExpectancyData-v2.csv`". You are to extract the data that's related only to Singapore and then plot a line graph on the `Life expectancy` over time. Again, plot another line graph to visualise the `Adult mortality` and `infant deaths` over time. Explain in what circumstances would the first line graph be useful (if at all) and What effect will infant and adult mortality rates have on life expectancy?

All visualisations (graphs) must be suitably labelled and formatted. Clarifications will be required because everyone of you will approach it differently. As a result, make use of the Moodle Forum for this reason, as it will stimulate peer-to-peer learning and may assist others who are pursuing a similar approach.

There will be penalties for late submission (as per University policy), incorrect submission format and/or unreadable submissions.

As extras, you can answer the following (not graded)

- *Should a country having a lower life expectancy value(<65) increase its healthcare expenditure in order to improve its average lifespan?*
- *What is the impact of schooling on the lifespan of humans?*
- *Does Life Expectancy have positive or negative relationship with drinking alcohol?*
- *Do densely populated countries tend to have lower life expectancy?*
- *What is the impact of Immunization coverage on life Expectancy?*

You can probably try to answer many other questions just from these datasets, for those who are interested, you can discuss among yourselves and use Ed forum for your discussion.

Marking Rubrics (Guideline ONLY)

Report	Appropriately formatted using Markdown (and HTML) and content	1 mark - Using at least 3 formatting codes (Markdown or HTML) **2 marks** - Good and easy to read submission, including introduction and conclusion.
Code	Reading and describing the file content	**2 marks** – Importing libraries, reading files and showing that they are read correctly, and basic statistics of the values in the files.
	Wrangling, merging the files into one DataFrame	1 mark – Using a list/tuple/other data structure to store the SEA countries, and explaining the choice. 3 marks – Aggregating, sub-setting, renaming, re-indexing, type manipulation (type casting), and merging (some evidence of the use of any of them) **4 marks** – Feature engineered "per capita GDP", neat and no duplicated fields in final DataFrame.
	Question 1	1 mark – Appropriately explained choice of graphing. 2 marks – Code and graph (logical and executable) **3 marks** – Some insights (subjective)
	Question 2	2 marks – Code and graph (logical and executable) 3 marks – Explain the issue with the basic bar graph **5 marks** – Challenge to create a more appropriate bar graph and also explain why purely using this data may not be appropriate.
	Question 3	1 mark - Code and graph for life expectancy (logical and executable) 2 marks – Code and graph for adult mortality and infant deaths (logical and executable) **4 marks** – The explanation on why each graph may or may not be useful.

Have Fun!

After completing this project, you should have a solid understanding of Drew Conway's Venn Diagram. By completing this assignment, you will have demonstrated your "hacking skills" (via your Python code), you should have touched on some basic statistics (though you did not use them effectively for understanding Machine Learning), and hopefully, you will have persuaded that you have some domain knowledge (e.g., South East Asian countries and Life expectancy– useful to know if you don't already!).