FIT5145 Assignment 1: Description

Due date: Friday 6 September 2019- 11:55pm

The aim of this assignment is to investigate and visualise data using various data science tools. It will test your ability to:

1. read data files **in Python** and extract related data from those files;
2. wrangle and process data into the required formats;
3. use various graphical and non-graphical tools to performing exploratory data analysis and visualisation;
4. use basic tools for managing and processing big data; and
5. communicate your findings in your report.

You will need to submit two separate files (Important Note: Zip file submission will have a **penalty of 10%)**:

1. A **report in PDF** containing your answers to all the questions. Note that you can use Word or other word processing software to format your submission. Just save the final copy to a PDF before submitting. Make sure to **include screenshots/images** of the graphs you generate in order to justify your answers to all the questions. (Marks will be assigned to reports based on their correctness and clarity. -- For example, higher marks will be given to reports containing graphs with appropriately labelled axes.)
2. The **Python code** as a Jupyter notebook file that you wrote to analyse and plot the data.

Assignment Tasks:

There are three tasks (Tasks A, B and C) that you need to complete for this assignment. Students that complete **only the questions that are not labelled as "Challenge"** can only get a **maximum of Distinction**. Students that **attempt three questions labelled as "Challenge"** will be showing critical analysis skills and a deeper understanding of the task at hand and can achieve the **highest grade**. You need to use Python to complete the tasks.

Task A: Investigating Natural Increase in Australia's population

In this task, you are required to visualise the relationship between the births, deaths, total fertility rate (TFR), net overseas migration (NOM) and net interstate migration (NIM) for the different Australian states/territories, and gain insights on how these relations and trends change over time. The data files used in this task were originally downloaded from Australian Bureau of Statistics. We have extracted the data from the original files and transform them into a simpler format. Please download the data from Moodle:

- Births.csv - This file contains yearly data regarding the recorded number of births by Australian state/territory of registration between 1977 and 2016.
- Deaths.csv -This file contains yearly data regarding the recorded number of deaths by Australian state/territory of registration between 1977 and 2016.
- TFR.csv - This file contains yearly data on the recorded average number of births per woman over her lifetime by each state/territory between 1971 and 2016.
- NOM.csv - This data file contains yearly data on the net gain or loss of population through immigration (migrant arrivals) to Australia and emigration (migrant departures) from Australia, for the period between 1977 and 2016.
- NIM.csv - This data file contains yearly data on the net gain or loss of population through the movement of people from one state or territory to another, for the period between 1977 and 2016.

A1. Investigating the Births, Deaths and TFR Data

1. Using Python, plot the number of births recorded in each state/territory for different Australian states over different years.
   a. Describe the trend in number of births for Queensland and Tasmania for the period 1977 to 2016?
   b. Draw a bar chart to show the number of births in each Australian state in 2016.

2. We will now investigate the trend in the total number of births over different years. For this, you will need to aggregate the total number of births registered in Australia by year.
   a. Fit a linear regression using Python to the above aggregated data (i.e., total number of births registered in Australia over time) and plot the linear fit.
   b. Does it look like a good fit to you? Identify the period time having any unusual trend(s) in your plot.
   c. Use the linear fit to predict the total births in Australia for the years 2050 and 2100.
   d. Instead of fitting the linear regression to all of the data, try fitting it to just the most recent data points (say from 2010 onwards). How is the fit? Which model would give better predictions of future population of Australia do you think and why?
   e. **Challenge:** Can you think of a better model than linear regression to fit to all of the data to capture the trend in the number of births.
      i. Describe the model you suggested and explain why it is better suited for this task.
      ii. Use your model to predict the total births for the years 2050 and 2100.

3. Inspect the data on Total Fertility Rate (TFR.csv) for Queensland and Northern Territory.
   a. What was the minimum value for TFR recorded in the dataset for Queensland and when did that occur? What was the corresponding TFR value for Northern Territory in the same year?

4. Next, plot the natural growth in Australia's population over different years. For this, you will need to aggregate the total births and deaths by year. (HINT: Natural growth in a population is the difference between the total numbers of births and deaths in a population, for instance, Natural Growth of Australia’s Population = Total Births in Australia - Total Deaths in Australia)
   a. Describe the trend in ***natural*** growth in Australian population over time using linear regression?

A2. Investigating the Migration Data (NOM and NIM)

1. Let’s look at the Net Overseas Migration (NOM) data in different states over time.
   a. Use Python to plot the NOM to Victoria, Tasmania and Western Australia over time. Explain and compare the trend in all three states (VIC, TAS and WA).
   b. Plot the Net Overseas Migration (NOM) to Australia over time. Do you find the trend strange? Explain the reason to your answer (Hint: You might go online to find contributing factors to this trend).
2. Now let's look at the relationship between Net Overseas Migration (NOM) and Net Interstate Migration (NIM).
   a. Use Python to combine the data from the different files into a single table. The resulting table should contain the NOM and NIM values for each of the states for a given year. What are the first year and last year for the combined data?
   b. Now that you have the data combined, we can see whether there is a relationship between NOM and NIM. Plot the values against each other using scatter plot. Can you see any relationship between NOM and NIM?
   c. Try selecting and plotting the data for Victoria only using scatter plot. Can you see a relationship now? If so, explain the relationship.
   d. Finally, plot the Net Interstate Migration (NIM) for Queensland and New South Wales over different years. Note graphs for both QLD and NSW should be on the same plot. Compare the plots for these two states. What can you infer from the trend you see for these two states?

A3. Visualising the Relationship over Time

Now let's look at the relationship between other variables impacting the population size and growth of Australian states/territories over time. Ensure that you have combined all the data from the different files (Births.csv, Deaths.csv, TFR.csv, NOM.csv and NIM.csv) into a single table.

1. Use Python to build a Motion Chart, that compares the role migration (overseas and interstate) plays towards population growth in each Australia state/territory over time. The motion chart should show the Net Overseas Migration (NOM) on the x-axis, the Net Interstate Migration (NIM) on the y-axis, and the bubble size should show the ***Total Population Growth***. (HINT: A Jupyter notebook containing a tutorial on building motion charts in Python is available here)
2. Run the visualisation from start to end. (Hint: In Python, to speed up the animation, set timer bar next to the play/pause button to the minimum value.) And then answer the following questions:
   a. Comment generally on the trend you see in Net Overseas Migration (NOM) and Net Interstate Migration (NIM) overtime. Is there any relationship between the two variables?
   b. Select VIC and NSW for this question: In which year(s) does VIC have a higher Net Overseas Migration (NOM) than NSW. Please support your answer with a relevant python code and motion chart screenshot.
   c. Which state has the highest Net Interstate Migration most of the years (for the period 1977 to 2016)?

Task B: Exploratory Analysis of Data

In this task, you are required to explore the crime statistics dataset and do data auditing and exploration on the crime statistics dataset. The data we will use in this task contains Suburb-based crime statistics for crimes against the person and crimes against property in South Australia and comes from the South Australian Government. The dataset is publicly available from data.sa.gov.au on a yearly basis. Please download the data from Moodle:

- Crime_Statistics_SA_2014_2019.csv - The Crime statistics dataset contains all offences against the person and property that were reported to police between 2014 to 2019 in South Australian suburbs. The dataset contains information about the crime reported date, suburb incident occurred, Postcode, 3 levels of description of the offence, and the offence count.

Have a look at the CSV file (Crime_Statistics_SA_2014_2019.csv) and then answer a series of questions about the data using Python.

B1. Daily number of crimes

1. For each suburb, calculate the number of days that at least 15 crimes have occurred per day. (Hint: your answer should contain all suburbs in the dataset together with a value showing the number of days that at least 15 crimes have happened)
2. Now which suburbs do have at least one day where the daily number of crimes are more than 15. Plot the number of days that at least 15 crimes have occurred for the suburbs you found in this step (step 2) using a bar graph.
3. Use an appropriate graph to visualize and detect outliers (extreme values) on the data from step 2 and remove them. Then, plot the data again using a bar graph.
4. Compare the bar graphs in step 2 and 3. Which bar graph is easier to interpret? Why?

B2. Challenge: Identify mistakes in data entry

There are some errors in the data entry in one of the columns.

1. identify the data entry errors and provide possible solutions.
2. Use Python to fix the errors.
3. Argue how your answers to part B1 might be changed after fixing the errors.

Task C: Exploratory Analysis on Other Data

Challenge: Find some publicly available data and repeat some of the analysis performed in Tasks A and B above. As discussed in the lectures, there are many publicly available datasets online. For example, the Australian, US, UK, Singapore and Indian governments all provide websites with links to datasets:

- https://www.data.gov.au/
- https://www.data.gov/
- https://data.gov.uk/
- https://data.gov.sg/
- https://data.gov.in/

And Kaggle, a private company which runs data science competitions, also provide a list of their publicly available datasets:

- https://www.kaggle.com/datasets

Please note:

1. Your dataset(s) should contain at least 100 records.
2. Your dataset(s) should contain **time component** (e.g., year, day, etc.) in one of the columns.
3. Your analysis should at least contain **visualisation**, **interpretation** of your visualisation and a **prediction task**.
4. Please include a link to your dataset in your report. You may wish to
   a. provide the direct link to the public dataset from the internet, or
   b. place the data file in your Monash student - google drive and provide its link in the submission