FIT5145 Assignment 1: Description

Due date: Sunday 2 September 2018

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 files:

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.

Tasks:

There are two tasks that you need to complete for this assignment, plus a third (optional) task C for higher credit. Students that complete **only tasks A and B** can only get a **maximum of Distinction**. Students that **attempt task C** 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 Population and Gender Equality in Education

In the task, you are required to visualise the relationship between the population in different countries, the income in different countries and the gender ratio (women % men, 25 to 34 years) in schools of different countries, and gain insights from how these relations and trends change over time. The data files used in this task were originally downloaded from Gapminder. We have extracted the data from the original files and put into a simpler format. Please download the data from Moodle:

- **Population.csv**: This file contains yearly data regarding the estimated resident population, grouping by countries around the world, between 1800 and 2018.

- **GenderEquality.csv**: This data file contains yearly data about the ratio of female to male number of years in school, among 25- to 34-years-olds, including primary, secondary and tertiary education across different countries around the world, for the period between 1970 and 2015.
- **Income.csv**: This data file contains yearly data of income per person adjusted for differences in purchasing power (in international dollars) across different countries around the world, for the period between 1800 and 2018.

A1. Investigating the Population Data

Have a look at the resident population data. You will see many columns representing different countries.

1. In Python plot the population growth of Australia, China and United States over time. (HINT: Removed)
    - Are the population values increasing or decreasing over time?
2. Fit a linear regression using Python to the Chinese population data and plot the linear fit. (HINT: Removed)
    - Does the linear fit look good?
    - Use the linear fit to predict the resident population in China in 2020 and 2100.
    - Instead of fitting the linear regression to all of the data, try fitting it to just the most recent data points (say from 1960 onwards). How is the fit? Which model would give better predictions of future population in China do you think?

A2. Investigating the Gender Equality Data

Now have a look at the gender equality data.

1. Use Python to plot the gender ratio (women % men) in schools for Australia, China and United States over time.
    - What are the maximum and minimum values for gender ratio in Australia over the time period?
    - How do you compare the trend in gender ratio (women % men) in schools for these three countries over the time period? Which two countries have similar growth trend?
2. Fit a linear regression to the gender ratio in schools in United States and plot it.
    - Does it look like a good fit to you? Would you believe the predictions of the linear model going forward?

A3. Investigating the Income Data

Now have a look at the Income data.

1. Use Python to plot the Income of Australia, China and United States over time.

    - What was the minimum income in China recorded in the dataset and when did that occur? What was the income in Australia in the same year?

A4. Visualising the Relationship between Gender Equality and Population

Now let's look at the relationship between gender ratio in schools and the population.

1. Use Python to combine the data from the different files into a single table. The table should contain population values, income and gender ratio in schools for the different years and different countries.
    - What is the first year and last year for the combined data?
2. Now that you have the data aggregated, we can see whether there is a relationship between gender ratio in schools and the population. Plot the values against each other.
    - Can you see a relationship there?
3. Try selecting and plotting only the data from India.
    - Can you see a relationship now? If so, what relationship is there?

A5. Visualising the Relationship over Time

Now let's look at the relationship between gender ratio in schools and income over time.

1. Use Python to build a Motion Chart comparing the gender ratio in schools, the income, and the population of each country over time. The motion chart should show the gender ratio in schools on the x-axis, the income on the y-axis, and the bubble size should depend on the population. (HINT: A Jupyter notebook containing a tutorial on building motion charts in Python is available here.)
2. Run the visualisation from start to finish. (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:

    - Which two countries generally have the lowest gender ratio (women % men) in schools?
    - Which country has the highest gender ratio during the whole period of time?
    - Is the gender ratio generally increasing or decreasing during the whole period of time? How about income? Explain your answer.
    - Select Cape Verde and Bolivia for this question: From which year onwards does **Cape Verde** start to have a higher gender ratio and a higher income from **Bolivia**. Please support your answer with a relevant python code and motion chart.
    - Is there generally a relationship between the amount of income and gender ratio (women % men) in schools in all countries during the whole period of time? What kind of relationship? Explain your answer.
    - Any other interesting things you notice in the data? Please support your answer with relevant python code and/or motion chart

Task B: Exploratory Analysis on Big Data

In this part, you are required to do some exploratory analysis on the health insurance marketplace data. The file **InsuranceRates.csv.zip** contains data on health and dental plans offered to individuals and small businesses through the US Health Insurance Marketplace. This data was originally prepared and released by the Centers for Medicare & Medicaid Services (CMS). The data was then published on Kaggle. The file we provide is an extract from the data on Kaggle. Unzipped, the file is over 500MB and contains the following fields:

COLUMN	DESCRIPTION
BusinessYear	Year for which plan provides coverage to enrollees.
StateCode	Two-character state abbreviation indicating the state where the plan is offered
IssuerId	Five-digit numeric code that identifies the issuer organization in the Health Insurance Oversight System (HIOS)
PlanId	Fourteen-character alpha-numeric code that identifies an insurance plan within HIOS
Age	Categorical indicator of whether a subscriber's age is used to determine rate eligibility for the insurance plan.
IndividualRate	Dollar value for the monthly insurance premium cost applicable to a non-tobacco user for the insurance plan in a rating area, or to a general subscriber if there is no tobacco preference.
IndividualTobaccoRate	Dollar value for the monthly insurance premium cost applicable to a tobacco user for the insurance plan in a rating area

Load the InsuranceRates.csv data in Python and answer the following questions:

1. How many rows and columns are there?
2. How many years does the data cover? (Hint: pandas provides functionality to see 'unique' values.)
3. What are the possible values for 'Age'?
4. How many states are there?
5. How many insurance providers are there?
6. What are the average, maximum and minimum values for the monthly insurance premium cost for an individual? Do those values seem reasonable to you?

B2. Investigating Individual Insurance Costs

Now let's look more in detail at the individual insurance costs.

1. Show the distribution of 'IndividualRate' values using a histogram.
   - Does the distribution make sense to? What might be going on?
2. Remove rows with insurance premiums of 0 (or less) and over 2000. (**Use this data from now on.**) Generate a new histogram with a larger number of bins (say 200).
   - Does this data look more sensible?
   - Describe the data. How many groups can you see?

B3. Variation in Costs across States

How do insurance costs vary across states?

1. Generate a graph containing boxplots summarising the distribution of values for each state.
    - Which state has the lowest median insurance rates and which one has the highest? (Hint: you may need to rotate the state labels to be able to read the plot.)
2. Does the number of insurance issuers vary greatly across states?
    - Create a bar chart of the number of insurance companies in each state to see. (Hint: you will need to aggregate the data by state to do this.)
3. Could competition explain the difference in insurance premiums across states?
    - Use a scatterplot to plot the number of insurance issuers against the median insurance cost for each state.
    - Do you observe a relationship?

B4. Variation in Costs over Time and with Age

Generate boxplots (or other plots) of insurance costs versus year and age to answer the following questions:

1. Are insurance policies becoming cheaper or more expensive over time?
    - Is the median insurance cost increasing or decreasing?
2. How does insurance costs vary with the age of the person being insured? (Hint: filter out the value 'Family Option' before plotting the data.)
    - In terms of median cost, do older people pay more or less for insurance than younger people? How much more/less to they pay?

Task C: Exploratory Analysis on Other Data

(Note: This additional task is for those students wishing to get higher grades for their assessment. It is not required to pass the assignment, but it is required to get higher credit. Please refer to the marking rubrics for more details)

Find some publicly available data and repeat some of the analysis performed in Tasks A and B above. Good sources of data are government websites, such as: data.gov.au, data.gov, data.gov.in, data.gov.uk, ...

Please note that your analysis should at least contain **visualisation**, **interpretation** of your visualisation and a **prediction task**.