FIT5145 Assignment 1: Description

Due date Sunday 18 January 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, Linux shell and R, 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 perform exploratory data analysis and visualisation;
4. use basic tools for managing and processing big data; and
5. communicate your findings in your report.

Instructions:

1. There are two tasks that you need to complete using Python, and a third task using Linux Shell and R.
2. You will need to submit these files:

    i. A **report in PDF** containing your answers to all the questions, numbered correspondingly. 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 or 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.)

    ii. The **Python code** preferably as a Jupyter notebook file that you wrote to analyse and plot the data. Remember to include a hyperlink for the data source you used in Task B.

    iii. Copies of all the bash command lines and R scripts you use.

3. Name your files as Student_ID.pdf and Student_ID.ipynb, etc.
4. Late submissions will have a penalty of 5% per day, including weekends and public holidays for up to 7 days. Assessment items handed in after 7 days will not be considered.
5. Make sure to submit your assignments that are in draft mode (not submitted). We will not accept the assignments that are not yet submitted.
6. Acknowledgement of sources: Plagiarism or unauthorised collaboration will result in an automatic fail.

Task A: Investigating Population, Income and the CO2 Emissions (45%)

You need to use Python to complete the tasks. You may also do some pre-processing of the data in Excel if you wish (e.g. for Task A4).

In the task, you are required to visualise the relationship between the population in different countries, the income in different countries and the CO_2 emissions (tonnes per person) 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. 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.

- **co2_emissions_tonnes_per_person.csv**: This data file contains yearly data about the Carbon dioxide emissions from the burning of fossil fuels (metric tonnes of CO_2 per person) across different countries around the world, for the period between 1800* and 2014.

  *Some countries' data only start in recent years, e.g. the data for Timor-Leste only starts in year 2002.

- **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 (8%)

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.
   - 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.
   - 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 CO_2 emissions Data (8%)

Now have a look at the CO2 emissions data.

1. Use Python to plot the CO_2 emissions (tonnes per person) for Australia, China and United States over time.
   - How do you compare the trend in CO_2 emissions for these three countries over the time period? Are these countries having similar growth trend? What are the maximum values for CO_2 emissions (tonnes per person) for Australia, China and United States respectively?
2. Fit a linear regression to the CO_2 emissions 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 (3%)

Now have a look at the income data.

1. Use Python to plot the Income 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 CO_2 emissions, Income and Population (10%)

Now let's look at the relationship between CO_2 emissions, income 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 CO_2 emissions for the different years and different countries. (HINT: Pandas contains a "transpose()" method and Excel can also be used to transpose data.)
   - What is the first year and last year for the combined data?
2. Now that you have the data combined, we can see whether there is a relationship between CO_2 emissions and the population. Plot the values against each other.
   - Can you see a relationship there?
3. Try selecting and plotting only the data for Australia, China and United States.
   - Can you see a relationship now? If so, what relationship is there?

A5. Visualising the Relationship over Time (16%)

Now let's look at the relationship between CO_2 emissions and income over time.

1. Use Python to build a motion chart comparing the CO_2 emissions, the income, and the population of each country over time. The motion chart should show the CO_2 emissions 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.) Briefly describe any pattern on the data you observed.

   - Is there generally a relationship between the amount of income and CO_2 emissions in all countries during the whole period of time? What kind of relationship? Explain your answer.
   - Is the CO_2 emissions generally increasing or decreasing during the whole period of time? How about income? Explain your answer.
   - Which country generally have the lowest CO_2 emissions?

- Which country has the highest CO_2 emissions during the whole period of time?

3. Some countries' data only start in recent years, e.g. the data for Timor-Leste only starts in year 2002. Discuss how this might affect your observations.
4. Now filter the data, use data from year 1990 to 2014 and choose only the countries Australia, China and United States. Run the motion chart again and discuss your finding.
5. Any other interesting things you notice in the data? Support your answer with relevant python code and/or motion chart.

Task B: Exploratory Analysis on Other Data (25%)

You need to use Python to complete the task.

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

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

Include in your submission a hyperlink to the data source for Task B.

Task C: Exploratory Analysis on Twitter Data (30%)

NOTE: data set for this assignment is in the Google shared drive:

https://drive.google.com/drive/folders/1p8WgVBZZcU84r9FbKpLEcGQ_Y_REozTP

C1. Investigating the Twitter Data using the Linux Shell (18%)

Download the file Twitter_Data_1.gz from the link above. This file is large (1 GB), so it is better to do this at a Monash computer studio. Use a Linux shell to manipulate the file and answer the following questions.

1. Decompress the file. How big is it?
2. What delimiter is used to separate the columns in the file and how many columns are there?
3. The first column is a unique identifier for a Tweet. What are the other columns?
4. How many Tweets are there in the file?
5. What is the date range for Tweets in this file?
6. How many unique users are there? [Hint: It could take a few minutes to sort such a big list, so be patient!]1

[1] If you don't want to be patient, redirect the output of the command to a file and run the command "in the background" by typing an ampersand character "&" at the end.

7. When was the first mention in the file of ‘Donald Trump’ and what was the tweet?

For the following questions, assume ‘Obama’, ‘Trump’ in the tweet are referring to Barack Obama and Donald Trump respectively.

8. How many times has ‘Trump’ appear in tweets? How did you find this?
9. What about Obama? That is, how many times does the term ‘Obama’ appear in tweets?
10. Do you think we have captured all the references to Donald Trump and Barack Obama? What other strings might we need to try? What problems might we face?

C2. Graphing the Data in R (12%)

1) *Background:* We want to consider how the amount of discussion regarding Barack Obama varies over the time period covered by the data file. To answer this question you will need to extract the timestamps for all tweets referring to ‘Obama’. For this question, assume Obama in the tweet are referring to Barack Obama. You will then need to read them into R and generate a histogram. [Hint: To read the data into R, first generate a file containing only the timestamp column as text. Then read the file into R as a CSV.] R will not recognise the strings as timestamps automatically, so you’ll need to convert them from text values using the strptime() function. Instructions on how to use the function is available here: (https://stat.ethz.ch/R-manual/R-devel/library/base/html/strptime.html). *Question:* You will need to write a format string, starting with “%a %b“ to tell the function how to parse the particular date/time format in your file. What format string do you need to use?

2) Once you’ve converted the timestamps, use the hist() function to plot the data. [Hint: you will need to set the number of bins sufficiently high to see the variation over time well.]

3) Briefly describe any pattern pattern before Feb 15 and what happens after that?

4) Plot a second histogram showing the distribution of “number of authors” over “tweets count”.