FIT5145 Assignment 1: Description

Due date: Sunday 26th April 2020 - 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 (Note: Submitting a zipped file will attract **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 code, the output and any 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** is a Jupyter notebook file (*idnumber_FIT5145_A1.ipynb*) that you wrote to analyse and plot the data. (*Note that the entire assignment should be completed using python*)

Assignment Tasks:

The way we supply and use energy in Australia is changing. To understand these changes, to plan for Australia's energy future, and to make sound policy and investment decisions, we need timely, accurate, comprehensive and readily-accessible energy data. The Department of Industry, Science, Energy and Resources is responsible for compiling and publishing Australia's official energy statistics and balances[1]. The is updated annually and consists of historical energy consumption, production and trade statistics.

In this task, you are required to explore the statistics covering all electricity generation in Australia. This includes by power plants, and by businesses and households for their own use, in all states and territories. This also includes both on and off grid generation. We have extracted the data from the original files and restricted it to a specific time period. Please download the dataset for this assignment from the following link:

https://lms.monash.edu/mod/folder/view.php?id=6720926 ➤ *energy_data.xlsx*

[1] https://www.energy.gov.au/government-priorities/energy-data/australian-energy-statistics

The Data File

The data file you have downloaded is in xlsx format. Each sheet contains the energy generation statistics of each Australian State/Territory in GWh (**Gigawatt** hours) for the year 2009 to 2018.

Field description

- State: Names of different Australian states.
- Fuel_Type: The type of fuel which is consumed.
- Category: Determines whether a fuel is considered as a renewable or nonrenewable.
- Years: Years which the energy consumptions are recorded.

There are two tasks (Task A and B) that you need to complete for this assignment. You need to use Python v3.5+ to complete the tasks.

Task A: Exploratory Data Analysis of the Energy Dataset

This assessment aims to guide you in exploring the Australian Energy generation data set through the process of exploratory data analysis (EDA), primarily through visualisation of that data using various data science tools. You will need to draw on what you have learnt and will continue to learn, in class. You are also encouraged to seek out alternative information from reputable sources. If you use or are 'inspired' by any source code from one of these sources, you must reference this.

A1. Investigating the Energy Generation data for Victoria

1. First, read the data for Victoria state into a dataframe. You will observe that some values for the fuel types (eg. Black coal etc.) are missing or have 'Nan'. To handle it, replace these values with zero (using appropriate python code) before proceeding with the rest of the questions.
    a. Using Python, plot the total energy generation in Victoria over the time period covered in the dataset (2009 to 2018). Describe the trend you see in the overall energy generation for the given time period.
    b. Draw a new plot showing the trend in total renewable and non-renewable energy generation for the same time period? What trend can you observe from this graph?
    c. Draw a bar chart showing the breakdown of the different fuel types used for energy generation in 2009 vs in 2018? Explain your observation.
    d. What was the most used energy resource (fuel-type) in 2015? Which renewable fuel type was the least used in 2015?
    e. Draw a plot showing the percentage of Victoria's energy generation coming from Renewable vs Non-Renewable energy sources over the period 2009 to 2018. What can you say about the trend you observe?
    f. Using a linear regression model, predict what percentage of Victoria's energy generation will come from Renewable energy sources in the year 2030, 2100? Do the predictions seem reasonable?

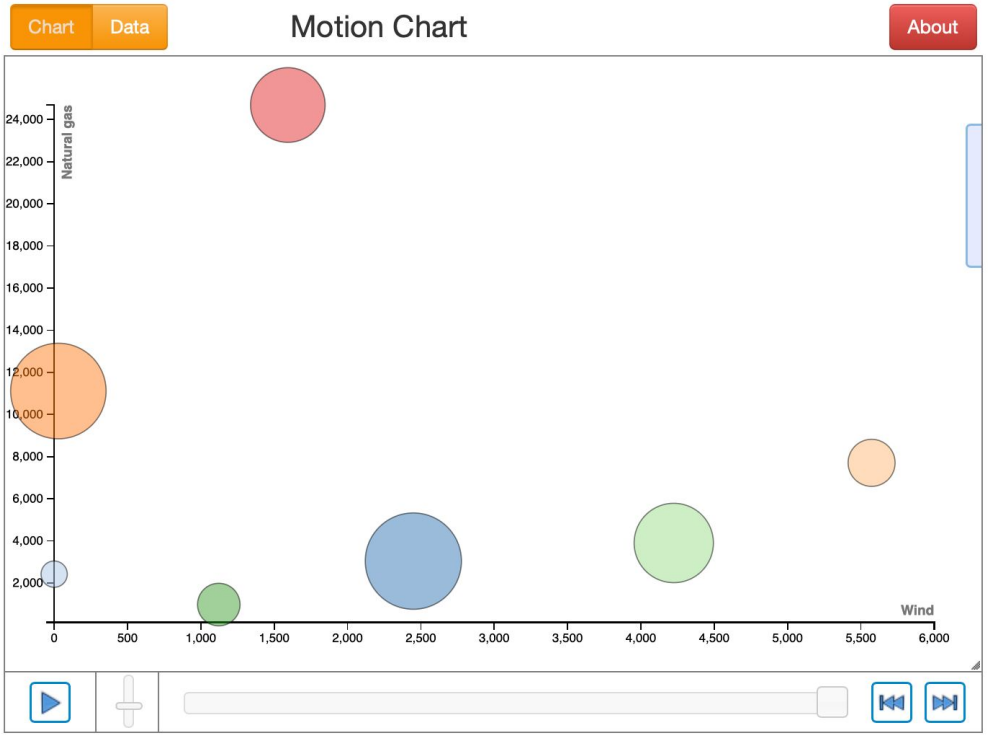
A2. Investigating the Energy Generation data for Australia.

1. Let's do some further investigation by combining the data for all the states and territories in Australia. Read the data for the rest of the states and merge them in a single dataframe. (Hint: you can use a combination of merge, melt or concat operators to get your data in a format suitable for answering the following questions)
    a. Plot a column chart showing the total energy generated in Australia by fuel type in the year 2018.
    b. Which state had the highest energy production in 2018? What is the ratio (percentage breakdown) of renewable vs non-renewable energy production for that state in 2018.
    c. Draw a plot showing the percentage of energy generation from renewable energy sources for each state over the period 2009 to 2018. From your graph, which state do you think is making the most progress towards adopting green energy? Provide a reason for your answer.

A3. Visualising the Relationship over Time

Now let's look at the relationship between all variables impacting the energy generation over time. Ensure that you have combined all the data from the different states. Ensure that your data is aggregated by year, state, the total energy produced (total_production), and has a separate column for each of the fuel types.

1. Use Python to build a Motion Chart, that visualises the energy production trend for Australia over time. The motion chart should show the units of energy production using *Wind* on the x-axis, the energy production using *Natural gas* on the y-axis, the colour represents the states/territories the bubble size should show the **total_production**. (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 the 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 on reliance on wind energy vs reliance on natural gas for each Australian state overtime. Is it logical to say if there is a relationship between the two variables?
    b. Which state relied most on natural gas for energy production in 2013? Please support your answer with any relevant python code and the motion chart screenshot.
    c. Comment on Queensland (QLD) states reliance trend on Natural gas between 2009 to 2018? What could be the reason contributing to this?



Sample snapshot of the expected Motion Chart

Task B: Exploratory Analysis of Data

In this task, you are presented with some pre-processed tweets about bushfires in Australia. The dataset is available via the following link:

https://lms.monash.edu/mod/folder/view.php?id=6720926 > twitter_data.csv

Please refer to Table 1 if you want to know the meaning of each feature/column. For example, nFollows shows the number of followers a user has. A user which has more than a thousand followers can be considered as a popular user. It should be noted that NOT every tweet in the data set is relevant to the bushfires in Australia, as represented by the value in the last column (1 denotes relevant and 0 irrelevant tweet).

Table 1: Description of Columns in the Data File

Column name	Description
text_score	Retrieval Score (based on some IR models)
text_score_expansion	Retrieval Score using expansion on topic
hashtag	If the tweet contains hashtag(s)
hasURL	If the tweet contains URL(s)
isReply	If the tweet is a reply to another tweet
length	The length (in characters) of the tweet
tweet_topic_time_diff	The time difference between the tweet and the query
semant ic_overlap	Overlap of named entities between the topic the tweet
#entityTypes	#types of Named-Entities (NE) extracted from tweet
#entities	#NEs extracted from tweet
organization_entities	#NEs with type of Orgranization extracted from tweet
person_entities	#NEs with type of Person extracted from tweet
work_entities	#NEs with type of Work extracted from tweet
event_entities	#NEs with type of Event extracted from tweet
species_entities	#NEs with type of Species extracted from tweet
places_entities	#NEs with type of Places extracted from tweet
nFollowers	#followers that the author has
n.Friends	#followees that the author has
nFavorties	How many times has the tweets been marked as favorite by others?
nListed	How many lists has the author been listed in?
isVerified	Whether the tweet is posted by a verified account or not?
isGeoEnabled	Is there a geolocation attached to this tweet?
twitterAge	How many years has been since the author signed up on Twitter?
#tweetsPosted	How many tweets has the author posted on Twitter?
relevanceJudge	Whether the tweet is relevant or not to the topic?

You are required to investigate the features of the twitter dataset. Please clearly label and comment your Python code used to answer each question.

B1. Investigating the Data

Please make sure to understand the data set and it's variables properly before answering the following questions. You need to have a good insight into the dataset to be able to understand some of the questions properly and avoid confusion.

1. How many tweets are there all together in the data file? How many of these tweets were posted from a verified account?
2. Draw a histogram showing the distribution of #entities extracted from the tweets. Set an appropriate bin size to present this information.
3. Compute the descriptive statistics (mean, std, quartile1, median, quartile3 and max) of #entities of relevant (ie. relevanceJudge = 1) and non-relevant (ie. with relevanceJudge = 0) tweets in the dataset. (Hint: You may use the describe() function for simplicity). Explain any interesting findings.
4. What is the average length of the tweets (in characters) that are judged as relevant? What is the average length of a non-relevant tweet?
5. To gain further insights into the twitter age of the users, it would be better to group the twitterAge in categorical bins. Create a new column twitter age group in your dataframe based on twitterAge by converting it into the following groupings or categories ['0-1','1-2','2-3','3-4', '4-5', '5+'] (***Hint:*** You can use the cut() method to bin (categorise) your data in these suggested categories)

    a. Generate boxplots summarising the distribution of each twitter age group against their median tweet length. What do you observe? Is there much variation in tweet length across the age groups?
    b. Which age group has the lowest median tweet length and which one has the highest? State these median values.
    c. According to the current bushfire tweet dataset, which age group is more active on twitter(has posted most tweets - from the current processed set tweets in your dataframe)? (Note: Each record in the dataframe is a tweet).
    d. Create a plot showing the total number of tweets posted by each age group (from Part [c] above).
    e. Which age group on average has the highest number of followers on twitter?

B2. Exploring correlation in the Data

In this task, you are required to explore the above (twitter) dataset and report on any interesting relationship/correlations you discover amongst the tweet variables. Your analysis should form a logical story. The answer should contain **visualisations** (plots to represent the trend or correlation), **interpretation** of your findings and an example of a **prediction** task (using simple linear regression).

[**Note:** *There should be a clear reason behind each visualisation you create, followed by a concise explanation of what message the visualisation is conveying.*]

All the Best !!!