FIT1043 Assignment 2: Description

Due date: Monday 26 September 2022- 11:55pm

Aim

The aim of this assignment is to investigate, visualise data and building machine learning models, **using Python** as a data science tool. It will test your ability to:

1. read a data file and extract related data from it;
2. use various graphical and non-graphical tools to perform data exploration, data wrangling and data analysis;
3. use basic tools for managing and processing data; and
4. communicate your findings in your report.

Hand-in Requirements

Please hand in the following two files. a **PDF file**[1] and a **Jupyter notebook file (.ipynb)**:

- PDF file should contain:
  1. Answers to the questions. In order to justify your answers to all the questions, make sure to
     a. Include **screenshots/images of the graphs or outputs** you generate (You will need to use screen-capture functionality to create appropriate images.)
     b. **Copy/paste of your Python code (not screenshots of your code**).

- Ipynb file should contain:
  1. **A copy of your working Python code** to answer the questions.

- You will need to submit two **separate** files (the PDF file and the ipynb file). **Zip**, **rar** or any other similar file compression format **are not acceptable** and will have a **penalty of 10%**.

Supportive Material:

- **Material:** In order to complete your assignment, you may want to use regressiondemo.py code used in week 5 tutorial. If you use this code, you do not need to upload the regressiondemo.py file in your final submission.

Assignment Tasks:

There are two tasks that you need to complete for this assignment. Students that complete **only Tasks A1-A8** can only get a **maximum of Distinction**. Students that **attempt tasks A9, B1 and B2** will be showing critical analysis skills and a deeper understanding of the task at hand and can achieve the **highest grade**.

[1] You can use Word or other word processing software to format your submission. Just save the final copy to a PDF before submitting or you can directly convert Ipynb file into pdf, if the Ipynb file includes everything required.

Data

We will explore two datasets in this assignment (Plus a dataset of your choice in Task B2):

1. **Australian Road Deaths (ARD) dataset:** you will work on this dataset in Task A.
2. **Song Popularity dataset:** you will work on this dataset in Task B1.

The Australian Road Deaths (ARD) dataset provides basic details of road transport crash fatalities in Australia as reported by the police each month to the State and Territory road safety authorities. It is published by the Bureau of Infrastructure, Transport and Regional Economics.

- The ARD dataset (Australian_Road_Deaths.csv file) contains basic demographic and crash details of people who have died in an Australian road crash.
- The data is collected from the year 2014 to 2021.
- The data is manipulated for assignment and is available on Moodle.
- The dataset contains the following information:

Column	Description
Crash ID	National crash identifying number
State	State of crash
YYYYMM	Year and Month of crash
Day of week	Day of week of crash
Time	Time of crash
Crash Type	The number of vehicles involved in the crash
Bus Involvement	Indicates involvement of a bus in the crash
Heavy Rigid Truck Involvement	Indicates involvement of a heavy rigid truck in the crash
Articulated Truck Involvement	Indicates involvement of an articulated truck in the crash
Road User	Road user type of a killed person
Gender	Gender of a killed person
Age	Age of a killed person (years)
Speed	Speed of a vehicle in the crash (km/h)
Driving Experience	Driving experience of a killed person (years)
National Remoteness Areas	ABS Remoteness Structure
SA4 Name 2016	Australian Statistical Geography Standard
National LGA Name 2017	Australian Statistical Geography Standard
National Road Type	National Road Type
Christmas Period	Indicates if crash occurred during the 12 days commencing on December 23rd
Easter Period	Indicates if crash occurred during the 5 days commencing on the Thursday before Good Friday
Age Group	Standard age groupings
Time of day	Indicates if crash occurred during the day or night

Task A: Data Wrangling and Analysis on ARD Dataset

In this task, you are required to explore the dataset and do some data analysis on the ARD dataset. Have a look at the csv file (Australian_Road_Deaths.csv) and then answer a series of questions about the data using Python.

A1. Dataset size

How many rows and columns exist in this dataset?

A2. The number of unique values in some columns

Count the number of unique values for National Remoteness Areas, SA4 Name 2016, National LGA Name 2017, and National Road Type in this dataset.

A3. Missing values and duplicates

There are some missing values: Unspecified, Undetermined, and blank (NaN) represent missing values.

1. How many rows contain missing values (Unspecified or Undetermined or blank) in this dataset?
2. List the months with no missing values in them.
3. Remove the records with missing values.
4. Remove duplicates as well after removing the missing values

Note: Use the dataset with missing values and duplicates removed from here onwards.

A4. Number of crashes in each month

List the number of crashes in each month. In which two months are the number of crashes at their largest?

A5. Investigating crashes over different months for specific road user

Now look at the Road User and YYYYMM columns and answer the following questions

1. Compute the average number of crashes against Month for car drivers. To do this,
    a. Extract Year and Month as separate columns
    b. Compute the number of crashes by both Year and Month for car drivers
    c. Based on task A5-1-b result, compute again the average number of crashes against Month. For each month, the average number of crashes is calculated over different years for which we have collected data for.
2. Draw a chart showing the average number of crashes over different months computed in task A5-1.
3. Discuss any interesting point in the chart.

A6. Exploring Speed, National Road Type, and Age

Now look at the Speed, National Road Type, and Age columns and answer the following questions

1. Draw a chart showing the average speed against National Road Type for car drivers
2. Due to **measurement error**, there are some counter-intuitive values in Age column. Identify those values and replace them with **zero**.

Note: Use the dataset from previous task (Task A6) and complete Tasks, A7-A9.

A7. Relationship between Age, Speed, and Driving Experiences

1. Compute pairwise correlation of columns, Age, Speed, and Driving Experiences for vehicle drivers (such as Motorcycle rider). Which two features have the highest linear association?
2. Now let's look at the relationship between the number of crashes and Driving Experiences. To do this, first compute the number of crashes against Driving Experiences for vehicle drivers and plot the values of these two features against each other. Is there any relationship between these two features? Describe it.

A8. Investigating yearly trend of crash

We will now investigate the trend in the crash over years. For this, you will need to compute the number of crashes by year.

1. Fit a linear regression using Python to this data (The number of crashes over different years) and plot the linear fit.
2. Use the linear fit to predict the number of crashes in 2022.
3. Can you think of a better model that well captures the trend of yearly crash? Develop a new model and explain why it is better suited for this task.
4. Use your new model to predict the number of crashes in 2022.

A9. Filling in missing values

Rather than replacing some counter-intuitive values with zero in task A6, use a better (e.g., model-based) approach to fill in the counter-intuitive values.

Task B: Decision Tree Classification on Song Popularity Dataset and K-means Clustering on Other Data

We have demonstrated decision tree classification and k-means clustering algorithm in week 7. Your task in this part is to apply decision tree on a Song Popularity Dataset provided on Moodle.

B1. Classification

We want to build a predictive model to predict song popularity for 5 popularity levels in the dataset based on features (You need to figure out which features to use). The song popularity column takes the following values: {1,2,3,4,5}, in which the higher the number the more popular the song is. 5 means very popular and 1 means less popular.

1. Divide the data set into a 75% training set and a 25% testing set using only the features relevant for classification.
2. Use feature scaling and train a decision tree model.
3. Using the test set, predict using the decision tree and compute the confusion matrix and the accuracy of classification.
4. Discuss your findings from the confusion matrix and accuracy. You should consider other performance metrics you learnt in lecture 7 to answer this question.

B2. Clustering

We have demonstrated k-means clustering algorithm in week 7. Your task in this part is to find an interesting dataset and apply k-means clustering on a dataset using Python. For instance, Kaggle is a private company which runs data science competitions and provides a list of their publicly available datasets:

https://www.kaggle.com/datasets

In particular you need to choose two numerical features in your dataset and apply k-means clustering on your data into k clusters in Python, where k>=2. Then visualise the data as well as the results of the k-means clustering, and describe your findings about the identified clusters. Ideally each cluster is shown in a different colour.

Please note you cannot use the same data set used in tutorials in this unit.

Please include a link to your dataset in your report. You may wish to:

1. provide the direct link to the public dataset from the internet, or
2. place the data file in your Monash student - google drive and provide its link in the submission.

Good Luck!