FIT1043 Assignment 1: Specification

Due date: Monday 28th August 2023- 11:55 pm

Aim

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

1. read a data file **in Python** and extract related data from it.
2. use various graphical and non-graphical tools for performing exploratory data analysis and visualisation.
3. use basic tools for managing and processing data and
4. communicate your findings in your report/video recording.

Data

The data we will use contains the number of monthly smartcard replacements by reason and type in Queensland and comes from the Queensland government open data initiative.

- The monthly smartcard replacements dataset (monthly_smartcard_replacements.csv) contains all recorded smartcard replacements in Queensland for different smartcard types and reasons each month.
- The information is given under variables; Month (including year and month), Transaction, Smartcard.Type, Action.Reason and Number.of.transactions.
- The file (monthly_smartcard_replacements.csv) is available on the unit Moodle site under Assessments.

Hand-in Requirements

Please hand in a **PDF file** containing your code, answers and explanations to questions and a **Jupyter notebook file (.ipynb)** containing your Python code to all the questions respectively:

- The PDF file should contain:
  - Answers and explanations to the questions. Make sure to include screenshots/images of the graphs you generate **and** your Python code to **justify your answers** for all the questions. (You may need to use screen-capture functionality to create appropriate images.) Please **do not** include screenshots of used code.
  - You can use Microsoft Word or other word processing software to format your submission. Alternatively, generate your PDF from your jupyter notebook formatted using markdown. Either way save the final copy to a PDF before submitting.

- The .ipynb file should contain:
  - **A copy of your work using python code** to answer all the questions.

- The video file should contain:
  - **A recording of yourself, explaining your answers to a subtask from Task A.**
  - You can use Zoom to prepare your recording.
  - Note each student is required to explain only one subtask from Task A. Please see Task B for more details.

You will need to submit **three separate** files (i.e., .pdf file, .ipynb file and your video file). Zip, rar or any other similar file compression format **is not acceptable** and will have a **penalty of 10%**.

Assignment Tasks:

Note: You need to use Python to complete all tasks.

Task A: Data Exploration and Visualisation

In this task, you are required to explore the monthly smartcard replacements dataset and perform analysis based on data subsets or groups with visualisations where required. Read the CSV file (monthly_smartcard_replacements.csv) in Python and then answer a series of questions about the data using Python.

A1. Exploring Smartcard Types

1. How many different (unique) smartcard types are recorded in the 'Smartcard.Type' column? What are those different smartcard types and how many instances are recorded for each type?
2. Plot a barchart of the smartcard types with the bars showing the count of each smartcard type. Which smartcard type is replaced the most?
3. Calculate the percentage of records for each smartcard type.

A2. Exploring Reasons for Smartcard Replacement

1. Convert data type of column 'Month' to a **datetime** format.
   Hint: Use pandas.to_datetime function to convert the type of 'Month' column to a datetime format as shown in one of your applied sessions.

2. What are the different reasons for smartcard replacements in the given data and how many instances are observed for each reason? **Hint:** Check the 'Action.Reason' column.

3. What is the total number of months in which 100 or more smartcard replacements are reported due to being "Faulty"?

A3. Investigating Annual Smartcard Replacements

1. Create a new column named 'Year' extracting the year from the 'Month' column.

Hint: you can extract year from column 'Month' using method **.dt.year** and create a new column for year as follows:

```
>>> your_dataframe['Year']=your_dataframe['Month'].dt.year
```

2. Create a line plot showing the total number of annual smartcard replacements (number of transactions) against year.
3. Explain the trend as observed from the chart. Are there any years that are different from others and if so, what is the reason behind it?

A4. Investigating Reasons for Smartcard Replacement

1. Plot a barchart to display the total number of transactions for each 'Action.Reason' using the available data.
2. What are the top three reasons for smartcard replacement?
3. Total number of transactions of which 'Action.Reason' is between 1000 and 2000?

A5. Investigating Reasons over Annual Smartcard Replacement

1. Find out the annual number of transactions for each 'Action.Reason' across different years for which data is available
2. For each action reason determine the number of years during which the number of annual transactions exceeds 10000.
3. Which action reasons have at least one year where the number of annual transactions exceeds 10000?
4. Create a histogram to analyse the distribution of the annual number of transactions per action reason as calculated in A5.1. Explain any observations and provide comments on the distribution.

Task B: Video Preparation

Presentation is one of the important steps in a data science process. In this task you will need to prepare a video of yourself (you can share your code on screen) and explain/present your answers to only one of the five subtasks in Task A (e.g., A3). In order to know which subtask you will need to prepare a video for, please take the last digit of your student ID (call it last_digit) and put it in the following formula to find the value of Question_number (do the calculations in python):

- Question_number = ⌈ (last_digit+1)/2 ⌉

  where ⌈⌉ is the ceiling function.

Then based on the Question_number you will get, you should explain only that subtask from Task A by recording a video of yourself explaining your answers to the questions in that subtask (e.g., how your Python code works, what are the inputs, outputs, what the graphs show etc).
For example if your student ID is 33333336, then take the last digit which is 6. Using the above formula, Question_number is equal to **4**. So you will need to prepare a video of Task **A4**.
Please make sure to keep your camera on (show yourself) during recording. You may want to share your screen with your code while you talk.

Good Luck! ☺