RMIT University
School of Computing Technologies

Practical Data Science with Python

Assignment 1: Data Cleaning and Summarising

Due: 23:59 on the 10th of April, 2022

This assignment is worth 25% of your overall mark.

Introduction

In this assignment, you will examine some data files and carry out the first several steps of the data science process, including the cleaning and exploring of data.

You will need to develop and implement appropriate steps, in IPython (Jupyter Notebook), to load a (some) data file(s) into memory, clean, process, and analyse it (them).

This assignment is intended to give you practical experience with the typical first several steps of the data science process.

The "Practical Data Science" Canvas contains further announcements and a discussion board for this assignment. Please be sure to check these on a regular basis – it is your responsibility to stay informed with regards to any announcements or changes. Login through `https://rmit.instructure.com/`.

Where to Develop Your Code

You are encouraged to develop and test your code in two environments: **Jupyter Notebook on Lab PCs** and **The Jupyter Notebook in Anaconda 3 environments that are suggested in Canvas Course Announcement**.

Plagiarism

RMIT University takes plagiarism very seriously. All assignments will be checked with plagiarism-detection software; any student found to have plagiarised will be subject to disciplinary action as described in the course guide. Plagiarism includes submitting code that is not your own or submitting text that is not your own. Allowing others to copy your work is also plagiarism. All plagiarism will be penalised; there are no exceptions and no excuses. For further information, please see the *Academic Integrity* information at `http://www1.rmit.edu.au/academicintegrity`.

Turnitin will be used for Plagiarism Review for this assignment in Canvas.

General Requirements

This section contains information about the general requirements that your assignment must meet. *Please read all requirements carefully before you start.*

- You *must* do the assignment in Jupiter Notebook that are available in Anaconda.
- Parts of this assignment will include a written report, this *must* be in *PDF* format.
- Please ensure that your submission follows the file naming rules specified in the tasks below. File names are case sensitive, i.e. if it is specified that the file name is `gryphon`, then that is exactly the file name you should submit; `Gryphon`, `GRYPHON`, `griffin`, and anything else but `gryphon` will be rejected.

Task 1: Data Preparation (10 marks)

The data you are required to prepare is statistical information about car buyers, which are in a set of CSV files (available in Canvas under the `Assignments/Assignment 1` section of the course Canvas.). Specifically, you will find several CSV files and each file stores the information of one attribute about car buyers. For example, 'Manufacturer.csv' stores the manufacturers of cars. Please note that there are two columns in each CSV file, and the first column represents the 'index' of observations and all information in CSV files with the same 'index' belongs to the same observation. You should study several of the input files in a text editor to get a feel for what you will need to process. You will quickly start to recognize the patterns in the structure of the data files so that you can quickly process all of them. The data is not clean, but it is well-formed. So this makes it very amenable to python data wrangling to extract and aggregate all of the data into a more usable form – a dataframe. pandas is a popular Python package that is used regularly in data science. So you will find that dataframes are a very useful way to organize and process columns of data similar to a database table.

First, you are required to extract and prepare the following information into one dataframe (namely, with each of the following information as one column in this dataframe):

- Manufacturer – Manufacturer.
- Model – Model.
- Price – Price.
- Transmission – Car Transmission.
- Power – Power (BHP) of the car.
- Engine CC – Engine size in CC.
- Fuel – Type of fuel.

error: fuel (peatrol, diasel, autometic)

- Male – Number of male owners.
- Female – Number of female owners.
- Unknown – Gender unclassified.
- Total – Total number of owners.

Moreover, being a careful data scientist, you know that it is vital to carefully check any available data before starting to analyse it. Your task is to prepare the provided data for analysis. Then, you need to clean the data by using the knowledge we taught in the lectures. You need to deal with all the potential issues/errors in the data appropriately and then write the cleaned data into a comma-separated values (csv) file named '**cleaned_car_buyers.csv**'.

Checkpoint (2 out 10 marks in Task 1): In week 5's Practical (Practical 4), you are required to demonstrate to your Lab Demonstrator that your assignment solution can generate this 'cleaned_car_buyers.csv' file. The Lab demonstrator will not assess how good is the generated file, and will only check whether your solution can generate the file or not.

Task 2: Data Exploration (8 marks)

Explore the provided data based on the following steps:

1. Explore the cars' total number of owners: Please analyze the composition of the total number of vehicle owners by gender for the top ten vehicles with the most owners.
2. Assuming that the data collector makes an entry error when collecting data, it can be ensured that the error occurred in the Price and Power columns, but it is not sure which car's information the error lies on. Please try to explore the error by visualization to identify how many errors there are and try to fix it.
3. Please analyze the relationship between the number of Male owners and the rest features (columns). Please use at least three other columns.

Note, each visualization (graph) should be complete and informative in itself, and should be clear for readers to read and obtain information.

Task 3: Report (7 marks)

Write your report and save it in a file called `report.pdf`, and it must be in PDF format, and must be **at most 6 (in single column format) pages (including everything, e.g. figures, references, appendix) with a font size between 10 and 12 points**. Penalties will apply if the report does not satisfy the requirement. Moreover, the quality of the report will be considered, e.g. clarity, grammar mistakes, the flow of the presentation.

Remember to clearly cite any sources (including books, research papers, course notes, etc.) that you referred to while designing aspects of your programs.

- Create a heading called "Data Preparation" in your report.
  - Provide a brief explanation of how you addressed the task. For the steps of dealing with the potential issues/errors, please create a sub-section for each type of errors you dealt with (e.g. typos, extra whitespaces, sanity checks for impossible values, and missing values etc), and also explain and justify how you dealt with each kind of errors.
- Create a heading called "Data Exploration" in your report.
  - For each numbered step in Task 2 above, create a sub-section with corresponding numbering.

What to Submit, When, and How

The assignment is due at

23:59 on the 10th of April, 2022.

Assignments submitted after this time will be subject to standard late submission penalties.

You need to submit the following files:

- The comma-separated values (csv) file - **cleaned_car_buyers.csv**.
- Notebook file containing your python commands for Task 1 and Task, 'assignment1.ipynb'. **Please use the provided solution template to organise your solutions**: *assignment1_TEMPLATE.ipynb*

For the notebook files, please make sure to clean them and remove any unnecessary lines of code (cells). Follow these steps before submission:

1. Main menu → Kernel → Restart & Run All
2. Wait till you see the output displayed properly. You should see all the data printed and graphs displayed.

- Your `report.pdf` file..

They must be submitted as ONE single zip file, named as your student number (for example, 1234567.zip if your student ID is s1234567). The zip file must be submitted in Canvas:

Assignments/Assignment 1.

Please do NOT submit other unnecessary files.