CITS4009 Project 1 — Exploratory Data Analysis

Worth 15% of the total assessment

Due: Friday, September 2nd, 2022, 11:59pm

1. Project Preamble

This is an individual effort project. In this project, you are to demonstrate your understanding of how exploratory data analysis is carried out using R functions and visualisation tools.

2. Data

To keep the unit relevant to real-world data analytics while fulfilling the educational goals, we would like to provide the following options for sourcing data for the project.

- A specified dataset
- Data from public repositories

Specified data

If you do not have any datasets or domain of interests in mind, we suggest you to use the **US Accident Injury** dataset, which can be obtained from *Data.gov*, published by *US Department of Labour*. The entire dataset spans across 15 years (2000 to 2015), and has a total of 202,814 observations, with a file size 140 MB. Even though it is not considered as big data, we suggest you to start with a smaller dataset to build your initial investigation with. So we provide you with a small dataset (2,000 observations). It is sufficient for this project to use the smaller dataset.

The download links to the US Accident Injury dataset (a .csv file) and the data dictionary (another .csv file) are provided on the *project* tab.

Data from public repositories

If you have specific domain of interests, for example, energy consumption, health sciences, transportation, sales, or sports, etc., you can browse some public dataset repositories to find a dataset that interests you. Note that the data needs to be in tabular form, i.e. not multi-media data, as we won't be able to deal with texts, images, videos, or audios. Also, the dataset should not be too simple. It should have continuous and discrete variables (columns) of various types (numerical, logical, character, etc). Your chosen dataset should have a similar level of complexity as the small (2,000 observations) US Accident Injury dataset.

A few well known public data repos are:

- **http://kaggle.com (http://kaggle.com)**
- **https://www.data.gov/ (https://www.data.gov/)**

- **https://www.data.gov.au/ (https://www.data.gov.au/)**
- **https://archive.ics.uci.edu/ml/datasets.html (https://archive.ics.uci.edu/ml/datasets.html)**

3. Exploratory Study of the Data

You are expected to produce an **HTML** file from your **R notebook** (a .Rmd file) that documents both the process and the R code that you used for your exploratory data analysis.

Using R functions to explore the data

Use R functions such as `str()`, `summary()` and `head()` to have a glance at the data. Document your interpretation of the data in the notebook.

Visualisation

Generate different types of plots and charts from the dataset for both single and multiple variable data exploration. Document any intuitions and observations you have in the notebook.

Data cleaning and transformation

Is there any missing values and data anomalies? Do you think it is important to do any data transformation? If so, document these in the notebook. Use visualisation to help justify the cleaning and transformation.

4. Marking Criteria

- **R functions (25%):** Correct use of built-in R functions, sensible interpretation of results, demonstrable proficiency in writing one's own functions.
- **Visualisation and Report Quality (30%):** At least 5 different types of plots are used correctly, with good justifications of the choice of *geoms* and annotations for meaningful readings of the data.
- **Data cleaning and transformation (35%):** Meaningful cleaning and transformation are performed on the data. Operations performed are well described and justified.
- **Process (10%):** Demonstrating a good understanding that EDA is an iterative process.

More details can be found in the **marking rubric** on the *project* tab.

Note: the quality of a data science project is not about just meeting these requirements, which are often considered as the bare minimum. One often has to go out of the way, demonstrating professionalism, proficiency, effort and thoroughness to obtain marks in the HD range (80%-100%).
Please see the model project from 2018 for a submission in the high HD range.

5. Submission

Generate an html file with the name **project1.html** from your notebook and submit it to **cssubmit (https://secure.csse.uwa.edu.au/run/cssubmit)**. Ensure that your submitted file can be opened and viewed in a web browser (e.g., firefox, chrome).

You should keep a record of your progressive work towards the final submission and also a copy of your latest work. You are encouraged to submit as often as you like to *cssubmit*. The latest submission will overwrite the previous version.

If you like version controlling your work, then you can keep your working copies on GitHub (https://github.com/) before the final submission to cssubmit; however, do make sure you keep your GitHub repo **private**.

Submission Check List:

- Your name and student number are visible at the beginning of your notebook. To make it easier for us in the marking process,
    - please ensure that you **have your student number written correctly**.
    - please ensure that you **use exactly the same name as shown on LMS**.
    - please put your **surname in uppercase letters**, e.g., Michael CHEN, John SMITH, Xiaolian HUANG.
- Submit only the generated **.html** file. No need to submit the original .Rmd file.
- If you are using data sourced from the web, make sure that you use the original URL in the data import function for your EDA.

6. Penalty on Late Submissions

See the URL below about late submissions of assignments:
https://ipoint.uwa.edu.au/app/answers/detail/a_id/2711/~/consequences-for-late-assignment-submission (https://ipoint.uwa.edu.au/app/answers/detail/a_id/2711/~/consequences-for-late-assignment-submission)

For example, if you get 70 marks (out of the total 100 marks) before applying the late penalty and if your submission is two days late, then you get 60 as your final mark (i.e., 10% of the total mark is deducted).