FIT5147 Project Proposal and Data Exploration Project

In this project, you are asked to analyse and explore data about a topic of your choice.

Please note that your project is subject to your tutor's approval. **Do not seek approval from the lecturers**.

It is an **individual assignment** and **worth 35%** of your total mark for FIT5147.

Relevant learning outcome

- Perform exploratory data analysis using a range of visualisation tools.

Overview of the tasks

1. Identify the project **topic**, **questions** that you want to address, and **data** source(s).
2. Submit **Project Proposal** in the Assessment block of Moodle by the end of Week 3.
3. Wait for approval before proceeding further. You will receive the feedback within Week 4.
4. Collect data and wrangle it into a suitable form for analysis using whatever tools you like.
5. Explore the data to answer your original question and/or to find something interesting using Tableau or R. The exploration should use appropriate visualisations and statistical analysis.
6. Submit a report detailing your findings and the method(s) that you used.
7. The **Data Exploration** is due on Monday of Week 6.

Project Proposal (2%)

Write a document consists of the following sections:

1. Project title.
2. Your identity (full name, student ID, tutor name).
3. 1-3 questions you wish to answer. The number of questions depends on the scope of the question itself. You can have one general question or three more detailed ones.
4. Data source(s) you plan to use to answer these questions, including a brief description of the data in each data source (kind of data: tabular, spatial, network, textual or other, number of records, URL).

Data Exploration (33%)

The report should contain the following structure:

1. *Introduction*
   Problem description, question and motivation.
2. *Data Wrangling*
   Description of the data sources with links if available, the steps in data wrangling (including data cleaning and data transformations), and tools that you used.
3. *Data Checking*
   Description of the data checking that you performed, errors that you found, your method to correct them, and tools that you used.
4. *Data Exploration*
   Description of the data exploration process with details of the statistical tests and visualisations you used, what you discovered, and tools that you used.
5. *Conclusion*

Summary of what you learned from the data and how your data exploration process answered (or didn't) your original questions.

6. *Reflection*
   Brief description of what you learned in this project and what in hindsight you might have done differently.
7. *Bibliography*
   Appropriate references and bibliography using APA or IEEE referencing system.

The written report should be **no more than 10 pages for all sections mentioned above**. Your written report will be the sole basis for judging the quality of the data checking, data wrangling and data exploration as well as the degree of difficulty. Thus, please include sufficient information in the report. It should, for instance, contain images of visualisations used for exploration and the results of any statistical analysis.

If you wish to provide further additional material an *Appendix* of up to 5 pages may be added at the end of the pdf document. However, the Appendix will not be graded. Therefore only use it to provide supplementary material that is not essential to the report or the readers understanding. Clearly title this section as Appendix.

Marking Rubric:

Project Proposal:

- *Project proposal* [2%]: Clear question(s) and identification of suitable data sources.

Data Exploration:

- *Data checking and wrangling* [5%]: appropriate checking, cleaning and reformatting, managing to get data into Tableau or R.
- *Data exploration* [10%]: completeness/thoroughness, use of appropriate visualisations and statistical measures, identification of trends or patterns etc and clearly articulated findings and limitations).
- *Degree of difficulty* [13%]: including, but not limited to, the use of non-tabular data, significant wrangling or cleaning required, large dataset, multiple data sets.
- *Written report* [5%]: quality of writing and use of images etc, logical structure, completeness.

Due dates:

- Submit the **PDF** version of the **Project Proposal** document to Moodle by **Friday, 21 August 2020, 5:00 PM**.
- Submit the **PDF** version of the **Data Exploration** document to Moodle by **Friday, 18 September 2020, 5:00 PM**.

Late submissions

- We encourage everyone to submit the proposal on time. We give **zero mark for late Project Proposal submission**. Everyone must submit the Project Proposal, even when the deadline has passed because your project must be approved before you can continue working on the Data Exploration.
- For Data Exploration, Assessments received after the submission deadline, or after the extended submission date for those with special consideration, will be **penalised at 5% of total mark [33%] per day for a maximum penalty period of ten (10) consecutive days.**
- If an **assessment is received after the penalty period**, then **zero marks** will be awarded.

- For further information on eligibility for **Special Consideration,** please refer to the relevant section on the Assessment page on Moodle.

Resubmissions

If you are retaking this unit from a previous semester, please ensure you choose a completely new topic and dataset.

Example of Project Proposal:

Project Proposal

Causes of serious bicycle accidents

Name : AAAAAA AAA

Student ID : 11111111

Tutor : TTT TTTTTT

Questions

1. What are the most common kinds of serious bicycle accidents?
2. How do lighting conditions affect these accidents?

Data sources:

a. ACT Road Cyclist Crashes, since 2012, which have been reported by the Police or the Public through the AFP Crash Report Form.
b. Canberra's sunrise and sunset times for 2018.

The data source a will allow me to answer question 1 at least for the ACT, while the combination of data source a and b will allow me to answer question 2.

Description of data sources:

1. Tabular data: 1K rows x 11 columns It has both spatial and temporal attributes as well as some simple text (https://www.data.act.gov.au/Justice-Safety-andEmergency/Cyclist-Crashes/n2kg-qkwj)
2. Tabular data in HTML: ~400 rows and 11 columns (http://members.iinet.net.au/~jacob/risesetcan.html)