Monash University

FIT5147 Data Exploration and Visualisation

Semester 1, 2023

FIT5147 Project Proposal and Data Exploration Project

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

Please note that your topic is subject to your tutor's approval.
Do not seek approval from the lecturer.

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, related questions that you want to address, and data source(s) that you will be using to answer those questions.
2. Submit your **Project Proposal** in the Assessment section of Moodle in **Week 3**.
3. Wait for approval before proceeding further. You will receive feedback in Week 4.
4. Collect data and wrangle it into a suitable form for analysis using whatever tools you like (e.g., Excel, R, Python).
5. Explore the data to answer your original question(s) and/or to find other interesting insights using **Tableau** or **R**. The exploration should rely on visualisations and visual analysis, but can include statistical analysis where appropriate.
6. Write a report detailing your findings and the method(s) that you used.
7. The **Data Exploration Project report** is due in **Week 8**.

Project Proposal (2%)

Write a **one page** document consists of the following sections:

1. Your full name, student ID, tutors' names, and tutorial number.
2. Project title.
3. Brief introduction of your topic and motivation (max 1-2 paragraphs).
4. Up to 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.
5. Data source(s) you plan to use to answer these questions, including a brief description of the data in each data source (e.g., number of rows; number of columns; type of data: tabular, spatial, network, textual or other; URL).
6. The bibliographical details of any references you have cited in the previous sections.

The topic and questions should allow for interesting and detailed analysis in the Data Exploration Project (DEP) and the subsequent Data Visualisation Project (DVP, due at the end of semester), which involves presenting the findings from your DEP in a narrative format.

It is strongly recommended that you **do not** include questions that are

- too easy to answer (e.g., what is the correlation between x and y, what is the average value of z variable, what are the top/bottom N values), or
- too difficult to do, or
- not relevant to the unit (e.g., training a machine learning model), or
- are not possible to find out from the data sources provided

Proposals with such questions will all be rejected (and not receive the *Suitability and Clarity* grade, according to the marking rubric). If in doubt, you should talk to your tutor during their consultation before the due date.

Good questions are general enough that they are not linked to specific parts of the data, allowing for more open-ended and exploratory analysis. For instance, asking "Where is the safest part of the network?" lets you explore various interpretations of how to link "where" and "safest" to the data about a network, whereas "Which LGA-code has the lowest value of number-of-deaths?" is very specific to the data and limits the exploration and visualisations possible.

The data must be accessible to the teaching staff, so the use of open data is encouraged (see the list of suggested data sources at the end of this document). Use of closed or proprietary data is allowed as long as explicit permission has been granted by the original source(s) for its use in the assignment. The data should include the last couple of years. Avoid common topics like COVID-19, unless you can think of some novel questions relating to it, perhaps by also using data on another related topic and from a different source. In such a case, you must discuss your intentions with your tutor before the due date.

Please ensure you read the rest of this entire document before deciding on your project, as the proposal is for the entire Data Exploration Project. See the end of this document for an example proposal and potential data sources that you may look at to get yourself started.

Data Exploration (33%)

The report should have the following structure:

1. *Introduction*
   Problem description, question, and motivation.
2. *Data Wrangling*
   Description of the data sources with direct urls to the data 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 and justification for how you corrected them, and the tools that you used. A comprehensive checking process is still expected, even if the data set is believed to be clean (i.e., to justify its correctness).
4. *Data Exploration*
   Description of the data exploration process with details of the visualisations and statistical tests (if applicable) you used, what you discovered, and what 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 you might have done differently in hindsight.
7. *Bibliography*
   Appropriate references and bibliography (this includes acknowledgements to online references or sources that have influenced your exploration) using either the APA and IEEE referencing system.

The written report should be **no more than 10 pages for all sections mentioned above**, excluding cover page, table of contents and appendix (see below). Your written report will be the sole basis for judging the quality of the data checking, data wrangling, 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. You should include any analysis that you carry out even if it is incomplete or inconclusive as it demonstrates that you have thoroughly explored the data set.

If you wish to provide additional material, an **Appendix** of up to 5 pages may be added at the end of the document. The Appendix will not be graded however. Therefore, you should only use it to provide supplementary material that is not essential to the report or the reader's understanding. Be sure to clearly title this section as Appendix.

Marking Rubric

Project Proposal (2%)

- ***Completeness and Timeliness*** [1%]: All components of the Proposal are included and it is submitted on time.
- ***Suitability and Clarity*** [1%]: Clear motivation, valid and suitable question(s) and data source(s).

You will be meeting with your tutor to discuss your Project Proposal and receive feedback during your Tutorial in Week 3. If your proposal is rejected, your tutor will specify the reason(s) they have done so and suggest areas for improvement.

Data Exploration (33%)

- ***Data Checking and Wrangling*** [5%]: appropriate checking, cleaning and reformatting; managing to get data into Tableau or R.
- ***Visualisation Design*** [5%]: visualisations that are appropriate for the intended purpose; readable and interpretable; appropriate labelling of axes; clear legends; saliency of patterns and trends.
- ***Analytical Methods and Interpretations*** [6%]: analysis that is appropriate for the intended purpose; justification and explanation of exploration process and use of statistical measures; identification of trends, patterns, and insights.
- ***Degree of Difficulty***:
    - ***Data Complexity*** [4%]: e.g., significant wrangling or cleaning required; good use of non-tabular data (e.g., spatial, relational, textual); large datasets (observations or dimensions) and/or multiple data sets; data scraping.

    - ***Advanced Analysis*** [2%]: e.g., clustering; dimensionality reduction; sophisticated aggregation and/or filtering; non-linear model fitting; correct use of statistical tests; complex timeseries analysis.
    - ***Visualisation Complexity*** [3%]: e.g., implementation difficulty; variety of good visualisations; attention to visual detail; complex visualisations.
    - ***Thoroughness of Interpretation*** [3%]: e.g., clearly articulated findings; awareness of limitations; deep exploration; thorough conclusions.
- ***Written Report*** [5%]: completeness; quality of writing and images; logical structure; correct referencing and use of figures and tables; correct academic referencing of sources.

Originality

As this is academic work, it must be original and must clearly indicate what elements were your work and what are based on someone-else's work. If you are including data, facts, opinions or any other written or graphical information from another source, you must cite the source and reference the bibliographic details for the source, using the APA or IEEE style guide. This includes any third-party programming code or software you use in your data exploration and analysis, and any definitions or descriptions of concepts or software. If you directly quote or replicate any material from a reference, you must do so in a manner appropriate to the APA or IEEE style guide.

If you are retaking this unit from a previous semester, please ensure you choose a completely new topic and dataset. The topic and dataset cannot have been used in any other unit. Likewise, you cannot reuse any code or written content that you have used in any previous assessment tasks for any units. The only self-plagiarism that is allowed is the questions you set in your proposal. The content of any previous or example assignments or sample report should not be reused.

Generative Artificial Intelligence (Generative AI) software or systems like ChatGPT or Midjourney cannot be used for any part of this assessment task, including (but not limited to) generating written or visual components of your submitted work.

If your work is believed to not be original, due to potential instances of plagiarism, collusion with other students or contract cheating, your academic integrity will be reviewed. If any breaches of the academic integrity are confirmed, penalties may be applied to your assignment, the unit and/or even your enrolment in your course.

Submission and due dates

- **Project Proposal**: Submit a one page **PDF.**
  Due Week 3. See Moodle for the date and time.
- **Data Exploration Project**: Submit a 10 page **PDF** (excluding cover page and appendix).
  Due Week 8. See Moodle for the date and time.

Late submissions

- There will be **zero marks for late Project Proposal submissions**. Everyone must submit the Project Proposal. Even if the deadline has passed, you must still submit a proposal (with a grade of 0) as your project **must be approved** before you can continue working on the Data Exploration Project.

- For the Data Exploration Project, submissions received after the deadline (or after an extended deadline for those with an extension/special consideration), will be **penalised at 10% of the total available mark [33%] per day up to a maximum of 7 days.** If submitted after 7 days, it will receive zero marks and no feedback will be provided.
- For further information on eligibility for **Extensions or Special Consideration,** see the Moodle unit page under Assessments > Extensions and Special Consideration

Example of Project Proposal

Project Proposal

Name: Jesse van Dijk, **Student ID:** 12345678, **Tutor:** Jo Bloggs and Alex Smith, Tutorial 0.

Project Title: Causes of serious bicycle accidents in Canberra

Introduction and Motivation

I am a keen cyclist and am concerned about cycling in Australia. I have recently moved to Canberra from the Netherlands where cycling is very safe and accidents linked with road vehicles is unusual. Recent media and industry reports indicate that Australian roads are becoming even more dangerous for cyclists [1,2] and I believe this is an important topic for many other audiences such as cyclists, road safety officers, and public health policy makers. Therefore I want to find out more about the factors that affect bicycle accidents in Canberra.

Questions

1. What are the most common kinds of serious bicycle accidents in Canberra, and how do these vary over the course of the day/week/year?
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 2021.

The data source A will allow me to answer question 1, while the combination of data sources A and B will allow me to answer question 2.

Description of data sources:

- 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)
- Tabular data in HTML: ~400 rows and 11 columns (http://members.iinet.net.au/~jacob/risesetcan.html)

References:

[1] Guthrie, Susannah (2020), *Report shows 'alarming spike' in cyclist deaths on Australian roads*, Car Advice, 05.08.2020. URL: https://www.caradvice.com.au/870483/cyclist-deaths-australia/ [Accessed on: 23.07.2021].

[2] Australian Automobile Association (2020), Benchmarking the Performance of the National Road Safety Strategy, July 2020, URL: https://www.aaa.asn.au/wp-content/uploads/2020/07/AAA_QBR_June_2020_Final_web.pdf [Accessed on: 23.07.2021].

Possible data sources (you are not limited to these)

The following is a list of data sources that you may take a look at to get started. Feel free to use these as sources of inspiration and ideas for your project. **You are not limited to what is listed here**.

- Data search tools and repositories, e.g.:
    - Google dataset search: https://toolbox.google.com/datasetsearch
    - Google Trends: https://www.google.com/trends/explore
    - Google Ngram Viewer: https://books.google.com/ngrams
    - Registry of Open Data on AWS: https://registry.opendata.aws/
    - Kaggle: https://www.kaggle.com
    - Science Hack Day: http://sciencehackday.pbworks.com/w/page/24500475/Datasets
- Open local and national government data portals, e.g.:
    - Victorian Government Data: http://data.vic.gov.au/
    - Australian Government Data: http://data.gov.au/
    - National Map: https://nationalmap.gov.au/ (Australian data)
    - Australian Bureau of Statistics: https://www.abs.gov.au/statistics
    - Atlas of Living Australia https://ala.org.au/
    - European Union Open Data: https://data.europa.eu/en
    - UK Government Open Data: https://data.gov.uk/
    - U.S. Government Open Data: https://www.data.gov/
- Humanitarian data sources, e.g.:
    - UNdata: http://data.un.org/
    - The World Bank Data Catalog: https://datacatalog.worldbank.org/
    - Our World in Data: https://ourworldindata.org/
    - Berkeley Library Health Statistics: http://guides.lib.berkeley.edu/publichealth/healthstatistics/rawdata
- Open corporate/industry data, e.g.:
    - Google Mobility Trends: https://www.google.com/covid19/mobility/
    - Apple Mobility Trends: https://covid19.apple.com/mobility
    - Uber: https://movement.uber.com/?lang=en-AU
    - Inside Airbnb: http://insideairbnb.com/get-the-data.html