Monash University
FIT5147 Data Exploration and Visualisation
Semester 1, 2025

Data Exploration Project

Part 1: Data Exploration Project Proposal
Part 2: Data Exploration Project Report

You are asked to explore and analyse data about a topic of your choice. It is an **individual assignment** and **worth 35%** of your total mark for FIT5147. Part 1 Project Proposal contributes 2% and Part 2 Project Report contributes 33%.

Relevant Learning Outcome

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

Overview of the Assessment Tasks

1. Identify the project topic, some related questions that you want to address, and the data source(s) that you will be using to answer those questions.
2. Submit your **Project Proposal** (Part 1) in the Assessments section of Moodle in **Week 3**.
3. Discuss with your tutor in your Week 3 Applied Session (*after* the submission in Moodle) and wait for approval from your tutor before proceeding further. Do not seek approval from the lecturer.
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 visually to answer your original questions and/or to find other interesting insights using **Tableau** or **R**. The exploration must rely on visualisations and visual analysis, but can analytical methods or statistical analysis where appropriate.
6. Write a report detailing your findings and the methods that you used. This must include properly captioned figures demonstrating your visual analysis (i.e. your visualisations must be referred to correctly in your report).
7. The **Project Report** (Part 2) is due in **Week 7**.

Read the rest of this document before deciding on your project topic, as the proposal is for the entire Data Exploration Project and Data Visualisation Project, which is the second major assignment of this unit. See the end of this document for an example proposal and potential data sources to get started. Be careful not to copy this proposal; it is an example proposal, not template text.

Choosing a Topic and Data

The choice of topic, data, and the questions you seek to answer 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 specifically designed narrative interactive visualisation format.

Good questions are general and 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?”is a good question that

lets you explore various interpretations of how to link terms like “where” and “safest” to the data about a network, whereas “Which region has the lowest value of number-of-deaths?” is not a very good question as it is very specific to the data, is easy to answer with one visualisation and therefore limits the exploration and visualisation possibilities.

It is strongly recommended that you avoid 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 answer (the work would take longer than the time available in the unit), or
- not relevant to the unit (e.g., training a machine learning model), or
- are not possible to answer from the available data.

Proposals with such questions will be rejected. If you are in doubt, talk to teaching staff during face-to-face teaching times or ask for confirmation on Ed.

How do you know if you have appropriate data? This depends on your topic and questions. You should ensure your data is big enough, i.e., has enough breadth and depth to invite interesting exploration. Combining data from different data sources is an ideal way to help add to the originality of the topic. To encourage different visualisation techniques your data will likely have a mixture of different data types. Time series (whether this be aggregated or detailed, such as months and years, or milliseconds) may be useful for your topic, and spatial, relational or text based data add useful complexity. If in doubt, talk to teaching staff during face-to-face teaching times or in a consultation before the due date.

The chosen topic should be topical and some of the data should be recently collected, ideally from the last two or three years. 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 for use in this assignment is granted by the original authors or copyright holders. If you have closed data, you must still make it available to your teaching staff to access, i.e., via a shared Google Drive.

Avoid common topics. Common topics including COVID-19, Netflix, AirBnB, car accidents, crime, house sales, car sales, world cup soccer, or electric vehicle sales should be avoided. Topics similar to the proposal example at the end of this document, i.e., traffic accidents and poor weather, must also be avoided. If you do have personal motivation for any of these mentioned common topics, you will need to propose a completely new angle to exploring the theme through novel questions with a mixture of new data sources. It is highly recommended to discuss your intentions with the tutor of your Applied Session prior to the proposal submission to avoid immediate rejection of the proposal.

Part 1: Project Proposal (2%)

Write a **one-page** PDF document consisting of the following sections:

1. **Project Title**
   A descriptive title for your project.
2. **Topic Introduction**
   One paragraph introducing the topic. This should include why it is a topical subject (for example, has it been in the news recently), and who might benefit from the insights you seek from your questions.

3. **Motivation**
   One paragraph describing why you personally are motivated to study this topic.
4. **Questions**
   Three questions you wish to answer using the data.
5. **Data source(s)**
   Briefly describe the data source(s) you will use. This should include: URLs of data source(s) and a description for each source: what is the data about, what is the size of the data (e.g., number of rows, number of columns), the type of data (e.g., tabular, spatial, relational, or textual), the type of attributes (e.g., categorical, ordinal, etc.) and the temporal intervals and period (e.g., monthly between 2019 and 2023).
6. **References**
   The bibliographical details of any references you have cited in the previous sections.

Include your full name, student ID, tutor names, and Applied Session class number. This can be in the document header or footer. There should be **no cover page.**

Part 2: Data Exploration (33%)

The report should have the following structure:

1. **Introduction**
   Topic detail, problem description, questions, and brief motivation.
2. **Data Wrangling and Checking**
   Description of the data and data sources with URLs of the data, the steps in data wrangling (including data cleaning and data transformations) and tools that you used. The data checking that you performed, errors that you found, your method and justification for how you corrected errors, and the tools that you used. A comprehensive checking process is expected to justify data correctness, even if the data set is believed to be clean.
3. **Data Exploration**
   Description of the data exploration process with details of the visualisations (including figures and descriptions of findings) and statistical tests (if applicable) you used, what you discovered, and what tools you used.
4. **Conclusion**
   Summary of what you learned from the data and how your data exploration process answered (or didn't answer) your original questions.
5. **Reflection**
   Brief description of what lessons you learnt in this project and what you might have done differently in hindsight.
6. **Bibliography**
   Appropriate references and bibliography (this includes acknowledgements to online references or sources that have influenced your exploration) using either the APA or IEEE referencing system.

Include your full name, student ID, tutor names, and Applied Session class number. This may be on a cover page, or in the header or footer of the first page.

The written report should be **not longer than 10 pages for all sections mentioned above**, excluding cover page, table of contents and appendix. 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, 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. However, the Appendix will not be marked. 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

Part 1: Project Proposal (2%)

- ***Completeness and Timeliness*** [1%]: All components of the Proposal are included and it is submitted on time.
- ***Suitability and Clarity*** [1%]: Motivation, Questions and Data Sources.

Motivation: A well-formulated project description with detailed information; a compelling and worthwhile topic to explore and visualise as a real-world problem.

Questions: Three well-crafted questions that can be clearly answered through data visualisations. Each question requires sophisticated analysis of relationships and patterns across multiple attributes and demonstrates potential for innovative visualisation approaches to reveal insights and complex patterns.

Data Sources: A clear description of data sources and datasets, including justification for which questions you will answer with each. The data must be sufficiently large or complex to require exploration and analysis. All datasets must be easily available, with URLs provided. For private and proprietary data, evidence of permission and a link to the dataset must be provided.

After submission you will meet with your tutor during the Week 3 Applied Session to discuss your Project Proposal, receive feedback and ideally approval to start. If your proposal is rejected, your tutor will specify the reasons and suggest areas for improvement. You will need to make these amendments to your proposal and get it approved by your tutor prior to commencing your project work.

Part 2: Project Report (33%)

Criteria	Below 50%	Pass (50%+)	Credit (60%+)	HD (80%+)
Data Complexity, Wrangling, Checking and Cleaning (7%)	Inappropriate checking, cleaning, or wrangling. 0 if no demonstration of data checking and cleaning.	Appropriate data cleaning and checking. Demonstrated ability to get data into R or Tableau;	Good choices and clear justifications for error checking, cleaning and transforming of non-tabular data (e.g. spatial, relational, textual); large datasets (observations or dimensions) and/or multiple data sets.	Excellence in data processing demonstrated and documented. Evidence of significant complexity in the wrangling, cleaning, transformation, or data collection (e.g. scrapping).
Data Visualisation and Design Choices (9%)	No visualisations; unsuitable or poor choice of visualisations; pixelated / poor quality images or illegible visualisations. 0 if not using Tableau or R.	Suitable visualisations, which are well presented, described, readable and interpretable.	Visualisations are appropriate for the intended purpose; appropriate labeling of axes and visualisations; clear legends when needed; saliency of patterns and trends.	Variety of high-quality, complex and/or creative visualisations with high attention to detail. Clearly justified design choices incl. visualisation idioms, choice of visual variables, layout and labelling.
Analytical Methods and Interpretations of Data and Topic Questions (9%)	Unsuitable analysis or misinterpretation of the data and topics questions. 0 if no data analysis is demonstrated.	Demonstrated suitable analysis and interpretation of the data and topic questions.	Analysis that is appropriate for the intended purpose; justification and explanation of the exploration process and use of statistical measures; identification of trends, patterns, and insights.	High quality of visual analysis demonstrated. Sophisticated and correctly used analytical methods such as clustering; dimensionality reduction; sophisticated aggregation and/or filtering; non-linear model fitting; correct use of statistical tests; or complex time series analysis.
Written Report: Quality and Completeness (8%)	Poor report, or missing sections.	Good report with logical structure with all the expected sections: Introduction, Data Wrangling, Data Checking, Data Exploration, Conclusion, Reflection, Bibliography. Referencing of sources, figures and tables. Correct grammar and spelling.	High quality of writing and figures/images with minimal errors. Correct referencing of figures and tables within the text, and correctly used academic referencing of sources.	Professional report with excellence of writing combined with high quality figures/images. Clearly articulated findings; awareness of limitations; deep exploration; thorough conclusions.

Originality

Since this is academic work, it must be original and clearly distinguish between your own contributions and those based on other's work. If you include data, facts, opinions or any other written or graphical information from another source, you must cite and reference it according to the APA or IEEE style guide. This includes third-party programming code, software used in data exploration and analysis, and any definitions or descriptions of concepts or software. Direct quotations or reproductions must adhere to the appropriate APA or IEEE style.

In your report you are encouraged to repeat the questions from your proposal. This is the only self-plagiarism that is allowed. If you are retaking this unit from a previous semester, you must choose a completely new topic and dataset. The topic and dataset cannot have been used in any other unit. You may not reuse any code or written content from previous assessment tasks for any unit. Additionally, content from previous assignments or sample reports cannot be used.

You may use Generative AI tools, such as ChatGPT, to improve writing and expression. However, your writing must be logically structured, clear and concise. Repetitive, poorly structured, or vague gibberish as often generated by Generative AI tools will result in a low grade. AI is generally unsuitable for data checking, cleaning, wrangling, exploration and visualisation of this level and should be avoided. It is important to remember that generated content can be biased. Any use of Generative AI in the preparation of your assessment must be acknowledged at the end of your submitted document.

If concerns arise regarding the originality of your work – whether due to plagiarism, collusion, contract cheating, or the use of unapproved software – your academic integrity will be reviewed. Confirmed breaches of academic integrity may result in penalties affecting your assignment mark, this unit, or even your enrolment.

Submission and Due Dates

Once you have completed your work, take the following steps to submit your work.

1. Save your proposal or report as a PDF document.
2. Name your file using the following structure**: Proposal_Surname_StudentID.pdf** or **DEP_Surname_StudentID.pdf**
3. Submit and upload your document.

- **Project Proposal**: Submit a one-page **PDF** in **Week 3**.
- **Project Report**: Submit a 10-page **PDF** (excluding cover page and appendix) in **Week 7**.
  See Moodle for dates and times.

Your assignment must show a status of "Submitted for grading" before it can be marked. Any submission in "Draft" mode will not be marked.

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. The proposal is a hurdle requirement. If it is not submitted and approved by your tutor, the mark for the Data Exploration Project is 0.

- For the Project Report, submissions received after the deadline (or after an extended deadline for those with an extension or special consideration) will be **penalised at 5% of the total available mark [33%] per calendar 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: https://www.monash.edu/students/admin/assessments/extensions-special-consideration

Example Data Sources

The following is a list of data sources to get started. Feel free to use these as a source of inspiration and ideas for your project. You are not limited to the data sources listed below.

- 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 **Note that using data from Kaggle exclusively is not acceptable, you must use at least one additional data source.**
  - 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.:
  - Uber: https://movement.uber.com/?lang=en-AU
  - Inside Airbnb: http://insideairbnb.com/get-the-data.html

Example Project Proposal

Please note this mock example is relatively old now. We expect your data to ideally include recent data, i.e., data from 2022, 2023 or even 2024. It is possible to complete this example project with only Data Source A and B, but C provides different opportunities and additional difficulty when doing the exploration and visualisations. If done well, this added depth and difficulty can gain extra marks but might take longer to complete. The student could use both datasets A and B to identify temporal aspects in the data, such as accidents near to sunset and sunrise across the whole dataset, but dataset C allows them to identify areas which are poorly lit and see if this correlates with the spatial pattern of pre-sunrise and post-sunset

accidents. Furthermore, whilst Data Sources A and C are currently tabular data, they can be converted to spatial features and spatial analysis can be carried out.

Name: Jesse van Dijk, **Student ID:** 12345678, **Teaching Associate:** Jo Bloggs & Alex Smith, Applied 01.

Project Title: Causes of Serious Bicycle Accidents in Canberra

Introduction

Recent media and industry reports indicate that Australian roads are becoming even more dangerous for cyclists [1,2]. I believe this is an important topic for many 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.

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. I have noticed it is difficult to see during sunset on a number of roads and would like to see if this pattern is evident in the data.

Questions

1. What are the most common kinds of serious bicycle accidents in Canberra, and how do these vary over different time periods (e.g. hour of day/day of week/month/season)?
2. How do lighting conditions affect these accidents?

Data sources

A. **ACT Road Cyclist Crashes** 2012 to 2021**,** which have been reported by the Police or the Public through the AFP Crash Report Form. This data is tabular data: ~1K rows × 11 columns. It has both spatial and temporal attributes including the geographical (latitude and longitude) location and a datetime stamp for the time of accident. Some numerical and simple text attributes relating to the incident. i.e. number of casualties, description of accident, including direction of traffic. (https://www.data.act.gov.au/Justice-Safety-andEmergency/Cyclist-Crashes/n2kg-qkwj)

B. **Canberra's sunrise and sunset times**, 2012 to 2021. Tabular data in HTML: ~365 rows × 4 columns for each year to be scrapped from sunrise website. Columns are simply date, time of sunrise, time of sunset and hours of daylight. https://sunrise.maplogs.com/canberra_act_australia.331491.html?year=2021

C. **ACT Streetlights, 2021**. Tabular data in CSV format with ~80K rows × 10 columns. These include latitude and longitude for the streetlight location and various text columns including lamp type, Luminaire, height and street and suburb name. There is no date column for the age of the lamp, but the source of the data is dated from 2017 and was last updated in Nov 2021. https://www.data.act.gov.au/Infrastructure-and-Utilities/ACT-Streetlights/cfpr-4tpw/about_data

Data Source A will be used to address Question 1, whilst A to C will allow me to answer Question 2.

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 (checked 25.02.2024)].

[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 (checked 25.02.2024)].