School of Computing and Information Systems
The University of Melbourne
COMP20008 - Elements of Data Processing, Semester 1, 2025

Assignment 2 – Road Crash and Injury Analysis

Release:	Monday 14 Apr 2025
Due:	• *Group contract*: Friday 2 May at 5 PM • *Code and Report submission*: Friday, 16 May at 5 PM • *Slides submission*: Friday, 23 May at 5 PM • *Oral presentation*: Week 12 (Monday 26 to Friday 30 May) • *Team-Evaluation:* Friday 30 May at 5 PM
Marks:	The Project will be marked out of 35 and will contribute 35% of your total mark[1].
Groups:	You should work in groups of 3 or 4 (Same groups as assignment 1)
Main Contact:	Hasti Samadi (hasti.samadi@unimelb.edu.au)

1. Overview

In this project, you will use the same dataset as Assignment 1. The goal is to uncover key patterns and risk factors contributing to road accidents and injury severity.

Through this project, you will:
- Develop a research question and investigate it
- Perform data and text processing to clean and structure the dataset.
- Conduct correlation analysis between weather, road conditions, and accident severity.
- Implement supervised learning models to predict accident severity and injury outcomes.
- Apply feature selection techniques to improve model performance.
- Use clustering methods to profile high-risk accident scenarios.

Your findings will be summarised in a technical report, supported by data visualisations and figures and code implementations.
You will present your report and analysis in an oral presentation and interview.

2. Assignment Structure

Group Contract – (Due: Friday 2 May at 5 PM) – Failure to submit leads to 2 marks penalty

You must submit a group contract outlining your team's goals, expectations, and policies for working on the project. A *group contract template* is provided. You are welcome to work with the provided template or customize it according to your preference. *Submit as a single PDF file via Canvas (Assignment 2: Group Contract).*

[1] As described in Week 1 lecture, each person's mark out of 35 will be moderated according to their individual bonus mark for workshop attendance.

You may vary your group contract throughout the semester, but proposed changes should be agreed to by all members. There are no marks directly allocated to the content of the Group Contract, but we may refer to it when assessing the relative contribution of each group member to resolve any dispute.

Code and Report Submission – 20 marks (Due: Friday, 16 May at 5 PM)

1. **Report**: Your report should consist of twelve to fifteen single-column A4 pages. Maintain a line spacing of exactly 1 with 2.54cm margins and ensure that the font size is 11pt or above. The page limit includes all the text, including references, captions, and any tables or images. All content should be readable if one was reading from a hard copy print out of the report.
   The group name W[XX]G[X] and all group members' names should appear on the first page after the title of the report. *Submit as a single PDF file through Canvas/Turnitin (Assignment 2: Group Report)*
2. **Code**: One or more programs written in Python, including all the code necessary to reproduce the results in your report (model implementation, data processing, visualisation, and evaluation). Your code should be executable and have sufficient comments to make it understandable. You should also include a README file that briefly details your implementation and describes how to run your code to reproduce the results in the report. *Submit as a single zip file through Canvas (Assignment 2: Code and Comments).*

Slides Submission (Due: Friday, 23 May at 5 PM)

You will need to submit the slides you are going to use for delivering your oral presentation. These slides should illustrate your insights derived from the data analysis task you've undertaken. *Submit as a single PowerPoint (.pptx) or PDF file through Canvas/Turnitin. (Assignment 2: Oral Presentation Slides). No other format is acceptable.*

In your presentation you will be required to use the exact slides that you have submitted. No updating allowed after submission.

Oral Presentation and assessment – 13 marks (Due: from Monday 26 May to Friday 30 May)

During week 12, all teams should deliver an oral presentation of their work and findings for assignment 2. This will happen at your usual workshop time. Some of the presentations will be conducted in the students' usual workshop room and some in other venues, which will be announced closer to the date. Two markers will assess the oral presentations. See section 6 for more details.

Teamwork evaluation – 2 marks (Due: Friday, 30 May at 5 PM)

For this part of the assessment, every team member needs to evaluate both their own contributions to the assignment and the contributions of their teammates. This evaluation should align with the expectations you set in your submitted "group contract".

The evaluation should be delivered via Feedback Fruits available on Canvas (Assignment 2: Teamwork Evaluation).

Your group members' evaluations will determine individual group member evaluation scores worth *two* marks. If any member is identified as a non-contributor, these scores may be used to adjust those individual's marks for the report component (worth 20 marks).

3. Data Sets

We are again using the same set of datasets that were used in assignment 1. The data is collected from Victoria Police Records including information regarding the road accidents. You will find the dataset overview report and all related data in the ZIP file provided on canvas.

4. Data Analysis Tasks

4.1. Research Question

The research question clarifies the purpose of your analysis. It identifies the problem or question being addressed, sets the context, and explains why the analysis is being conducted.

In your report, it is essential to introduce (at least) ONE research question clearly and explicitly. Here is a list of a few examples of possible research questions. Your team can either choose from this list and refine to make it more specific or design another research question.

- How do weather and road conditions influence accident severity?
- What factors contribute most to fatal accidents and injury severity?
- Can we predict accident severity using machine learning models?
- How do seatbelt usage, seating position, and vehicle type relate to injury outcomes?
- What is the impact of road types, intersections, and speed zones on accident risk?
- Can clustering techniques identify high-risk accident scenarios?
- How do vehicle characteristics affect accident likelihood and severity?
- Which data features are most important for improving accident prediction models?

While the possibility exists to explore more than one research question, it's important to note that the pursuit of several questions is not necessarily desirable or likely to lead to greater marks. We will primarily evaluate the quality of your work by assessing the depth of your analysis, and the insights it yields, rather than simply covering a larger quantity of content or material.

The following sections outline some required activities as part of investigating your research question.

4.2. Data Pre-processing Component (mandatory)

Throughout this subject, you've encountered various data preparation techniques, including handling missing values, reshaping data, scaling, encoding, discretising, merging datasets, and feature engineering. You've also explored dimensionality reduction and text processing for simplifying complex data.

In this assignment, you must apply at least ONE data preprocessing task, though you're encouraged to use multiple if they enhance your analysis. Your chosen method(s) should align with your research question, and you should justify your selection in your report and presentation.

Possible Data Preprocessing Tasks:

- Feature Engineering - Create new variables such as an Accident Severity Index, Weather-Road Risk Score, or Vehicle Risk Score by combining relevant attributes.
- Data Integration & Merging - Combine multiple datasets (Accident, Person, Vehicle) using shared identifiers and enrich data with location-based attributes.
- Encoding & Transformation - Convert categorical data into numerical format, normalise numerical features, and discretise continuous variables like speed zones.
- Outlier Detection & Handling - Identify and manage anomalies in accident severity, vehicle characteristics, and speed limits using statistical methods.

You may select a task from this list or propose your own, provided it supports your research question. Remember, there's no single expected solution here. The depth of your analysis and understanding of the data will be a foundation for the rest of your project.

4.3. Correlation and Causal Analysis Component (mandatory)

Understanding relationships between variables is an important step in understanding the accident data. In this component, you will explore how different factors—such as road conditions, weather, vehicle attributes, and driver behaviour—correlate with accident severity and injury outcomes.

You must perform (at least) ONE correlation analysis, though you are encouraged to investigate multiple relationships where relevant. Clearly justify your chosen approach in your report and presentation.

Possible Correlation Analysis Tasks:

- Identify Key Risk Factors – Determine which factors (e.g., road, weather, vehicle, driver) are most associated with severe accidents.
- Examine Vehicle Impact – Explore how vehicle characteristics (e.g., age, type, safety features) relate to accident outcomes.
- Assess Road & Environmental Influence – Investigate how road conditions, intersections, and weather contribute to accident likelihood.
- Analyse Human Behaviour Trends – Study the effects of seatbelt usage, time of day, and accident type on injury severity.

You may select a task from this list or propose your own, provided it supports your research question.

Remember, correlation does not imply causation. Be mindful of confounding variables, spurious correlations, data biases, and reverse causality when interpreting results. Acknowledge these limitations in your report to ensure a more accurate and critical analysis.

4.4. Supervised Learning Models and Evaluation Component (mandatory)

In analysing this dataset, machine learning models can be used to predict accident-related outcomes based on factors such as road conditions, weather, vehicle attributes, and driver behaviour. You must implement (at least) TWO supervised learning models and compare their performance. Clearly justify your model choices and evaluation approach in your report and presentation.

Possible Prediction Tasks:

- Accident Severity Prediction – Classify accidents as minor, serious, or fatal based on road, weather, and vehicle features.
- Multi-Vehicle Crash Prediction – Determine whether an accident is likely to involve multiple vehicles based on location, time, and road type.
- Weather-Based Accident Risk – Estimate the probability of an accident occurring under certain atmospheric conditions.
- Road Condition Impact – Predict the severity of an accident based on surface conditions, lighting, and traffic control measures.
- Emergency Response Need – Predict whether an accident will require hospital transport based on injury reports and accident severity.

You are welcome to use any supervised learning model covered in lectures or any other model, provided you can *justify your choice* and *defend your implementation*.

You should evaluate your models based on appropriate performance metrics and provide a meaningful comparison of their effectiveness. Consider classification metrics such as accuracy and F1-score to assess overall performance or use a confusion matrix to analyse misclassification patterns and identify specific errors. Use appropriate methodology for splitting data intro training and testing as part of evaluation.

Consider potential biases in data, class imbalances, and overfitting risks when interpreting results. Acknowledge these limitations in your report

NOTE: You are welcome (and indeed strongly encouraged) to make use of any relevant existing Python libraries (such as *sklearn* or *scipy*) in your work.

4.5 Data Clustering & Risk Profiling (ONLY required for groups of FOUR)

Clustering techniques can be used to identify hidden patterns in accident data. By grouping similar accidents, drivers, vehicles, or road conditions, we can gain insights into high-risk scenarios and develop targeted safety interventions.

For this project, groups of four must perform (at least) ONE clustering analysis, though you are encouraged to explore multiple approaches where relevant. Clearly justify your clustering approach, choice of features, and interpretation of the results in your report and presentation.

Possible Clustering Analyses:

- Driver Demographics & Risk Profiles – Group drivers based on age group, gender, road user type, and injury level to identify patterns in accident involvement.
- Time & Location-Based Clustering – Identify accident hotspots by clustering based on the time of accident, day of the week, and location coordinates.
- Vehicle Characteristics & Crash Patterns – Cluster vehicles based on engine power, weight, construction type, and fuel type to analyse their involvement in severe accidents.
- Road & Weather Condition Clustering – Group accidents based on surface conditions, atmospheric conditions, and road geometry to uncover risk factors for hazardous driving conditions.
- Collision Type Clustering – Categorise accidents based on event types, initial impact points, and number of vehicles involved to detect common crash scenarios.
- Traffic Control & Accident Frequency – Cluster accidents based on traffic control measures (e.g., signals, stop signs, roundabouts) to analyse their impact on accident rates.

There is no single correct clustering approach—your goal is to identify meaningful patterns in the data and interpret their implications. Consider potential biases in the clustering process, the impact of feature selection, and whether additional factors could improve the analysis. Your report should discuss the significance of your findings and their potential use in improving road safety policies.

5. Report

Your primary deliverable for this assignment is your report. The report should follow the structure of a technical paper. It should describe your approach and observations, both in data preparation, and the data processing algorithms you tried. Its main aim is to provide the reader with knowledge about the chosen research question, providing critical analysis of your results and discoveries.

The following is the expected structure of the report for this assignment.

- **Executive Summary:** A concise overview of the entire report, summarising the objectives, methods used, key findings, and recommendations. This section provides a high-level snapshot of what you have done.
- **Introduction**: This section introduces the purpose of the report, the problem or question being addressed, and introduces the data sources used. It sets the context and explains why the analysis was conducted.

- **Methodology**: Detailed explanation of the methods, techniques, and tools employed for data preparation, analysis, and interpretation. When writing this section, you can assume that the reader is familiar with the technical terms.
- **Data Exploration and Analysis:** Present the results of your data analysis. This section may include descriptive statistics, evaluation results and visualisations gained from exploring the data. Use charts, graphs, and tables to illustrate patterns, trends, and relationships.
- **Discussion and Interpretation:** Provide a list of interesting findings and an in-depth interpretation of them. Bullet points or numbered lists can help highlight these findings. Explain the significance of the patterns observed. Explain why these findings are interesting and valuable. Discuss any unexpected or interesting insights that emerged. (This is the most important section of your report)

  Remember we are more interested in seeing evidence that you have thought about the task and can identify possible reasons for your results. You should also connect them back to your research question. You can also add complementary experiments and their results in this section.
- **Limitations and improvement opportunities:** Address the limitations of the analysis, such as data constraints, potential biases, or assumptions made. Explain what else could be done as future work to improve your analysis.
- **Conclusion:** Summarise the main points of the report and reiterate the key findings and recommendations. Emphasise the value and potential impact of the analysis.
- **References:** List any sources, references, or citations used in the report, especially if you've drawn upon external research or literature to inform your analysis.

We've supplied a template for the report via the assignment page. You are welcome to work with the provided template or customize it according to your preference.

6. Oral Presentation and Assessment

You need to conduct an oral presentation explaining what you have done for assignment 2. Your presentation should encompass the key components below:

1. *Introduction of Research Question*: Begin by introducing the research question that guided your assignment. Explain briefly why it is relevant to road accident policy makers or implementors.
2. *Methods, Techniques, and Tools*: Elaborate on the methods, techniques, and tools you employed for both data preparation and data analysis. Explain how you gathered, cleaned, and structured the data, as well as the analytical techniques and machine learning techniques you utilized.
3. *Presentation of Results*: Share the outcomes derived from your data analysis. Provide a concise overview of the insights you gained through your analytical process.
4. *List of Findings and In-Depth Interpretation*: Present a list of the findings from your analysis. Then provide an interpretation of these findings, shedding light on the significance and implications they hold in relation to your research question.

5. *Limitations and Improvement Opportunities*: Address the limitations encountered during your study, discussing any constraints or challenges that might have influenced the results. Furthermore, demonstrates suggested potential areas for improvement and development.

The presentation requirements are as follows:

- **Timing**: Your presentation should take exactly **8 minutes**. If your presentation doesn't finish on time the markers will interrupt and stop you and it will also negatively impact your mark. There will be a further **15-20 minutes** of questions and answers with the markers.
- **Presenters**: In person attendance at the presentation is mandatory for all team members unless they have been granted an exemption by the teaching staff. Each member of the group is expected to contribute to the presentation content.
- **Slides**: To ensure fairness for all groups and prevent last-minute modifications based on other teams' work, when presenting you will be asked to use the exact version of the slides that you submitted to Canvas.

6.1. Oral Assessment

After the presentation, there will be an oral assessment of all team members' knowledge of the assignment. During this Q&A session, each member will be evaluated individually. Tutors will ask questions about the **entire** report rather than focusing on your specific sections. All members are required to respond independently to oral questions regarding both the report and the code.

7. Teamwork

As mentioned previously, two marks for this assignment are determined by the results of your teamwork evaluation task. Furthermore, based on these evaluations and any other relevant information, we reserve the right to adjust assignment grades to ensure fair outcomes.

The group contract outlines the expectations and responsibilities of each group member. It's crucial that every member actively participates in this assignment. Remember, your understanding of the entire project will be assessed during the oral evaluation.

If you encounter any challenges with inactive team members who aren't responsive to your inquiries, please reach out to Hasti Samadi for assistance in finding a solution.

8. Assessment Criteria

The report will be marked according to the rubric published on the assignment page. The oral presentations and oral assessments will also be marked according to their published rubrics.

Although your code is not assessed directly, you must submit the code that produced the results presented in your report. If you do not submit executable code that supports your findings, we reserve the right to give your team **zero** marks for the report section.

9. Terms and Conditions

9.1 Changes/Updates to the Assignment Specifications

We will use Canvas to advertise any (hopefully small-scale) changes or clarifications in the assignment specifications. Any addendums made to the assignment specifications via Canvas will supersede the information contained in this version of the specifications.

It is your responsibility to ensure you are adhering to the latest iteration of these specifications should updates be announced.

9.2 Late Submissions

Due to the group-based nature of this assignment and its potential overlap with oral assessments, we strongly discourage extension requests.

However, if one or more group members experience severe unforeseen circumstances that significantly impact the group's ability to submit by the due date, the group may apply for an extension. In such cases, you must refer to the FEIT extension policies outlined on the subject's Canvas page.

Please note that even short extensions (up to 3 working days) are unlikely to be granted unless there is sufficient evidence to demonstrate the severity of the situation.

9.3 Academic Honesty

While it is acceptable to discuss the assignment with others in general terms, excessive collaboration with students outside of your group is considered cheating. Your submissions will be examined for originality and will invoke the University's Academic Misconduct Policy where either an inappropriate level of collaboration or plagiarism appears to have taken place.

We highly recommend (re)taking the academic honesty training module in this subject's Canvas. We will be checking submissions for originality and will invoke the University's Academic Misconduct policy where inappropriate levels of collusion or plagiarism appear to have taken place.

9.4 Policies on use of Generative AI tools

Generative AI (GenAI) may be used to assist with your development of code, provided it is appropriately cited.

GenAI must not be used to write any phrases or sentences or paragraphs or sections of your report. It should not be used to translate any part of your report from another language. It should not be used to "polish" any writing in your report.

Any misuse of GenAI, including failure to acknowledge its use, will be considered academic misconduct. The oral assessment will evaluate your understanding of your report. If we find major differences between your written work and your explanations, we reserve the right to investigate further.

9.5 Data Acknowledgement

The data used in this assignment is extracted from the datasets provided on the State of Victoria's open data platform under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

When using this data in your report, you must provide proper citation in accordance with the licensing requirements. Failure to do so may result in academic integrity concerns. Below is an example of how you can cite the dataset in your report:

Example citation:

State of Victoria. (2025). *Victoria Road Crash Data*. Retrieved from https://discover.data.vic.gov.au/dataset/victoria-road-crash-data.