School of Computing and Information Systems
The University of Melbourne
COMP20008 - Elements of Data Processing, Semester 2, 2025

Assignment 2 – Victorian Integrated Survey of Travel and Activity Analysis

Release:	Friday 12 Sept 2025 at 2 PM
Due:	• *Group contract*: Friday 26 Sept at 11:59 PM • *Code and Report submission*: Friday, 10 Oct at 11:59 PM • *Slides submission:* Friday 17 Oct at 11:59 PM • *Oral presentation*: Week 12 (Monday 20 to Friday 24 Oct) • *Team-Evaluation:* Friday 24 Oct at 11:59 PM
Marks:	The Project will be marked out of 35 and will contribute 35% of your total mark.
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 Victorian Integrated Survey of Travel and Activity (VISTA). The dataset contains detailed information about households, individuals, trips, stops, and journeys across Victoria. It provides insights into how Victorians travel for work, study, and other purposes.

Through this project, you will:

- Develop a research question and investigate it
- Perform data and text processing to clean and structure the dataset
- Conduct a correlation analysis between household, demographic, and travel attributes
- Implement supervised learning models to predict travel-related behaviours or outcomes
- Apply feature selection techniques to improve model performance
- (Groups of 4 only) Use clustering methods to profile household or travel behaviour

Your findings will be summarised in a technical report, supported by data visualisations and code implementations. The audience of your report should be the teaching team, so you can assume that the basic technical terms are known. You will present your report and analysis in an oral presentation and respond to questions after the presentation.

2. Assignment Structure

This assignment has multiple components; please read this carefully to adhere to all its elements.

Group Contract – (Due: Friday 26 Sept at 11:59 PM) – Failure to submit mean 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 customise it according to your preference. *Submit as a single PDF file via Canvas (Assignment 2: Group Contract).*

You may vary your group contract throughout the semester, but proposed changes should be agreed to by all members.

Report and Code Submission – 20 marks (Friday, 10 Oct at 11:59 PM)

1. **Report**: For groups of three students, the report should be 12–13 single-column A4 pages in length, while for groups of four students, it should be 14–15 single-column A4 pages. Maintain a line spacing of exactly 1 with normal margins and ensure that the text font size is 11pt. The page limit includes all the text, including references, any appendices, captions, and any tables or images. Tables and image content should be readable and sensible in size.

   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 via Canvas (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 enough 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 via Canvas (Assignment 2: Code and Comments).*

Slides Submission (Due: Friday 17 Oct at 11:59 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 via Canvas. (Assignment 2: Oral Presentation Slides). No other format is acceptable.*

You will be required to use the exact slides that you have submitted for your presentation.

Oral Presentation and assessment – 13 marks (Due: Monday 20 to Friday 24 Oct)

During week 12, all teams should deliver an oral presentation of their work and findings for Assignment 2. 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. See section 6 for more details.

Teamwork evaluation – 2 marks (Due: Friday, 24 Oct at 11:59 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 will 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 (worth 20 marks).

3. Data Sets

We are using the same set of datasets that were used in Assignment 1. The data is collected from the Victorian Integrated Survey of Travel and Activity (VISTA), which includes multiple tables providing detailed information on households, individuals, trips, stops, and journeys. 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 it to make it more specific for your analysis or design another research question.

- How does household income influence travel mode choice?
- What factors affect the likelihood of working from home?
- How much wasted travel time is observed in commuting journeys, and what are the trends?
- How do travel patterns differ between work-related and study-related trips?
- Can clustering techniques identify distinct household travel behaviour profiles?
- Which demographic features are most important for predicting transport mode?

4.2. Data Pre-processing

Throughout this subject, you have learned various data preparation techniques, including handling missing values, reshaping data, scaling, encoding, discretising, merging datasets, and feature engineering.

In this assignment, you must apply an appropriate number of preprocessing tasks, justified by your research question. It is essential to explain why you selected these methods, as well as discuss alternative methods considered and why they were not chosen.

Possible Data Preprocessing Tasks (You may select a task from this list or propose your own):

- Feature Engineering – Create new variables such as a Travel Efficiency Score, Commute Burden Index, or Wasted Time Ratio by combining relevant attributes.

- Data Integration & Merging – Combine multiple datasets (Household, Person, Trips, Stops, Journeys) using shared identifiers to enrich data with additional attributes.
- Encoding & Transformation – Convert categorical data (e.g., travel mode, education level) into numerical format, normalise numerical features, and discretise continuous variables like trip duration or distance.
- Outlier Detection & Handling – Identify and manage anomalies in travel times, distances, or household attributes using statistical methods.

There is no single expected solution here. What matters is the depth of your analysis, your ability to justify choices, and your critical reflection on alternatives.

4.3. Correlation Analysis

Understanding relationships between variables is an important step in analysing travel behaviour. In this component, you will explore how different factors—such as household characteristics, travel attributes, and demographic features—correlate with travel outcomes.

You must perform an appropriate number of correlation analyses, justified by your research question. It is essential to explain why you selected these methods and discuss the alternatives not chosen.

Possible Correlation Analysis Tasks (You may select a task from this list or propose your own):

- Income and Mode Choice – Determine how household income relates to the likelihood of using private cars versus public transport.
- Distance and Work-from-Home – Explore how travel distance correlates with remote working likelihood.
- Household Size and Travel Frequency – Assess whether larger households make more daily trips.
- Education Level and Commute Patterns – Analyse correlations between educational attainment and trip purposes or modes.

Remember, correlation does not imply causation. Consider confounding variables, spurious correlations, and data biases when interpreting results. Be explicit in acknowledging these limitations.

4.4. Supervised Learning Models and Evaluation

Machine learning models can be applied to predict travel-related outcomes based on household, demographic, and trip attributes. You must implement (at least) TWO supervised learning models and compare their performance. All choices should be justified, including the selection of the used models, hyperparameters, evaluation methodology and evaluation metrics. Do not use the same algorithm with just different hyperparameters (e.g., different k values for kNN). The algorithms need to be different.

Possible Prediction Tasks:

- Travel Mode Prediction – Classify trips as car, public transport, cycling, or walking based on demographic and trip features.
- Work-from-Home Prediction – Predict the likelihood of remote work based on household and person attributes.
- Journey Purpose Prediction – Classify whether a journey is related to work, education, or other purposes.

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*.

Each machine learning algorithm should follow the same evaluation approach and metric to enable a fair comparison in your report. Consider potential biases in data, class imbalances, and overfitting risks when interpreting results. Acknowledge these limitations in your report.

You should present the results with at least THREE metrics, such as confusion matrix, precision, recall, accuracy, F1-score, root mean squared error, mean absolute error, etc. The chosen evaluation methods should be justified, i.e., explain why they are appropriate for your question, methods and/or data.

Every algorithm should have at least ONE hyperparameter to be tuned. The range of hyperparameter values tried and the optimal value should be reported.

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

Clustering techniques can be used to identify hidden patterns in travel data. By grouping individuals, or trips, we can gain insights into travel behaviour profiles.

For this project, you must perform an appropriate number of clustering and profiling, justified by your research question. Clearly justify your clustering approach, choice of features, and interpretation of the results.

Possible Clustering Analyses:

- Household Profiles – Cluster households based on income, car ownership, and trip frequency.
- Trip Patterns – Cluster trips based on distance, duration, and mode to uncover travel patterns. You can connect them to household profiles.
- Education Journeys – Cluster student journeys by education level, mode of travel, and time of day. You can connect them to household profiles.
- Regional Travel Profiles – Identify clusters of travel behaviour by geography and demographics.

There is no single correct clustering approach. Your goal is to identify meaningful patterns in the data and discuss their implications.

5. Report

Your **primary** submission 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 machine learning algorithms you tried. Its main aim is to provide the reader with knowledge about the problem, in particular, a critical analysis of your results and discoveries.
The following is the expected structure of the report for this assignment:

- Executive Summary
- Introduction
- Methodology
- Results Exploration and Analysis
- Discussion and Interpretation
- Limitations and Improvement Opportunities
- Conclusion
- References
- Appendices (if needed)

The details of each section are covered in the provided template.

6. Interactive Oral Assessment

An interactive oral assessment is included in this assignment. It will take place in person and consists of both an oral presentation and answering questions face-to-face. We will provide the specifications and rubrics for this section closer to the date.

7. Teamwork Evaluation

As mentioned previously, two marks for this assignment are determined by the results of your teamwork evaluation task. However, based on these assessments and past records, we will identify any non-contributing members and adjust the overall assignment grade accordingly.

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 comprehension 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 (hasti.samadi@unimelb.edu.au) for assistance in finding a solution.

8. Assessment Criteria

The report will be marked according to the rubric published via the assignment page. The interactive oral assessment will also be marked according to their published rubric.

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 modifications 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 of the entire group.

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) tools may be used for **code assistance only** if appropriately cited. However, GenAI use for writing the report and assistance during the oral is prohibited. Regarding using GenAI for translation purposes, you may use it for minor tasks, such as refining sentence structure. However, writing your report in another language and **using AI to translate** it is **not allowed**. Additionally, it is strictly prohibited to use it to generate any part of your report. This includes using AI to write paragraphs or sections based on prompts. 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 significant differences between your written work and your explanations, we may conduct further investigations.

9.5 Data Acknowledgement

The data is provided by the State of Victoria under Creative Commons Attribution 4.0. You must properly cite the dataset in your report. Below is an example of how you can cite the dataset in your report:

Example citation:

State of Victoria. (2025). Victorian Integrated Survey of Travel and Activity (VISTA). Retrieved September 11, 2025 from https://discover.data.vic.gov.au/dataset/victorian-integrated-survey-of-travel-and-activity-vista.