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

Assignment 2 - Group Contract

Group Name: W12G8
Workshop: Tuesday - 12:00PM
Tutor: Farnaz Pirasteh

Group Members:

- Shadman Rahman Shaan - sshaan@student.unimelb.edu.au
- Zizhao Li- zizhaol3@student.unimelb.edu.au
- Dinh Tung Phan- dinhtungp@student.unimelb.edu.au

Research Question:
Which demographic features are most important for predicting transport mode, and how does work and education travel pattern differ from each other?

Project Overview:

1) Data processing

- Merge required tables; keep analysis units at the trip level with linked demographic/household features.
- Clean and impute (median for numeric, most-frequent for categorical); remove impossible values (e.g., negative duration).
- Encode categorical (one-hot); standardise numeric for linear models.
- Plot quick EDA: mode share by demographics, distributions of distance/duration, and time-of-day histograms to spot outliers and class imbalance.

2) Correlation analysis

- Quantify associations between demographics and travel outcomes (mode, duration, distance).
- Use Cramér's V for categorical–categorical, Spearman/Pearson for numeric–numeric, and point-biserial/ANOVA for mixed types.
- Visualise with a compact correlation heatmap (mixed-type matrix) and ranked bar charts of strongest links.
- Note practical significance (effect sizes) and highlight candidate predictors for modelling.

3) Machine learning model

- Task: predict transport mode (multiclass) from demographics; compare models on the same split.

- Models: 1. Multinomial Logistic Regression (with L1/L2) for interpretability; 2. Tree ensemble (Random Forest or Gradient Boosting) for non-linearities.
- Training & evaluation: stratified train/validation/test; address imbalance (class weights). Report F1-score and a confusion matrix.
- Feature selection & importance: L1-logistic screen + Mutual Information cross-check.
- Robustness: 5-fold CV for tuning; identical preprocessing via a single pipeline to ensure a fair comparison.

Roles and Responsibilities:

Member Name	Role and Responsibilities
Zizhao Li	Data cleaning and preprocess on trips and stops.csv, implement Pearson correlation and visualise with a compact heatmap. Train and test Multinomial Logistic Regression as described in project overview.
Shadman Rahman Shaan	Data cleaning and preprocess on household and person.csv, quantify associations between demographics and travel outcomes, slides final checkpoint.
Dinh Tung Phan	Data cleaning and preprocess on journey to work and journey to education.csv, implement point-biserial/ANOVA for mixed types data, note and highlight predictive factor for modelling. Train and test Random Forest models as described in project overview, report final checkpoint.

Communication Plan:

We have two main communication methods, first it is through an online group chat via WhatsApp, this is for deadline reminder and after office hours contact, we also hold online meeting twice a week for progress update and Q&A session. Secondly, we hold a 2-hours on campus meeting every Monday, where the members will review each other work and help each other finish their tasks as well.

Meeting Schedule:

- Frequency: Once to twice a week for in person meeting from 1-3pm on Monday and Thursday, twice a week for online meeting(30 mins).
- Expectation: Group members are expected to complete their given task before every in-person meeting and ready to take the next task, during that time we also work on the assignment together to solve issues and push the progress of the task. In online meeting, group members will share their thoughts, and issue will others to resolve and discuss about the question.
- Participation: Every member must attend at the meeting as regard to the agreement

Decision-Making Process:

- All decision are made through major vote, if there is a conflict or decision there to be make, the group will approve through voting.

Work Plan and Timeline:

Provide a tentative timeline for key project milestones, including data preparation, analysis, report drafting, and presentation preparation. Allocate time for review and revisions.

Task	Deadline
Data Cleaning & data preprocessing	27/09/2025
Correlation Analysis & data cleaning review	01/10/2025
Machine learning models & correlation analysis review	05/10/2025
Draft report	06/10/2025
Final report	08/10/2025
Draft Slide	13/10/2025
Final Slide	15/10/2025

Code of Conduct:

- **Respectful collaboration:** We listen, disagree politely, and include all voices. No dismissive or discriminatory behaviour.
- **Active participation:** Come prepared, contribute in meetings/chat, and respond to group messages within 24 hours (weekdays).
- **Accountability:** Own your tasks and deadlines; if blocked, flag it early and renegotiate before the due time.
- **Underperformance protocol:** (1) Private check-in and agree a short support plan with clear checkpoints; (2) if missed, reassign critical work to protect the deadline; (3) if the pattern continues, document and escalate to the tutor; (4) contribution statements will reflect actual work.

Disagreements or non-responsiveness:

1. **Direct contact (within 24h):** The concerned member(s) message the person privately (chat or 1:1) outlining the issue, desired outcome, and a proposed next step.
2. **Team review (next meeting or within 48–72h):** If unresolved, the item is added to the agenda. The group agrees on a written action plan (tasks, checkpoints, deadlines) and assigns a buddy for support.
3. **Mediation (if still unresolved):** When checkpoints are missed or disagreement persists, the team emails the workshop tutor.
4. **Escalation:** If mediation fails, escalate to the head tutor/subject coordinator.

Non-responsiveness rule: No reply on a critical task for **>**24h (weekdays) **->** task is reassigned to protect deadlines; the original owner may rejoin later. Repeated non-response triggers step 3-4

Good-faith expectation: All members engage respectfully, meet agreed checkpoints, and accept outcomes recorded in minutes (including contribution statements reflecting actual work).

Signature:

By signing below, each group member acknowledges their commitment to adhere to the terms outlined in this contract.

Student 1: _____ Dinh Tung Phan_____ Date: ___24/09/2025_____
Student 2: _____ Zizhao Li _____ Date: ___24/09/2025_____
Student 3: _____ Shadman Rahman Shaan _____ Date: ___24/09/2025_____

This contract is intended to establish clear expectations and promote effective collaboration among group members throughout the duration of the data analytics project. Any amendments to this contract should be discussed and agreed upon by all group members.