The Impact of Household Income on Travel Modes

Group W01G6

Caitlin Spizzirri	Shenghan Hou	Ji Zhang	Bill Luo
Caitlin.Spizzirri@student.unimelb.edu.au	shenghanhou@student.unimelb.edu.au	jiz8@student.unimelb.edu.au	bill.luo@student.unimelb.edu.au

Executive Summary

This study investigates the relationship between household income and travel mode preferences using data derived from a regional transportation survey. The analysis aims to determine whether income levels significantly influence travel behaviour among households. The project was divided into four analytical stages: data preprocessing, clustering, correlation analysis, and supervised learning. Data preprocessing ensured consistency and removed invalid or duplicate records. Clustering using the K-Means algorithm suggested three coarse groupings with substantial overlap (k≈3), not cleanly separable based on normalised income and travel mode tendencies. Correlation analysis revealed weak linear associations between income and travel modes, though a strong negative relationship was observed between private and active travel. Supervised learning models, including logistic regression and decision trees, achieved limited predictive performance (F1 ≈ 0.3), indicating that income alone is insufficient to predict travel choice. Overall, while household income may shape mobility options indirectly, other demographic and geographic factors likely play a greater role. Future research should incorporate multidimensional features to better understand socio-economic influences on transport behaviour.

Introduction

Understanding the relationship between household income and travel behaviour is essential for developing equitable and sustainable transportation systems. Income often influences access to vehicles, residential location, and the affordability of public or active travel options. However, the degree to which income determines mode choice remains uncertain and may vary across socio-economic groups. This study explores how household income affects travel mode preferences by analysing survey data that captures households’ main modes of travel and reported income levels.

The research is motivated by the need to inform transport policy and infrastructure planning through data-driven insights. The dataset was first cleaned and standardised to ensure reliability, followed by clustering analysis to uncover latent groupings among households based on their income and travel mode patterns. Correlation analysis was then applied to examine the linear relationships between income levels and specific travel modes, while supervised learning models were used to assess the predictive power of income in determining travel behaviour.

Through this multi-stage analysis, the project aims to identify whether income is a strong determinant of travel mode choice, and to highlight the limitations of single-variable approaches in explaining complex mobility patterns.

Methodology

This project adopted a four-stage analytical framework—**data preprocessing**, **clustering**, **correlation analysis**, and **supervised learning**—to explore the impact of household income on travel mode choices. Each stage was designed to progressively refine the data and extract deeper patterns, transforming raw survey information into interpretable behavioural insights.

Data Preprocessing

The preprocessing phase focused on transforming raw, heterogeneous survey data into a consistent analytical format. The original datasets contained both household-level and individual-level information, including multiple journey descriptions and income brackets. Income fields were initially expressed as categorical ranges (e.g., "$65,000–$77,999") or open-ended categories ("$190,000 or more"). Regular expressions were used to extract numerical values, and each range was replaced by its midpoint to create a single continuous numeric variable (hhinc_group).

Travel mode descriptions, which appeared under multiple columns such as mainmode_desc_1, mainmode_desc_2, etc., were normalised through text cleaning and grouped into three broad categories for interpretability: **private** (car, taxi, or other personal vehicles), **active** (walking or cycling), and **public** (bus, tram, or train). Duplicated or incomplete household records were removed, and households with missing or invalid travel mode entries were excluded. The result was a clean dataset containing unique household IDs, a numeric income measure, and a dominant mode of travel per household. This processed dataset provided the foundation for all subsequent analyses.

Clustering

Unsupervised learning was employed to identify potential behavioural groupings among households. The **K-Means** algorithm was selected due to its efficiency and interpretability when applied to low-dimensional continuous data. Income values were first normalised to the [0, 1] interval to mitigate scale bias, while categorical travel modes were numerically encoded (private = 0, active = 0.5, public = 1). To determine the optimal number of clusters, the **Elbow Method** was implemented by computing the within-cluster sum of squares for k values ranging from 1 to 9. A clear inflection point was observed at k = 3, suggesting three coarse groupings with substantial overlap (k≈3), not cleanly separable. These clusters roughly corresponded to low-, medium-, and high-income households, each exhibiting different dominant travel behaviours. This step provided exploratory insights into income-related segmentation within the population.

Correlation Analysis

To further quantify linear relationships between household income and travel mode choices, a **Pearson correlation** analysis was conducted. Prior to computation, travel mode variables were converted into dummy variables using one-hot encoding. The resulting correlation matrix included the household income variable and all mode indicators. Comprehensive visualisations were produced using heatmaps, boxplots, and scatter plots to complement the statistical results. Although the Pearson coefficients between income and travel modes were near zero, indicating negligible linear correlation, strong negative correlation was observed between the *private* and *active* categories ($r \approx -0.93$), confirming that these modes are largely mutually exclusive. Boxplots further illustrated that income distributions across different travel modes were highly overlapping, suggesting that income alone is not a strong determinant of travel preference.

Supervised Learning

The final stage tested whether household income can meaningfully predict main travel mode using two models: multinomial Logistic Regression (with standardisation) and a Decision Tree. The data were split 80/20 with stratified sampling to preserve class proportions. Hyperparameters were tuned via 5-fold CV on the training split only (LR: $C \in \{0.5,1,2\}$; DT: max_depth $\in \{3, 5, \text{None}\}$); the best settings were then refit on the full training data and evaluated once on the hold-out test set using accuracy and macro-averaged F1. On the test set, the Decision Tree achieved ≈0.698 accuracy (and the higher macro-F1), slightly above Logistic Regression (≈0.683 accuracy). Both models heavily favoured the majority class (private), with very low recall for active and public. These results show that income alone lacks sufficient signal for reliable mode prediction, motivating a multi-variable approach that incorporates demographic, geographic, and accessibility features alongside imbalance-aware training.

Results Exploration and Analysis

This section presents the findings from descriptive analysis, clustering, correlation exploration, and predictive modelling. The visualisations and statistical results together demonstrate how household income relates to travel mode preferences within the dataset.

Descriptive Statistics

The initial step examined the number and share of households across each travel mode category. The results are summarised below:

Travel Mode	**Average Household Income**	**Total Households**	**Proportion**
Public	163,222.22	9	0.029
Private	162,022.33	215	0.691
Active	161,200.01	87	0.280

These figures indicate that **private transport** is the dominant mode, accounting for approximately 69% of households. **Active modes** such as walking and cycling make up 28%, while **public transport** is used by only 3% of households. Despite the differences in proportion, the **average household income across all three groups is nearly identical**, around $160,000. This suggests that income level alone does not significantly explain travel mode selection.

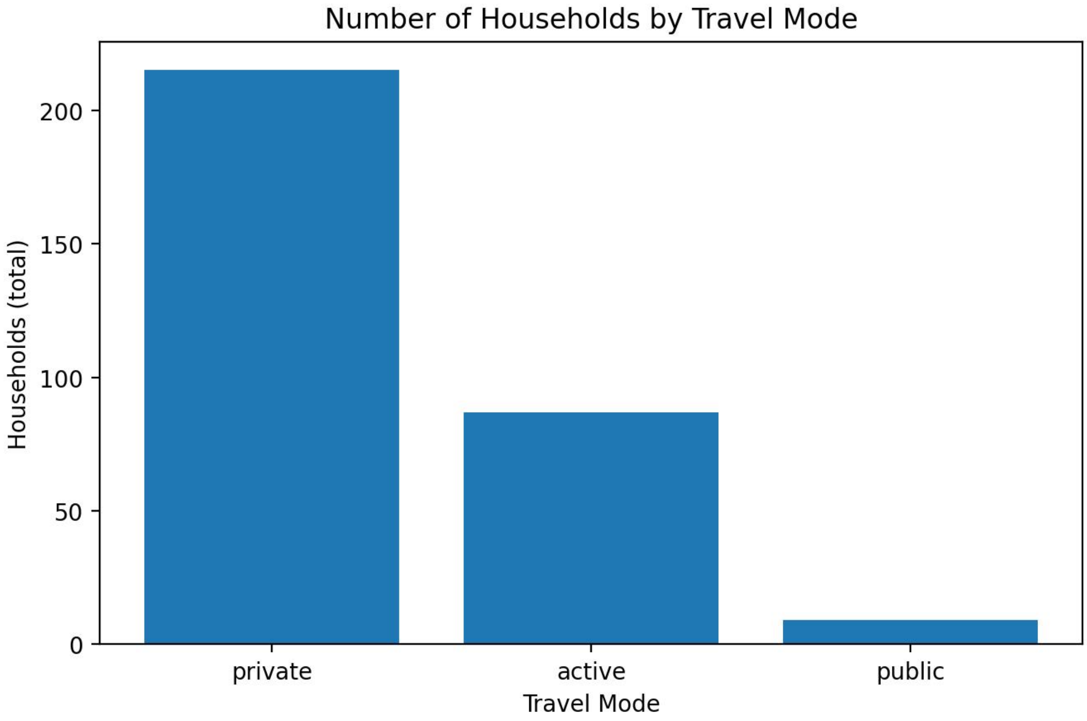

Figure 1: households_by_mode_totals

1) Data Preprocessing

Class balance is highly skewed (private > active > public), foreshadowing majority-class bias (Fig.1).

Income by mode shows large overlap; medians are similar; many outliers across all modes (Fig.2–3).

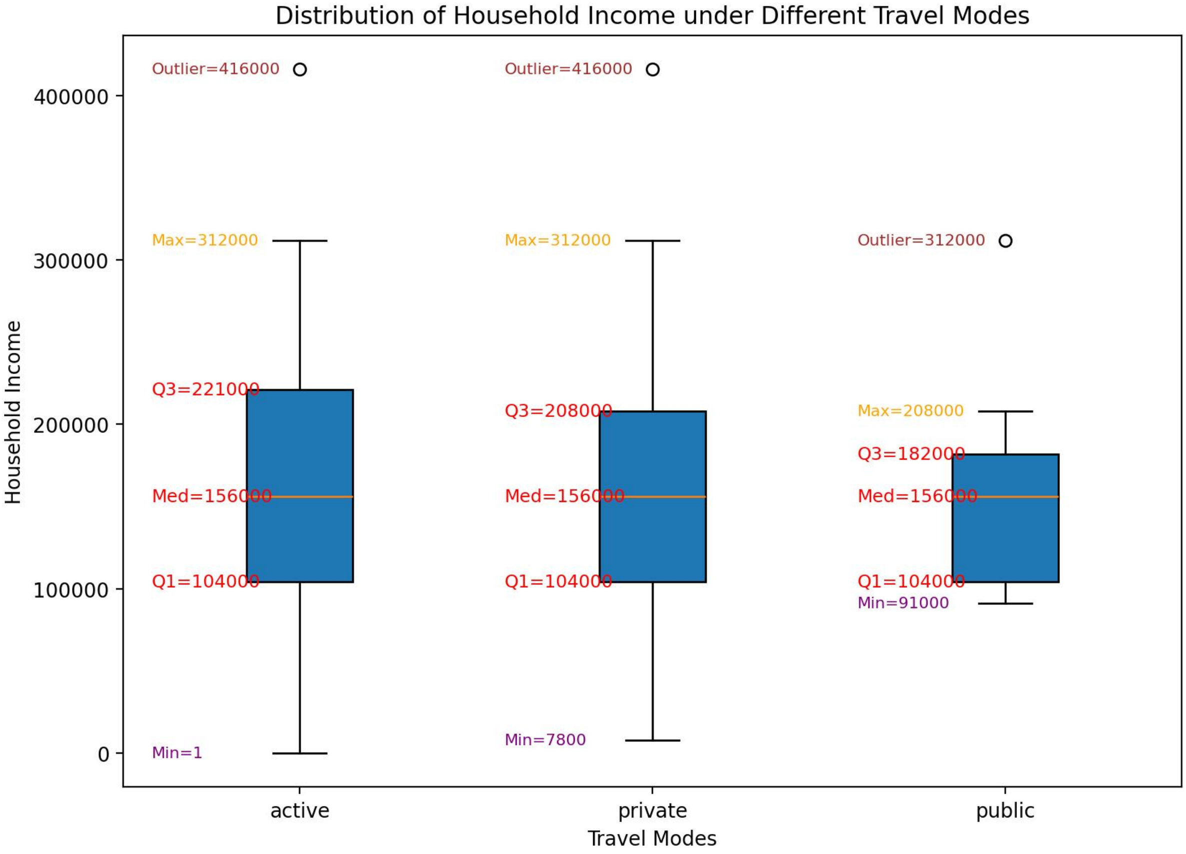

Figure 2:income_boxplot

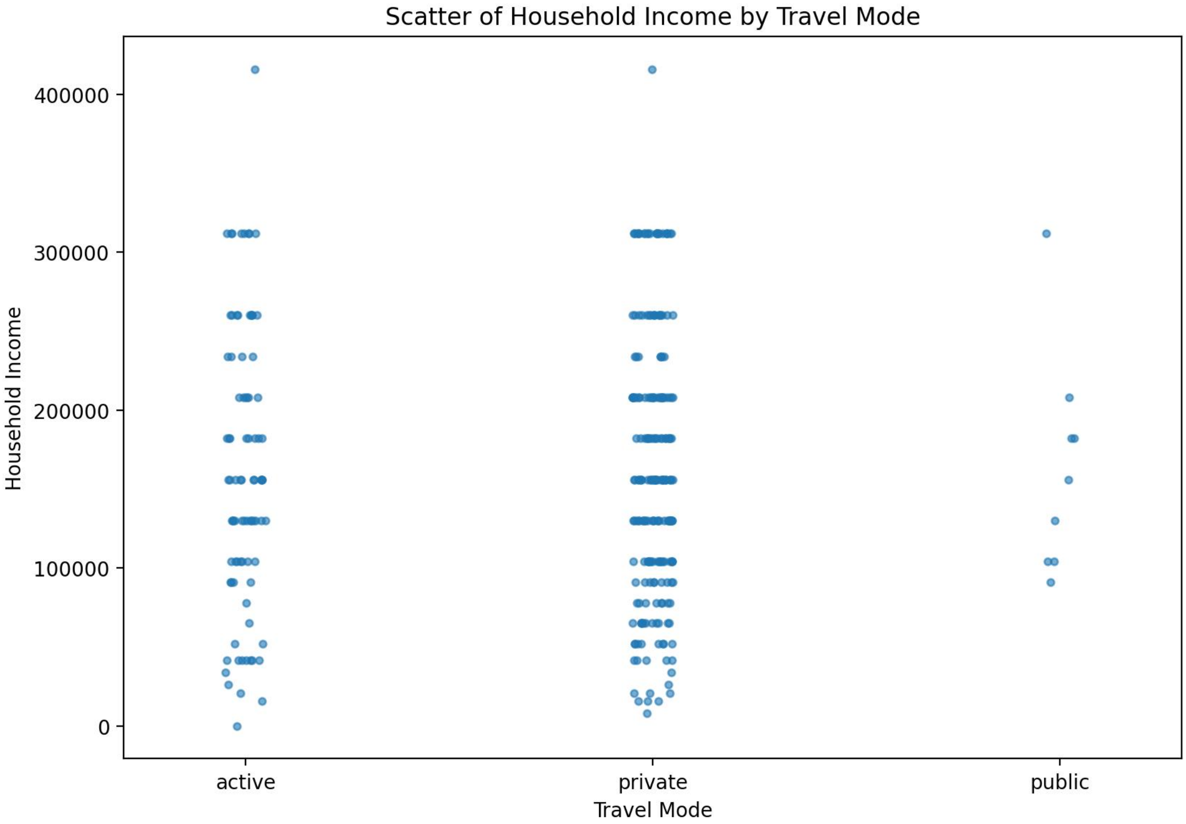

Figure 3:income_scatter_by_mode

Implication: Income alone is unlikely to separate classes; macro-averaged metrics and confusion matrices are necessary.

2) Correlation Analysis

Pearson heatmap shows near-zero correlation between hhinc_group and mode indicators (≈ −0.005 to 0.003).

Strong negatives among one-hot labels are by construction (mutual exclusivity) (Fig. 4).

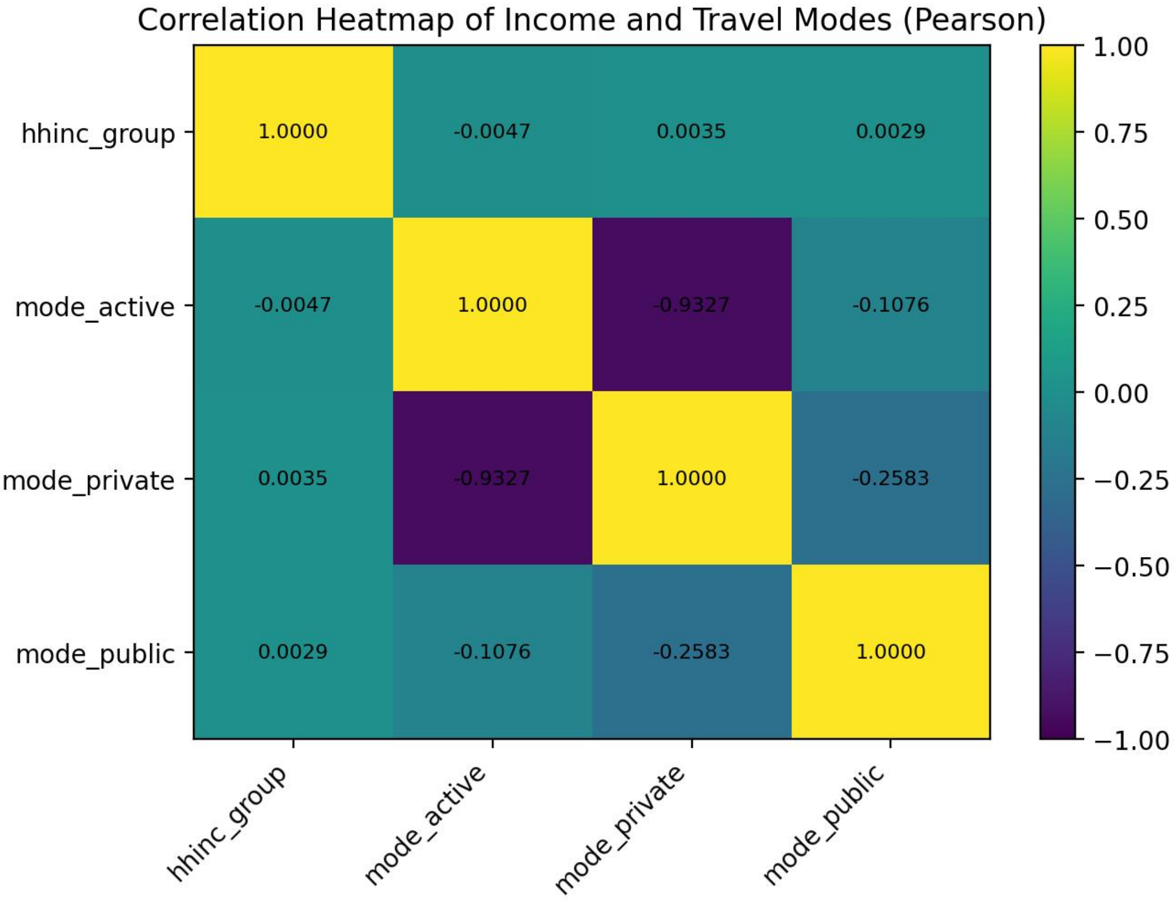

Figure 4:correlation_heatmap

Implication: Linear signal from income to mode is weak.

3) Data Clustering & Profiling

The elbow method showed that the optimal k value was 3 as past that there were diminishing returns(Fig.5).

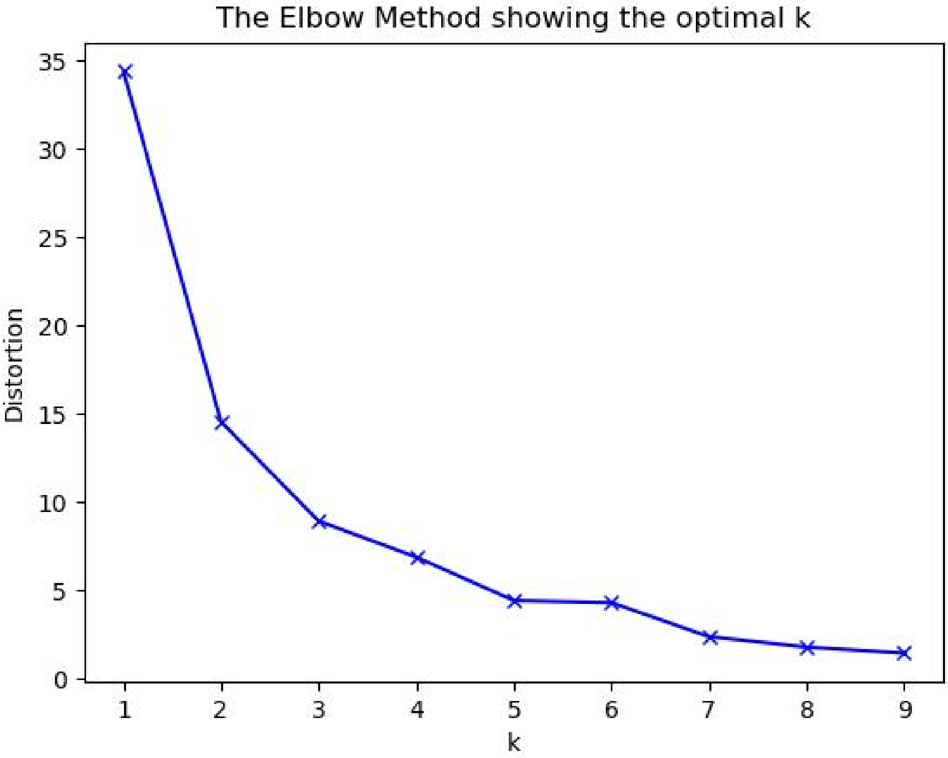

Figure 5: elbow_method

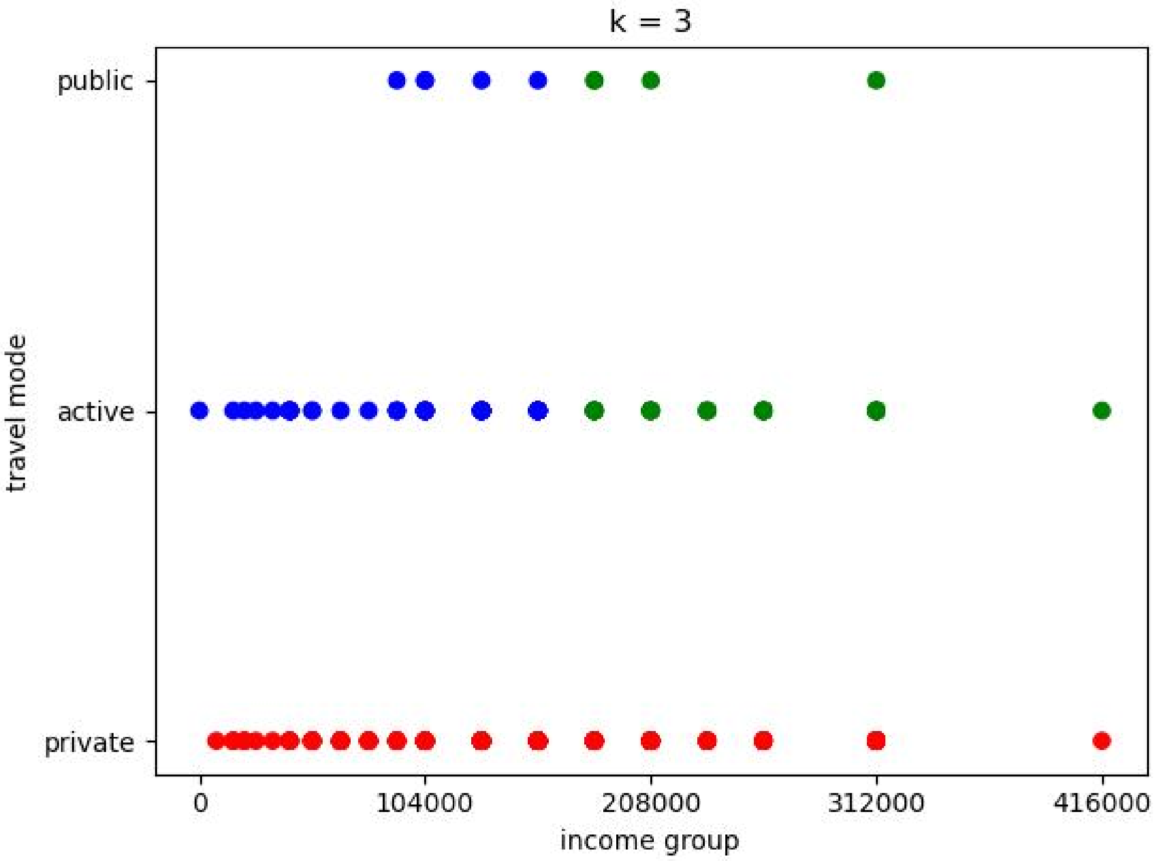

Figure 6:k-means

As shown in Fig 6 there are 3 groups however each group appears arbitrary and mainly joined by similar income. K-means are ill-suited with income-only features; clearer structure likely requires richer predictors.

4) Supervised Learning & Evaluation

Models

Multinomial Logistic Regression (standardized) & Decision Tree (unscaled).
Test metrics:

Decision Tree: Acc ≈ 0.698, Macro-P ≈ 0.456, Macro-R ≈ 0.363, Macro-F1 ≈ 0.335.

Logistic Regression: Acc ≈ 0.683, Macro-P ≈ 0.228, Macro-R ≈ 0.333, Macro-F1 ≈ 0.270. Confusion matrices: LR predicts almost all as private; DT recovers a few active but still misses public (Fig. 7-8)

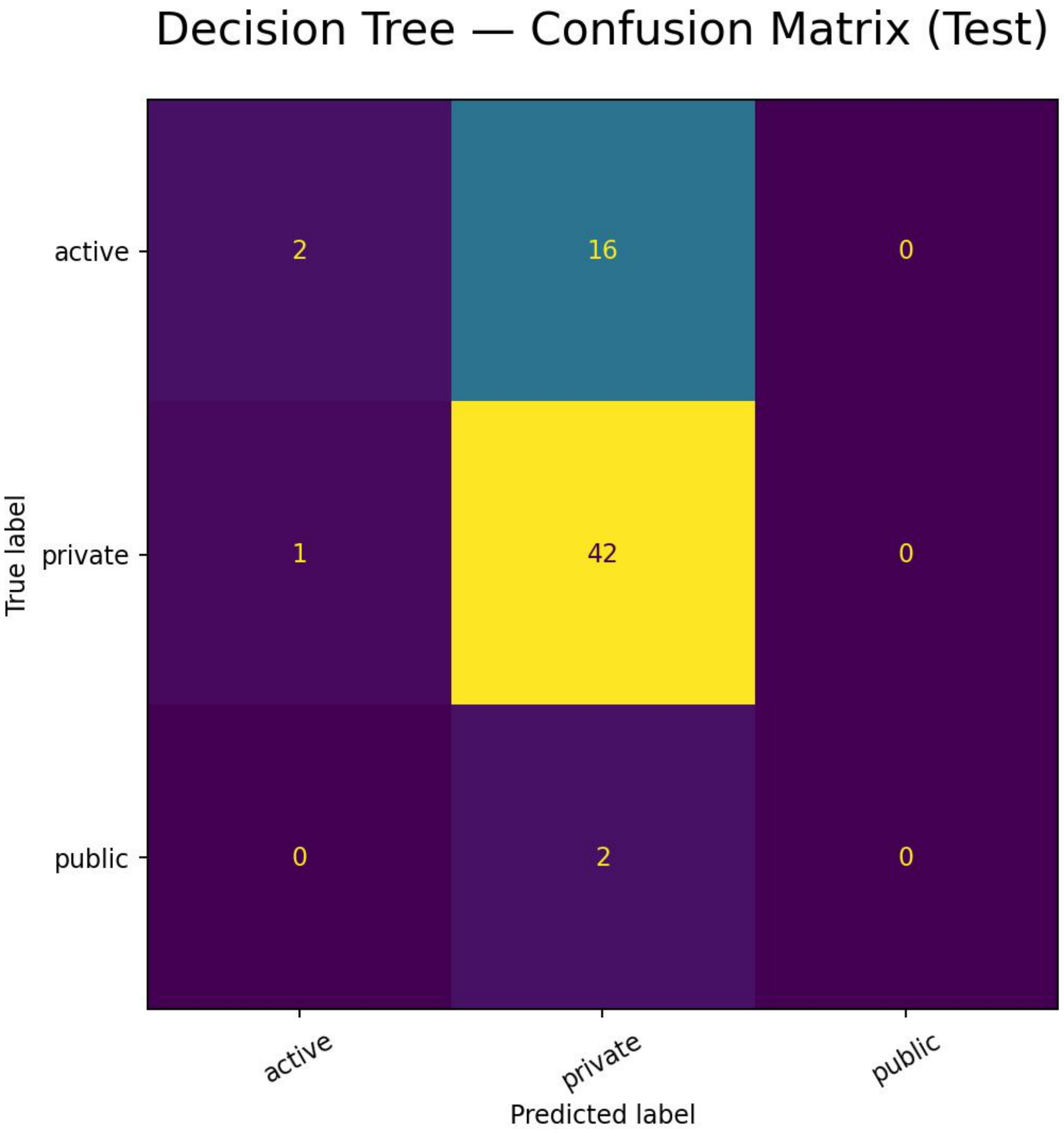

Figure 7:dtree_cm

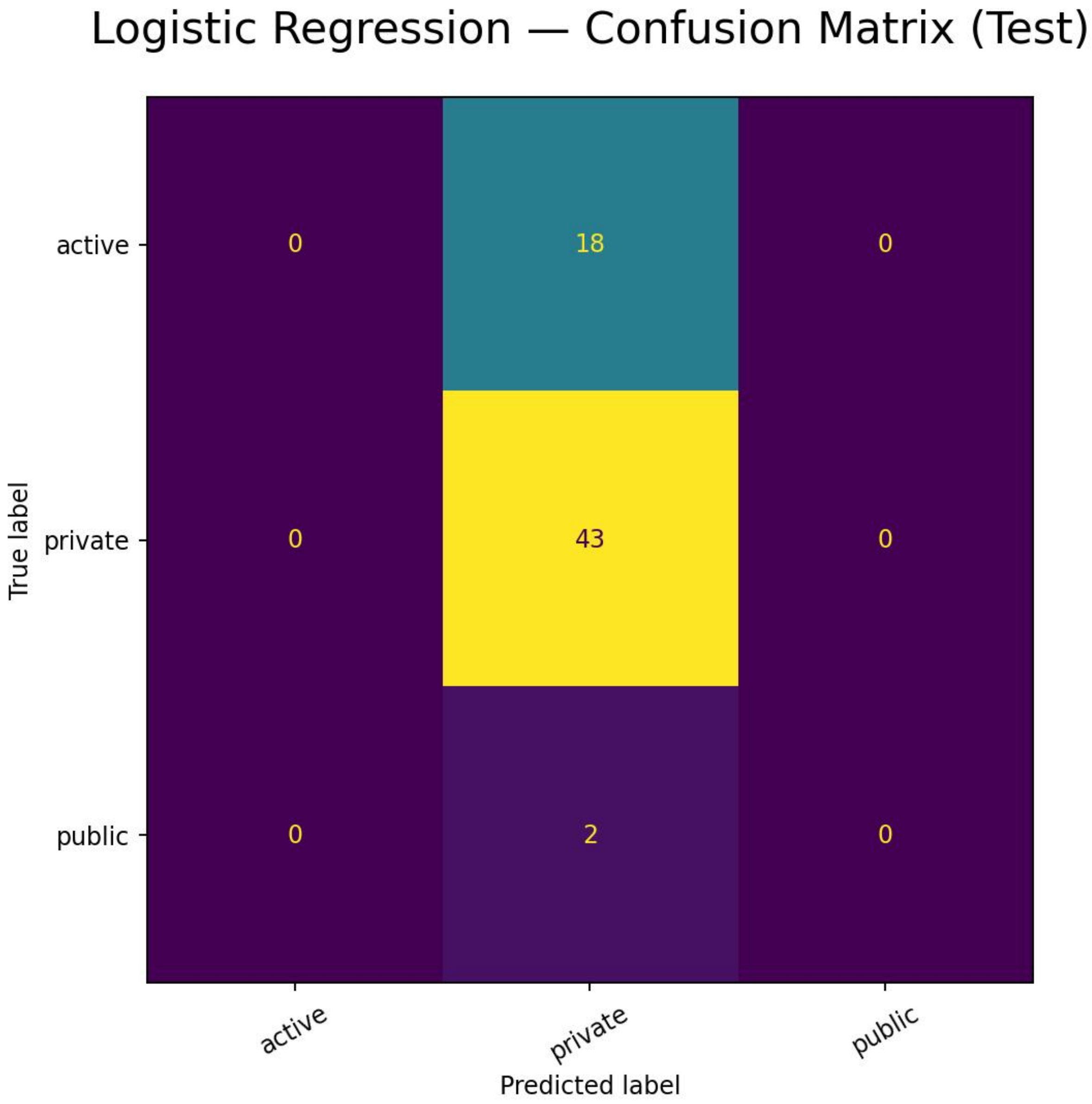

Figure 8:logreg_cm

Implication: With one feature and imbalance, both models favor the majority class; DT's thresholds give a small macro-F1 edge.

5) Summary of Findings

Overall, the results across all analytical methods consistently indicate that household income has minimal influence on travel mode choice. Despite minor income variations among users of different transport types, there is no strong linear or predictive relationship between the two. The prevalence of private transport (≈70%) highlights behavioural inertia and infrastructure bias rather than economic constraints. Policymakers aiming to shift households toward public or active travel modes should therefore prioritise accessibility, reliability, and infrastructure improvement instead of relying solely on income-based interventions.

Discussion and Interpretation

Cross-part coherence
The four components tell a consistent story. Preprocessing shows severe class imbalance (private > active > public; Fig. 1) and substantial overlap of income across modes (Fig. 2–3). Correlation confirms near-zero linear association between income and mode indicators (Fig. 4). Data Clustering & Profiling (K-Means) finds only coarse groupings with strong overlap and no label-like partitions (Fig. 5–6). On top of this evidence base, Supervised Learning yields moderate accuracy but low macro-F1 and confusion matrices dominated by the majority class (Fig.7–8). The pieces align: income alone carries weak discriminative signal, and imbalance drives majority collapse.

What income explains vs. what it doesn't
Income is plausibly associated with a tendency to use private travel (ownership, cost tolerance), but association ≠ separability. The boxplots and scatter (Fig. 2–3) show broad within-class variance and overlapping distributions; any monotonic trend is drowned by overlap and outliers. From a decision-boundary perspective, a single axis (hhinc_group) cannot carve three well-separated regions—hence the near-zero per-class recall for minorities in LR, and only a tiny active recovery in DT.

Metrics matter under imbalance
Accuracy looks passable because predicting private "wins" often. Macro-averaged precision/recall/F1 weighs minority classes equally and exposes the real issue: near-zero minority recall (LR) or very low recall (DT). The confusion matrices (Fig.7–8) visualize this—entire public and almost all active land in the private column.

Model behaviours

Logistic Regression (multinomial) with one standardized feature is effectively fitting a single linear boundary along income ranks; with overlapping distributions, it collapses to the majority class.

Decision Tree uses axis-aligned thresholds, which can catch a few active cases (e.g., a small income range where active is relatively more likely), hence the slight macro-F1 gain. But with only one feature and tiny minority support, the tree cannot generalize to public.

Clustering & profiling as context, not labels
K-Means suggests that the continuous space spanned by income admits only weakly separated clusters (Fig.5-6). Profiling those clusters (e.g., cluster-level income summaries) may reveal small shifts but does not replicate mode labels—consistent with the supervised results and the correlation heatmap.

Robustness and interpretability
We fixed a random seed and used a stratified 80/20 split to stabilize class proportions in the test set (≈63 samples, with public ≈2). While this improves reproducibility, it does not overcome sampling variance for minorities; conclusions should emphasize patterns (majority bias, weak signal) rather than exact point values. The pipeline remains interpretable: simple models, clear metrics, and inspection-friendly confusion matrices.

Implications for analysis & policy
With income alone, any "policy" inference about mode choice would be fragile and inequitable, as minorities are poorly captured. For practical utility, we need richer features capturing behavior and accessibility (vehicle ownership, vehicles per household, trip length, distance to transit, CBD proximity, household size, network access indices), and imbalance-aware training. Only then will per-class performance (especially active/public) become informative and fair.

Limitations

Data and target composition

Severe imbalance: public has only ~2 test cases; active is also small. Minority-class metrics therefore have high variance, and small misclassifications swing macro-F1.

Single predictive feature in SL: We intentionally constrained supervised learning to hhinc_group as a baseline. This limitation explains majority collapse and low macro-F1; it is a design choice, not an algorithm failure.

Label granularity & reliability: main_mode_group aggregates behaviors; misreporting or temporal variability could blur signal.

Potential selection bias: Dropping rows with missing target prevents leakage but may bias the sample if missingness is systematic.

Methodology and evaluation

One split, small test set: A single stratified 80/20 split (seed=7) is reproducible but noisy for minorities. Repeated/bootstrapped CV or nested CV wasn't used; model-selection uncertainty is under-quantified.

Tiny hyperparameter grids: LR C ∈ {0.5,1,2}; DT max_depth ∈ {3,5,None}. This meets the minimum requirement but may under-explore capacity/regularization. No threshold tuning or cost-sensitive optimization was attempted.

Metric scope: We report Accuracy and macro-averaged metrics plus confusion matrices—appropriate, but no per-class PR curves, AUROC (one-vs-rest), or calibration analyses, which could further diagnose minority performance.

Modeling assumptions

Logistic Regression: Linear decision surfaces in a one-dimensional feature space; cannot capture non-linear or interaction effects.

Decision Tree: Axis-aligned splits; with one feature and tiny minority support, depth-limited trees remain brittle and prone to overfitting small pockets.

K-Means (clustering & profiling): Assumes roughly spherical clusters and comparable scales; with overlapping class structure and few features, cluster partitions won't align with labels.

Correlation: Pearson correlations can miss non-linear or interaction-driven associations; the near-zero values don't rule out complex relationships mediated by unobserved variables.

External validity and fairness

Context dependence: Results pertain to this dataset/time period and the specific cleaning/label schema. Transfer to other regions/years is untested.

Fairness & equity: Majority-class bias implies that conclusions/predictions may systematically underserve active/public households; we did not run subgroup fairness audits (e.g., by region or income decile).

Computational & reproducibility details

Versioning: Minimal environment; library versions and random seeds matter. While we fixed a seed, we didn't provide full environment lockfiles.

Figure coverage: We prioritized the final figures listed (Fig.1–8). Any intermediate diagnostic plots not included could add nuance but were omitted for brevity.

Mitigations / Future work

Feature enrichment: add behavior/access variables (car ownership, vehicles per HH, transit distance, network access, household size, trip length/duration), and simple engineered features (income quantiles, interactions).

Imbalance handling: class_weight='balanced', focal loss, or straightforward resampling (e.g., under/over-sampling) evaluated via macro-F1 and per-class PR.

Validation upgrades: repeated stratified CV or nested CV; report uncertainty (CIs via bootstrap), and sensitivity to test_size/seed.

Model set: after feature enrichment, compare linear SVM/KNN and shallow boosted trees under the same unified protocol; keep interpretability by reporting thresholds/feature contributions.

Clustering & profiling: profile clusters with added features; consider GMM or hierarchical clustering if distributions are not spherical; align profiles back to mode outcomes for actionable insights.

Fairness checks: per-subgroup metrics (e.g., by geography/income decile) to detect disparate performance, and mitigate with reweighting or constraints.

Conclusion

Across the full pipeline—Preprocessing → Correlation → Data Clustering & Profiling → Supervised Learning—the evidence is consistent: income alone provides weak signal for predicting a household's main travel mode. Correlations are near zero; K-Means shows overlapping groups; and on a stratified 80/20 split the supervised baselines favor the majority class (private), yielding low macro-F1. Decision Tree slightly outperforms multinomial Logistic Regression by recovering a few active cases, but both models struggle on minorities, especially public.

Interpretation

The task is constrained more by feature insufficiency and class imbalance than by algorithm choice.

Next steps

Incorporate behavioural and accessibility variables, apply imbalance-aware training, expand tuning with stronger validation, and reassess cluster structure and correlations with the enriched feature set. These changes are expected to raise minority-class recall, narrow the gap between accuracy and macro-F1, and deliver conclusions that are both more accurate and more equitable across travel modes.

Reference

Victorian Integrated Survey of Travel and Activity (VISTA) - Victorian Government Data Directory. (2023). Vic.gov.au.
https://discover.data.vic.gov.au/dataset/victorian-integrated-survey-of-travel-and-activity-vista