The impact of household income on travel modes

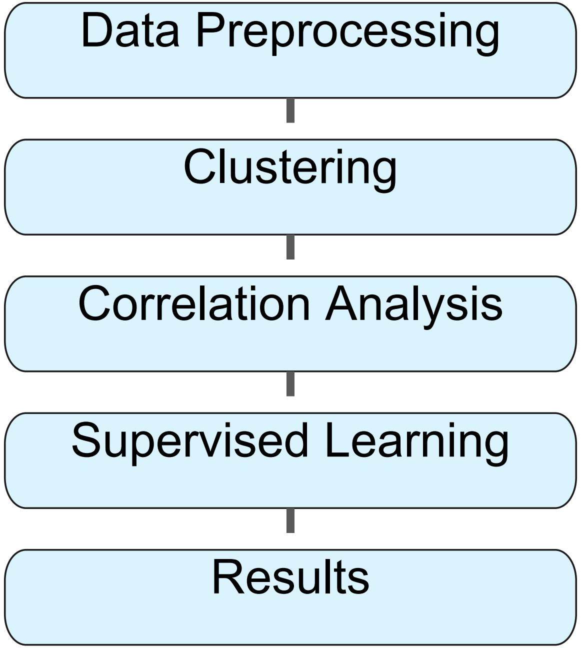
Group W01G6

By Catlin Spizzirri, Shenghou Hou, Ji Zhang, Bill Luo

Research Question

This study investigates the relationship between household income and travel mode preferences using data derived from a regional transportation survey conducted by the Victorian Integrated Survey of Travel and Activity.

Methods and Techniques



Data Preprocessing

Clustering

Correlation Analysis

Supervised Learning

Results

Data Preprocessing

Income data formatting:

Raw:

Cleaned:

"$1,250-$1,499 ($65,000-$77,999)" —> midpoint(65000, 77999) = 71499.5

“"$8,000 or more ($416,000 or more)" —> 416000 (as a maximum income)

Travel mode data formatting:

* Extracted keywords from columns labeled as “mainmode_desc_01” through to mainmode_desc_12”

1. Private (Car, taxi, etc)

1. Active (walking, cycling, etc)

1. Public (bus, tram, train, etc)

Clustering

The **K-Means** algorithm was selected due to its **efficiency and interpretability** when applied to low-dimensional continuous data.

private = 0

active = 0.5

public = 1

Fig 1: elbow_method

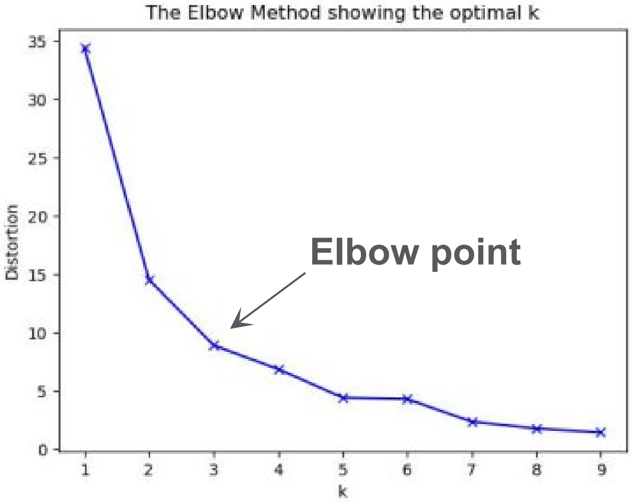
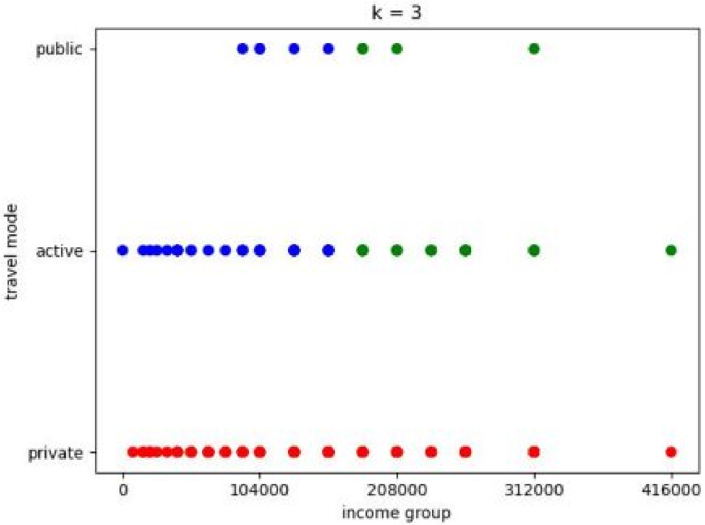
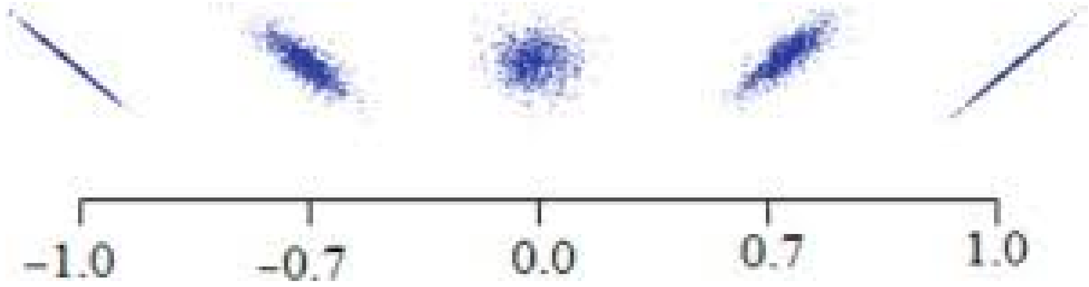


Fig 2:k-means



Correlation Analysis

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, box plots, and scatter plots** to complement the statistical results.



Pearson Correlation Coefficient Scale

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. 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.

Varied parameters:

Logistic Regression: C ∈ {0.5,1,2}

Data Tree: max_depth ∈ {3, 5, None}

Presentation of Results

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.

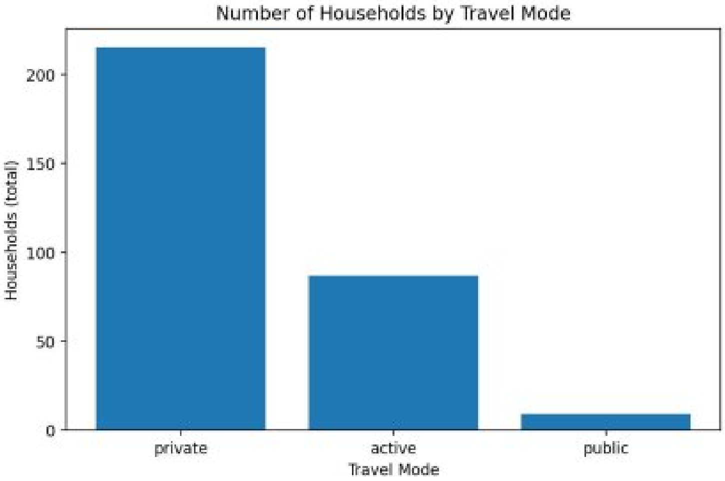
Descriptive Statistics

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	251	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.



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).

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

Fig 1: households_by_mode_total

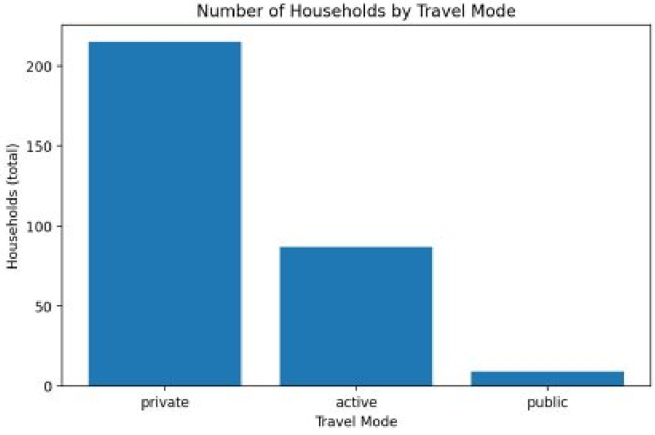


Fig 2: income_boxplot

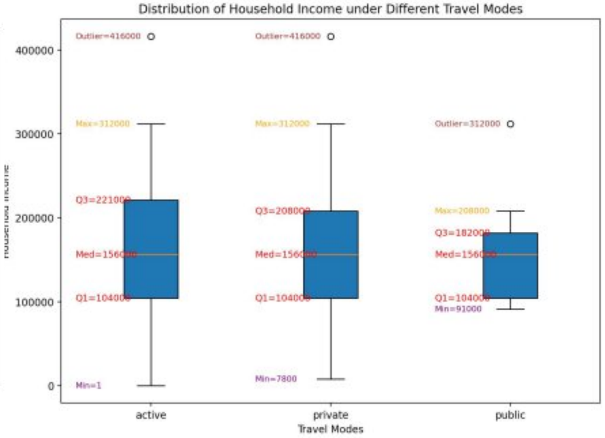
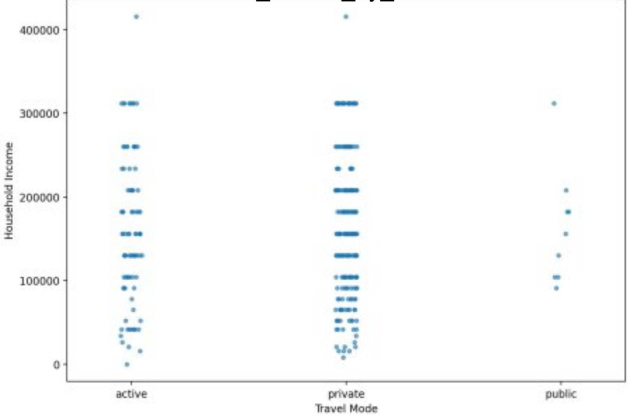


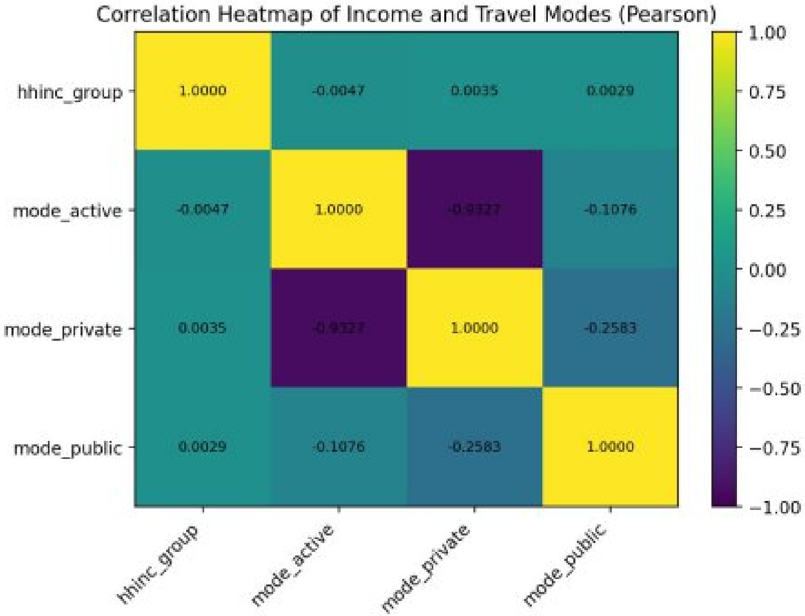
Fig 3: income_scatter_by_mode



Correlation Analysis

Pearson heatmap shows **near-zero correlation** between hhinc_group and mode indicators (ranging from around −0.005 to 0.003).
Strong negatives among one-hot labels are by construction (mutual exclusivity) (Fig. 1). Implies Linear signal from income to mode is weak.

Fig 1: correlation_heatmap



Data Clustering & Profiling

The **elbow method** showed that the optimal k value was 3 as beyond that there are clearly diminishing returns (Fig.1).

As shown in Figure 2, 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.

Fig 1: elbow_method

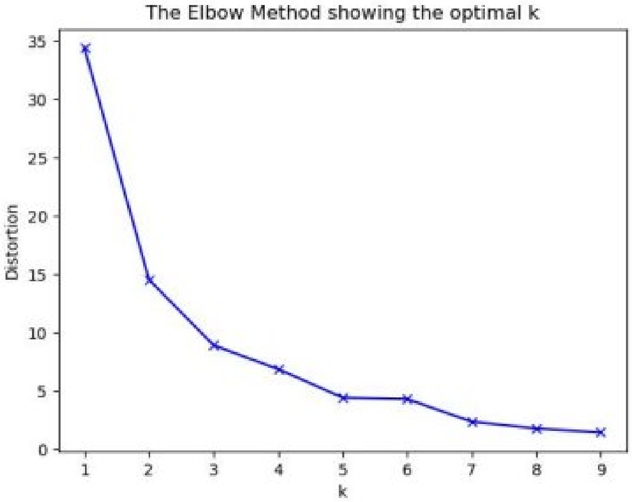
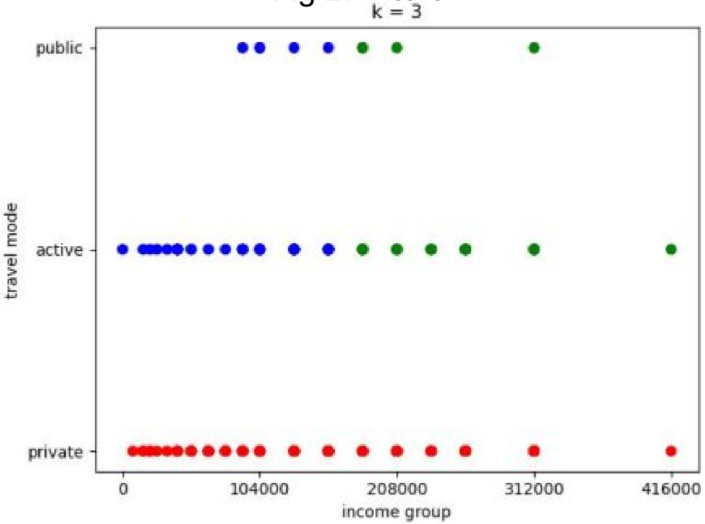


Fig 2:k-means



Supervised Learning Metrics

Decision Tree metrics				Logistic Regression metrics			
Mode classification	Precision	Recall	F1-score	Mode classification	Precision	Recall	F1-score
Public	0.0	0.0	0.0	Public	0.0	0.0	0.0
Private	0.7	0.9767	0.8155	Private	0.6825	1.0	0.8113
Active	0.6667	0.1111	0.1905	Active	0.0	0.0	0.0

On the test set, the Decision Tree achieved approximately 0.698 accuracy and the higher macro-F1, **slightly above** Logistic Regression with around 0.683 accuracy.
The models ended up heavily **favouring the majority class,** that being the ‘private’ group, with a very low recall for active and public.

Overall metrics		
	Decision Tree	Logistic Regression
Accuracy	0.698	0.683
Precision-macro	0.456	0.228
Recall-macro	0.333	0.333
Macro-F1	0.335	0.270

These results demonstrate that income as a factor alone **lacks sufficient signal** for reliable mode prediction, and a more conclusive result would be seen with a **multi-variable approach** that incorporates other important details like demographic, geographic location, etc.

Supervised Learning Matrix

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

Logistic Regression model predicts almost all as private– Decision Tree recovers a few active but still misses public (Fig. 1-2).

Fig 1: dtree_cm

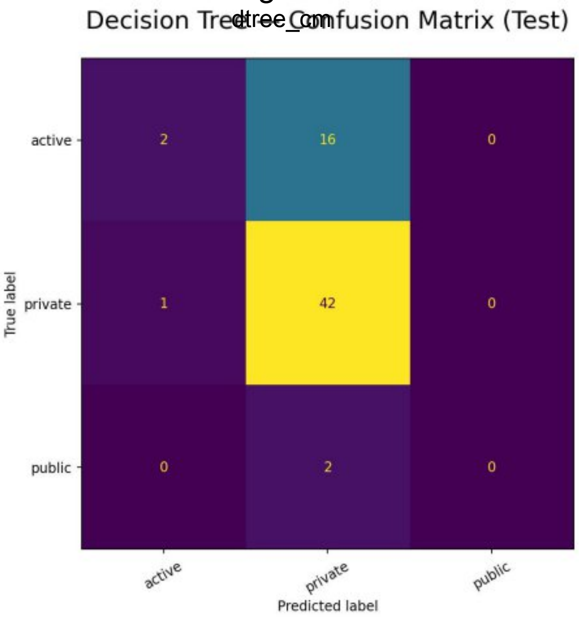
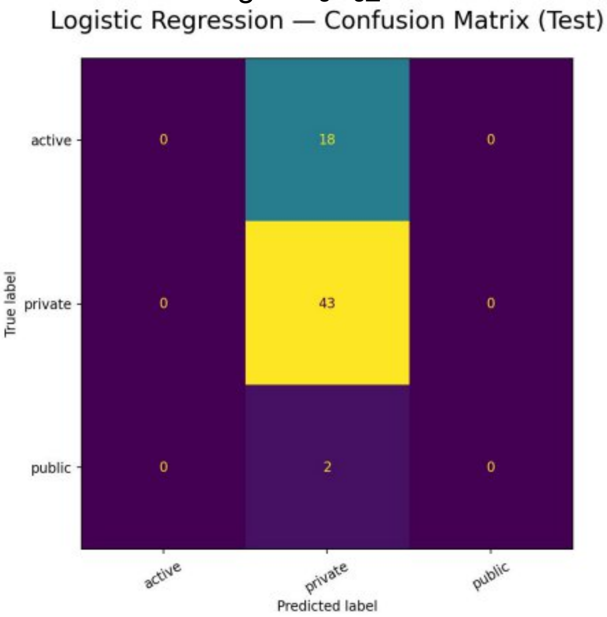


Fig 2: logreg_cm



List of Findings

1. private transport dominates ≈70%
2. **Average household income** is roughly $160,000 a year across *all modes*
3. there is **strong negative trend** (r =-0.93) between private and active
4. Pearson correlation for income and travel mode is **near zero** (≈−0.005 to 0.003)
5. **3 clusters** for k-mean clustering appeared but had **large overlaps.**
6. The decision tree and logistic regression models **almost always chose private or active**
7. Majority mode made accuracy better than it should have been

Interpolation

- Income is plausibly related to travel mode however it alone **cannot explain travel mode.**
- Accuracy for many of the methods isn’t bad but this is mostly due to predicting ‘private’ working most of the time.
- **Logistic regression** with one standardization feature **collapses to the majority class.**
- Active and especially public travel modes had few members suggesting **something other than income** explained travel mode.
- This all highlights a need for a higher dimensional analysis.

Limitations & Improvements

1. The dataset heavily **favours private** transport
2. Public dataset **too small**
3. Minority class metrics have **high variance** because of how small public was
4. Main_mode_group **combines the household together** which might misrepresent groups
5. Dropping rows with no travel mode might add biases to the dataset

To improve:

1. A simple possible improvement could be to **draw from more data** from the given dataset, this might get rid of majority class imbalance as well as add more dimensions
2. Try using agglomerative clustering, or divisive clustering
3. Use repeated cross validation / bootstrap confidence intervals