MONASH University
Engineering

Modelling Electric Vehicle Ownership in Melbourne and Geelong

ENG4702: Final Year Project - Final Report

Author(s): Daniel Lawson (30574218)

Supervisor(s): Le Hai Vu, Bob La

Date of Submission: 22/10/2023

Project type: Research

1 Executive Summary

Currently we are living through the electric vehicle (EV) revolution and with this growth comes challenges. The smooth transition to electrification will require policy makers to be informed about what and where EV infrastructure is to be built, which relies on understanding EV ownership. This project has created two distinct methods for predicting EV ownership in Melbourne and Geelong. Both methods use a synthetic population which includes every household in Melbourne and Geelong as well as the geospatial location and household attributes of this population. The data generated by this project lays the groundwork for future research.

Method 1 was developed to predict which households in this population are most likely to own EVs, without using any EV data. This was achieved by using K-means clustering to group the population into similar households. Each one of these groups were then individually inspected to identify the one who's household attributes are most similar to those of EV drivers establish by prior research. Due to not having access to EV data method 1 greatly overpredicted the number of EVs in Melbourne and Geelong by 24 times, predicting 148,394 EVs when the true total is 6,215. All households predicted to own EVs had household attributes strongly associated with EV ownership, because of this the model has been reinterpreted as a prediction of household who are likely to buy EVs in the future rather than a prediction of current owners.

Method 2 used the synthetic population as well as EV data that detailed the number of EVs in each Postcode Area (POA). Both the EV data and the synthetic population were used to train a Multilayer Perception (MLP) network. This MLP can predict the number of EVs in an area by providing it with the households in that area. The MLP predicted 5,507 EVs when given household data aggregated to the POA level resulting in an error of approximately 10%. When predicting EVs given household data aggregated to SA1 the model predicted 5,898 EVs and had an error of approximately 5%.

The overall contribution of this project is the development of two methods for assigning EVs to the most likely households in a synthetic population. Method 2 is the preferred method for predicting EV ownership due to its greater accuracy. The trained model does not need EV data to predict EV households and has been proven to work with geospatial data from POA down to SA1.

Contents

2 Introduction

Electric Vehicles (EVs) are increasingly becoming more common as they are perceived as an eco-friendly alternative transportation option (LaMonaca & Ryan, 2022). With the rise of urbanisation coupled with the use of the Internal Combustion Engines (ICE) in cities, the air quality in many metropolitan areas has deteriorated (Abdel-Rahman, 1998). Additionally, as the understanding of the impacts of Greenhouse Gases becomes more widespread, many countries are introducing policies that will encourage EV adoption (Plötz et al., 2014). This includes the Australian Government who have their own goal of being net-zero by 2050 and are committed to building out Australia's EV charging infrastructure [4].

As the transition to EVs increases gains momentum, so does the demand for electricity, which poses a challenge to the power grid (Deb et al., 2017). By understanding EV ownership through modelling, resources can be better allocated to policies and infrastructure that will smooth the transition of ICE vehicles to EV. Modelling EV usage involves understanding the driving habits of EV owners, the accessibility of charging infrastructure and the cost of charging Hjorthol (2013). Consequently, it has become crucial to be able to identify EV drivers and their attributes. This knowledge will serve as a foundation for future researchers to build EV usage models upon.

Because of varying methods and data sources, studies into EV usage between regions are not easily comparable (Hjorthol, 2013). Most of the current research concerned with vehicle ownership has been conducted in North America, Asia and Europe (Ma & Ye, 2019). This is why it is important for more research to be conducted in Melbourne and Geelong.

By leveraging the knowledge gain from prior research, this project aims to develop EV ownership models that are tailored to Melbourne and Geelong's unique characteristics. Researching and using data specific to Melbourne and Geelong, will enhance our current understanding of EV ownership and usage. Furthermore, it will help guide the direction of EV adoption.

3 Aims and Objectives

3.1 Research Question

How can electric vehicle owning households be identified by their household attributes?

3.2 Aims

Aim 1: To develop a model that can identify EV owning households in Melbourne and Geelong without the use of EV data.

Aim 2: To develop a model that can identify EV owning households in Melbourne and Geelong by utilising EV data.

The final model developed should produce a dataset containing information about household attributes, location, and EV ownership that can then be utilised by external projects.

3.3 Objectives

To successfully complete this project various milestones and objectives will need to be completed.

1. Establish which household attributes are strongly associated with EV ownership through a comprehensive review of prior studies.
2. Segment the households of Melbourne and Geelong into like groups.

3. Identify the group whose household attributes are most representative of the researched EV ownership profile.
4. Obtain reliable data that contains the number of EVs in Melbourne and Geelong. Preferably including the geospatial distribution of these EVs.
5. Create a comprehensive dataset that combines household attributes and the number of EVs in Melbourne and Geelong.
6. Develop a robust model that can effectively utilise both EV data and household attributes to accurately predict which households in Melbourne and Geelong will own EVs.
7. The final objective is to be able to take a synthetic population of Melbourne and Geelong's households and assign EV ownership to the most likely households.

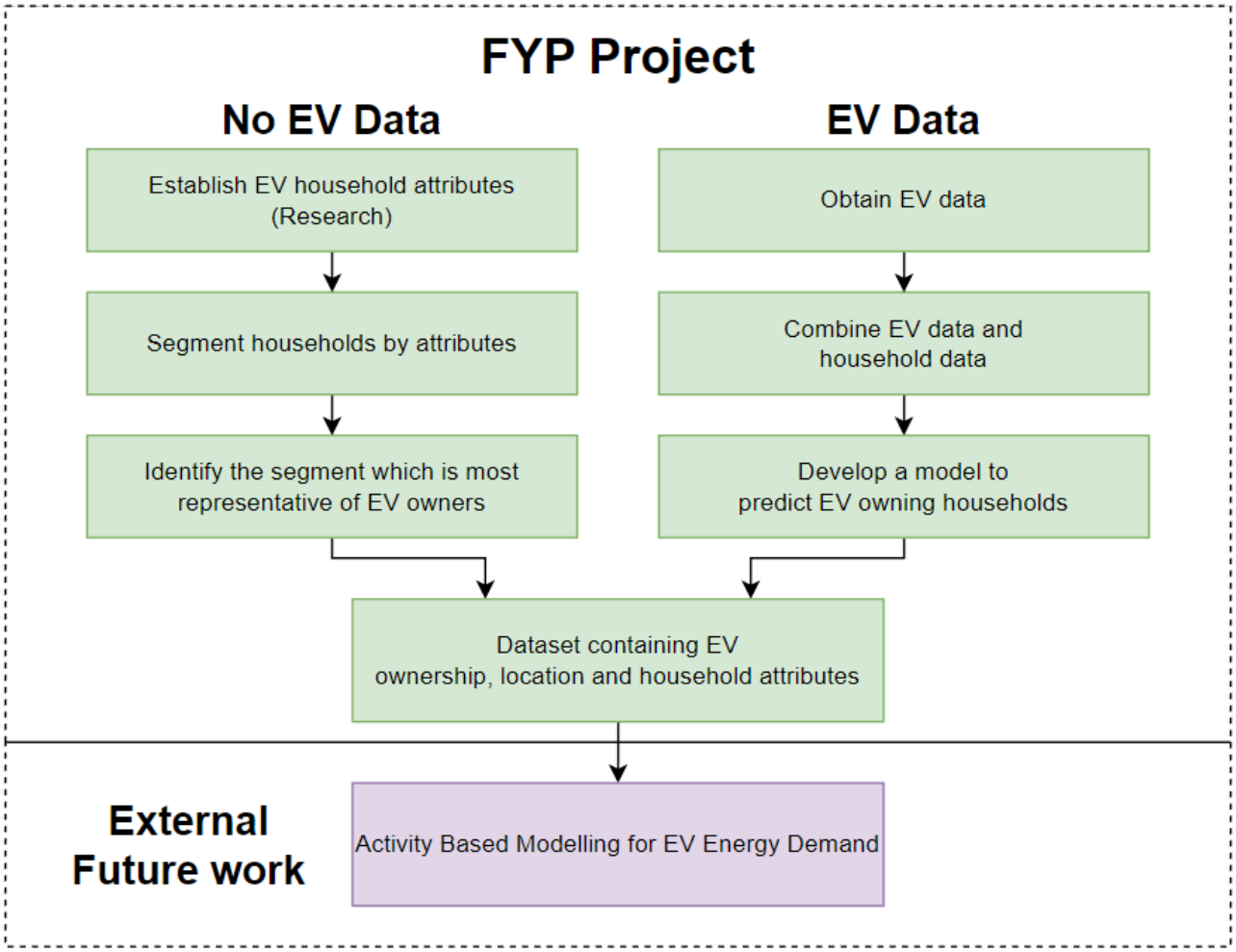


Figure 3.1: Project Objectives

4 Literature Review

To create models that represent the EV population and explore their habits there is a large field of research to draw upon. The first step in understanding EV usage in Melbourne is having a way to represent the population. Unfortunately, due to privacy concerns, survey limitations and cost constraints having a full representation of a population is often not possible (Kirill, 2011). Government collected data such as censuses are often not available in full, rather a "public-use samples" is released where data is rounded, aggregated, removed and collected infrequently (Kirill, 2011). Moreover, this data comes with other issues such as not containing specifics on vehicle power source (E.g., electric, diesel and petrol). Often the data is aggregated to a location or has had identifying features removed preventing links between data points being observed. Privacy concern is of particular importance when the data is about where people live, where they travel, when they travel and their household demographics (Ossama & Virginia, 2019). This information is essential for creating travel itineraries that can be used to model energy demand. To combat the issue of aggregated data, population synthesis can be employed to disaggregate this data. Moreover, population synthesis can be used to generate a more granular and comprehensive representation of Melbourne's EV drivers.

Population synthesis combines several different pieces of data to try and give a fuller representation of a population. Its primary goal is to disaggregate the data while maintaining the marginal distribution of the data (Kirill, 2011). For the example of aggregated household data, population synthesis would produce individual households with their own attributes while maintaining the total number of households and number of each attribute. Recent years has seen a rise in population synthesis for several reasons including, improved computational power, greater storage and more data about individuals (Harland et al., 2012). Population synthesis can make use of travel diary surveys, land use data, and census data to create an EV representative population (Ossama & Virginia, 2019). Population synthesis is a powerful tool that activity based-modelling can utilise to simulate a population (Harland et al., 2012).

Activity-based modelling allows a population to be studied in greater detail. Activity-based models (ABMs) are used extensively in both private and public sectors to simulate agents movements (Both et al., 2021). They are often used to model transport demand and congestion but have also been used to model EVs energy demand. An ABM was used in Flanders, Belgium to predict EV energy usage based on location, time and trip purposes of individuals (Knapen et al., 2012). The study used the Feathers activity-based modelling to predict travel schedules (Bellemans et al., 2010). The Feathers model was not used to determine vehicle type, rather this was determined from the predicted travel schedule using a Bayesian network (Knapen et al., 2012). To accurately represent a population through simulation it is essential to assign agents attributes that affect their behaviour. Agents demographic information such as age are associated with transport mode choice and consequently their travel behaviour (Both et al., 2021). ABMs can incorporate logit and nlogit models to predict mode choice(Both et al., 2021). When creating a model of Melbourne's EVs electricity demand there is three key steps that have been outlined, generate agents demographics, assign activity patterns, and assign locations to activities (Wang et al., 2021). Furthermore, research has shown that a transportation model of Melbourne could be developed with just publicly available data from the Victorian Integrated Survey for Transportation (VISTA) and the Australian Bureau of Statistics (ABS) (Both et al., 2021). It can then be inferred that an EV specific model could be produced if given additional data such as EV registrations. EV trip usage can be determined by using a decision tree (Knapen et al., 2012). For example, if the trip is within the EV's range the household will use an EV. The study in Flanders uses a decision tree that incorporated the following questions: is it a work trip, is the vehicle privately own or a work vehicle and can the car be charged at work. ABMs are capable of predicting energy demand for every minute of the day for thousands of area zones (Knapen et al., 2012).

EV ownership modelling is an emerging field of study (Ma & Ye, 2019). There are several studies that attempt to capture EV ownership. However, studies in EV ownership are not easily compared between countries due

to differences in consumer attitudes and government policies (Hjorthol, 2013). The EV market is rapidly growing, predicted to be 130 million EVs worldwide in 2050 (LaMonaca & Ryan, 2022). Past studies have identified men aged 30-50, with families, high educations, and high income to be the most likely to buy an EV (Hjorthol, 2013; Plötz et al., 2014). This is in part because of their high socio-economic status enables it and they own a large share of vehicles (Plötz et al., 2014). As EVs are an emerging market the diffusion of technology is greatly dependent on "early adopters". If a technology is embraced early on by a large group of consumers, the technology is far more likely to succeed (Rogers, 2003). Research suggests that for government policies to be efficient and effective in increasing EV adoption they need to align with the needs of early adopters (Plötz et al., 2014). This highlights the importance of understanding EV ownership characteristics to implement the most impactful policies. Adoption enticing policies examples can be seen in Norway such as free parking, driving in bus lanes, free driving on toll roads, reduced vehicle tax and reduced company tax (Hjorthol, 2013).

EV owners often travel a significant number of kilometres and cite work commutes as the primary reason for owning an EV (Hjorthol, 2013). Although EVs are more expensive than ICE vehicles, they have lower running cost which compensates for the high upfront cost (Plötz et al., 2014). EVs limited range of approximately 345km (*Range of full electric vehicles*, 2023) and the resulting "range anxiety" is often seen as a drawback to EVs. However, amongst EV owners, driving range is often not cited as an issue. This is because driving range to work is a prerequisite for purchasing an EV and this is their primary function for the vehicle (Hjorthol, 2013). Furthermore, practical experience in driving an EV is likely to reduce anxiety and increase likelihood of purchase. Despite reduced range anxiety due to experience EV drivers prefer not to use their EV for leisure activities as this requires more planning (Hjorthol, 2013). Moreover, Pluggable Hybrid-Electric Vehicles (PHEV) have been shown to use more electricity than Battery Electric Vehicles (BEV) despite BEVs only source of power being electricity. This is because PHEV are able to exhaust the full range of the EV whereas BEVs must drive less than the anxiety-reduced range (Knapen et al., 2012).

The availability of chargers have been identified as a key factor for EV adoption (LaMonaca & Ryan, 2022). Since EVs typically require longer charging times compared to ICE vehicle refuelling and have a reduced range, access to charging infrastructure can significantly affect ownership experience. EV ownership requires that drivers have access to both public and home charging infrastructure in order to feel confident in their transition (LaMonaca & Ryan, 2022). However, EV drivers report they use public charging infrequently. This is due to chargers being unreliable, difficult to locate, sometimes reserved for professional fleets and poorly maintained (Hjorthol, 2013). There are three main categories of chargers each with their own benefits. Home chargers typically use level 1 (120V 3.3kW AC) and level 2 (240V 7.4 – 22 kW AC) chargers and make charging overnight easy, however they are often not as quick as public EV chargers. Level 3 charger also referred to as DC Fast Chargers (CDFC) are typically found at commercial charging locations (LaMonaca & Ryan, 2022). When comparing charging times for 100km, the three different levels vary greatly. Level 1 can be as slow as 8 hours while DCFC can be as quick as 10 minutes. Understanding charging times, locations and battery range are essential factors for modelling energy usage of EV owners.

Currently there is no dataset publicly available that gives a breakdown of the number of EVs in Melbourne and the spatial distribution of these EVs. For this reason, it is essential for this project to have a way to estimate the number of EVs in Melbourne as well as where they are located. Prior studies into early adopters of EV vehicles have demonstrated that clustering can be used to identify EV owners from their attributes. A study in Birmingham used hierarchical clustering to identify the geographic distribution of individuals who fit the profile of an EV owner (Campbell et al., 2012). The study used a literature review to identify six attributes (age, homeowner, detach/semi-detached house, drives a car to work, car/van ownership and socio-economic status) strongly associated with EV owners and used these features to cluster the population. Through further research the following attributes have been determined to have a significant positive relationship with EV ownership. High education (Bjerkan et al., 2016), high income (Qian & Soopramanien, 2011), two or more vehicles (Peters & Dütschke, 2014). Furthermore, a survey of 1,257 EV owners in Maryland was conducted

to determine EV ownership attributes which additionally found links between age, household size and number of children (Noyce David, 2019). Understanding the relationship between these attributes will enable clustering of households to identify EV owners.

Hierarchical clustering is just one method of unsupervised learning that can be used to cluster data. To produce the most accurate clusters it is important to experiment with more than one clustering algorithm. A more common method is K-means clustering which clusters data into k groups by minimising the distance between data points and their assigned cluster (Bonaccorso, 2018). An extension to this method is Gaussian Mixed Model (GMM) which uses the a variance matrix to produce more complex decision boundaries (Patel & Kushwaha, 2020). To determine the ideal number of household clusters the Elbow Rule can be utilised. This method involves plotting the inertia (the summation of the squared distances of each point to its centroid) against the number of clusters (Bonaccorso, 2018). The Elbow is the point in which the inertia gradient significantly changes identifying an ideal combination of reduced inertia, while avoiding over fitting.

To improve the accuracy of household clustering, pre-processing the data is essential. Removing outliers and standardising values to a specific range is essential as clustering algorithms are dependent on distances between features in the feature space (Patel & Mehta, 2011). Standardising to a normal distribution is a common technique however, its accuracy depends on the underlying distribution of the data and selecting an appropriate transformation. If the data has a multinomial or a Poisson's distribution it can be more appropriate to use a quantization transformation (Developers, 2023). Furthermore, to compare the performance between clustering methods and cluster sizes, a comparison metric is required. The Silhouette Index is an internal clustering index that will enable this project to identify the best model to identify EV owning households.

5 Methodology and Methods

5.1 Methodology

This project will use a quantitative research methodology as it seeks to understand the relationship between household attributes and EV ownership. The project has developed two models that can be used depending on data availability. In the absence of EV data, a model utilising unsupervised learning specifically clustering has been developed. The second model has been developed to incorporate EV data for improved model predicting. The model developed in the presence of EV data utilises machine learning, specifically a multi-layer perception (MLP) network. Both models will predict which households in Melbourne and Geelong will own EVs, thus resulting in a dataset containing all households in Melbourne and Geelong, their household attributes and if they own an EV.

The project relied on a parallel external project for the generation of a synthetic population. Furthermore, the external project has undertaken Pearson Feature Selection to explore the correlation between EV ownership and household attributes. The overarching structure of the project is shown in Figure 5.1.

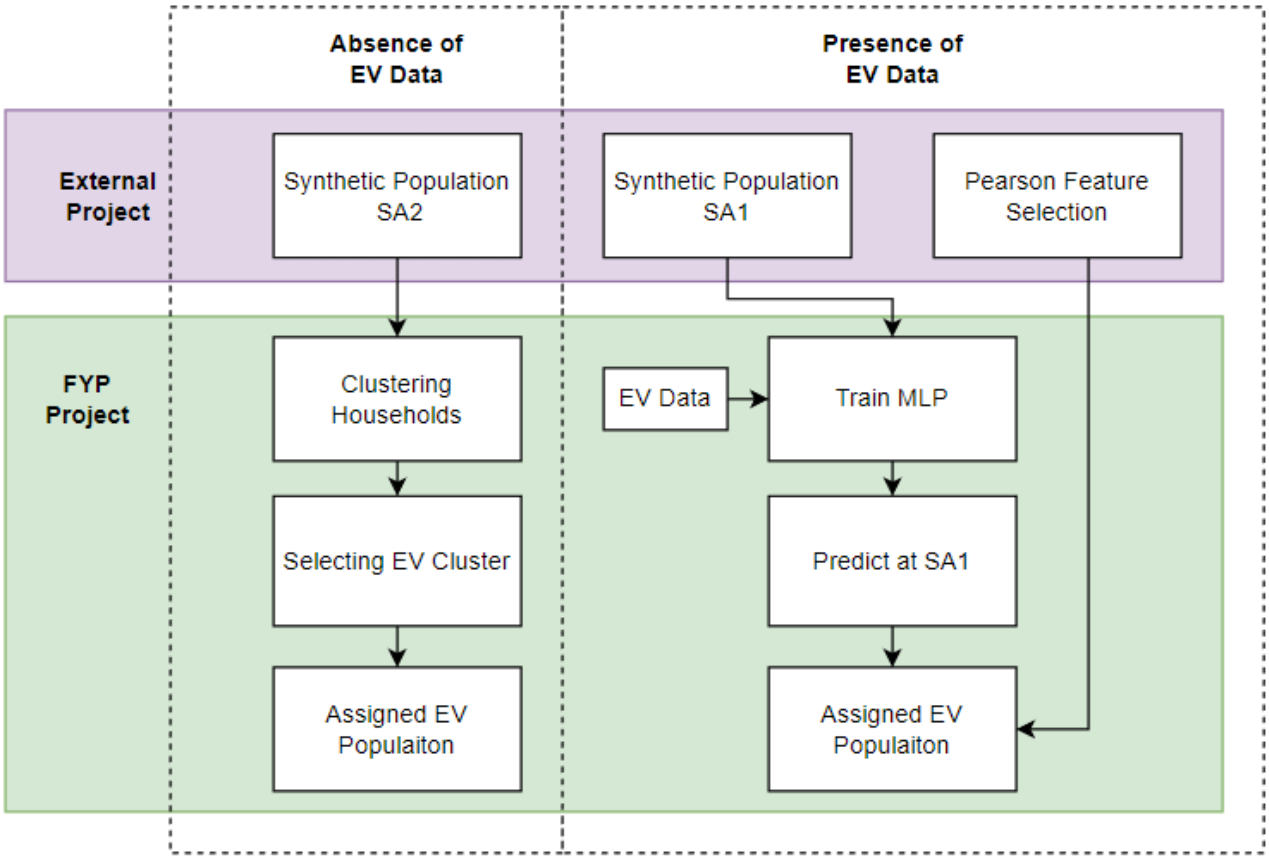


Figure 5.1: Project Flow

5.2 Method

5.2.1 Method 1: Absence of EV Data

Data preparation

Firstly, the synthetic population was cleaned to remove missing datapoints, handle outliers and transform variable distributions. The data was generated for all households in Victoria; however, this project is only investigating households in Melbourne and Geelong. For this reason, the dataset was reduced based on each households Statistical Area 2 (SA2) code. When inspecting the data there were several households which did not record their income either through error or refusal to. As there is no clear way to impute these missing values and they only accounted for 0.02% of the data, households with missing attributes were removed from the dataset.

Handling Outliers

Before clustering it is important to handle outliers to prevent cluster centres shifting to capture outlying results. Due to the nature of income the distribution is right skewed, meaning that there is a small number of high-income households and many middle to lower-income households. This difference means that clustering that involves income will have its cluster centres shifted to minimise loss when capturing the impact of the 1% of income. This issue can be solved by several approaches such as removing outliers using the inter quartile method, trimming or Winsorizing. Using the interquartile range provides a statistical basis for removing outliers from datasets in which recording errors are possible. Similarly, using trimming indiscriminately removes values beyond a certain threshold. However, removing incomes by thresholding is inappropriate as this project is interested in high income houses, as they are likely to own EVs (Bjerkan et al., 2016). Winsorizing was used as this method transforms the outlying data points instead of removing them. Hence maintaining the statistical information of these data points. Winsorizing involved moving data points

in the top and bottom 1% to the top and bottom 1%. By doing this all households are maintained in the population while reducing their effect on the distribution.

Redistribution of Data

K-means clustering works by minimising the distance between data points and their assigned cluster. For this reason, all attributes needed to be on the same scale. For example, income data is in the thousands whereas total vehicles owned is usually single digit. If these attributes were used to cluster the data, the vehicle count would be insignificant in affecting the distance to the cluster and all clusters would be created primarily based on income. Hence, this project mapped all numeric data between 0 to 1.

Due to the skewness of the income data, it was redistributed to a uniformed distribution. This was done using a quantile transformation. Quantile transformations divide the data into intervals so that each interval has the same number of datapoints. These intervals are referred to as quantiles. Each quantile is given an index then these indices are scaled between 0 to 1, creating a uniformed distribution.

Clustering Data

The project experimented with three different clustering methods: K-means, Gaussian Mixture Model (GMM) and Hierarchical Clustering. Before using these models, two different feature spaces were created. The following attributes have been identified in the literature review and are contained in the synthetic population dataset as seen in Table 5.1.

Table 5.1: Dataset Variables

Feature	Variable Type
Household Size	Integer
Total Vehicles	Integer
Household Income	Integer
Dwelling Type	Binary
Dwelling Ownership	Binary
Age Profile*	Binary

> ***Note:*** *Age profile was created to capture the effect age has on EV ownership. Households who have a member older than 30 and have less than three children, are more likely to own an EV. If both these conditions were met, Age Profile is assigned to 1.*

This project experimented with clustering based on a Boolean dataset and a continuous/discrete dataset. In the Boolean dataset each households' attributes were set to a 0 or 1 based on if they met the EV driver profile established in research. For instance, a high-income household's income value would be set to a 1 and 0 if low income. This method was experimented with based on the case study in Birmingham which used this approach (Campbell et al., 2012). Clustering was also done using the continuous data. Results between clustering methods were compared using a Silhouette score.

Cluster Selection

Selecting the EV household cluster was achieved by inspecting the cluster that best fit the research criteria. This was made possible by increasing the number of clusters until an EV cluster could be identified.

5.2.2 Method 2: Presence of EV Data

Data Preparation

The EV data was provided by the Department of Transport and Planning in the form of vehicle registration data. Because this data is structured at the POA level whereas the synthetic population is structured at the SA1 level. It is necessary to develop a mapping between these two structures. The POA structure is a non-

ABS structure and is typically significantly larger than SA1 regions which are an ABS structure. Mapping between ABS structures is simpler because they share common boundaries and furthermore statistical areas are subsets of each other, with SA1 being the smallest region. An example of the SA1 code containing Monash University is given in Table 5.2. Furthermore, the boundary types used in this project ate shown in Table 5.3.

Table 5.2: Example SA1 21205156702

State/Territory	SA4	SA3	SA2	SA1
2	12	05	1567	02

Table 5.3: Boundary Types

Boundary Type	Structure
2021 Statistical Area Level 1 (SA1)	ABS Structure
2021 Statistical Area Level 2 (SA3)	ABS Structure
2021 Statistical Area Level 3 (SA3)	ABS Structure
2021 Statistical Area Level 4 (SA4)	ABS Structure
2021 Significant Urban Area (SAU)	ABS Structure
2021 Postal Area (POA)	Non-ABS Structure
2021 Local Government Area (LGA)	Non-ABS Structure

Significant Urban Areas were used to select the region encompassing Melbourne and Geelong. Subsequently, this area was then used to select for SA1, POA and LGA regions within the Melbourne/Geelong vicinity. Each SA1 was assign to a POA based on the area of intersection with the POA. If a SA1 intersected multiple POA, it was given to the POA it had the largest intersection with. This was also done to map SA1 to LGA, which is necessary as LGA is a non-ABS structure.

Training MLP

To predict the number of EVs in each regions an MLP was developed. The model was trained on the department of transport EV data and the synthetic population. To develop a labelled dataset, households in the synthetic population were aggregated by POA. Households were aggregated at the POA level so that EV data also aggregated at POA could be linked to the household responsible for the total EVs in each POA. To accurately represent the population a histogram of each household's attributes was produced. Histograms were chosen as the independent variable as they can capture more information about households in each region that other options. Options such as measures off central tendencies (mean, median or mode) cannot encapsulate the uniqueness of each region. A histogram was generated for each attribute and appended to form a single array of length 42 to be used as feature data. The training data consisted of the total number of EVs in each POA and an associated array.

Table 5.4: Attribute Bins

Attribute	Number of Bins
Household Income	21
Vehicles	5
Household Size	7
Dwelling Type	4
Dwelling Ownership	5
Total	**42**

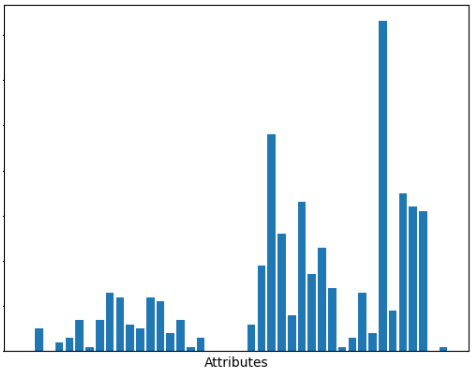


Figure 5.2: Histogram of Attributes

Once the feature space was created, the MLP architecture could be design. The architecture was developed through an iterative process of adjusting the MLP structure and fine tuning hyperparameters after each training run of the model. The dataset was split into a training set, validation set and test set consisting of 72%, 18% and 10% of the data respectively. The breakdown is shown in Figure 5.3.

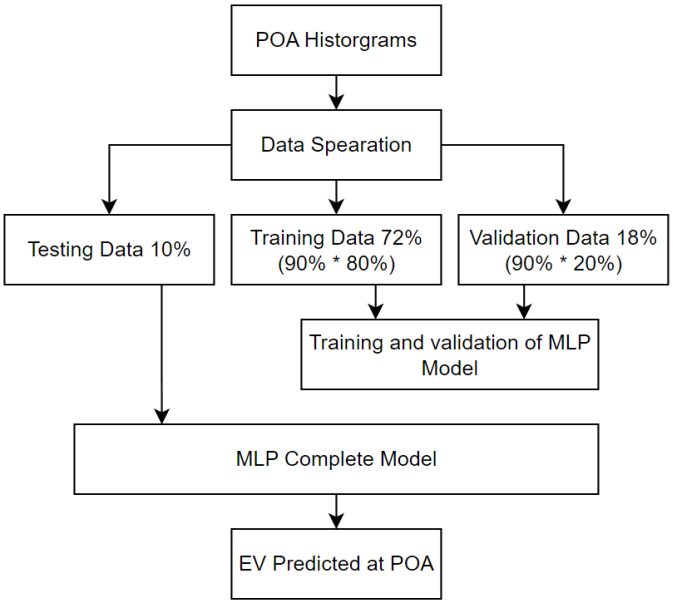


Figure 5.3: MLP Training

Using MLP

The trained MLP can then be used to predict the number of EVs at different geospatial levels. By creating histograms at the SA1 level the model can predict the total number of EVs expected in each SA1. Figure 5.4 depicts the process of predicting at SA1; however, the model can take histograms generated at any geospatial level to predict directly at that geospatial level.

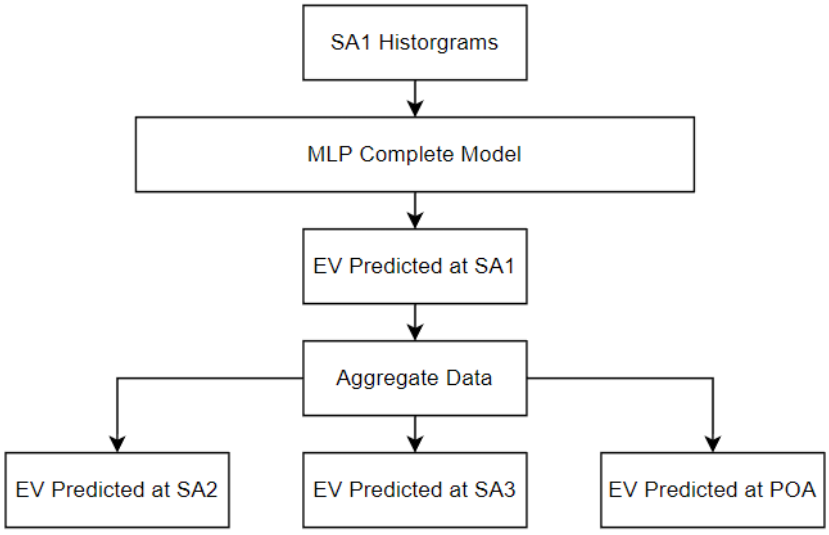


Figure 5.4: Predicting with MLP

Truncate, Replicate and Sample

The MLP predicts a continuous number of EVs for each region. This becomes a problem when trying to assign a non-integer number of EVs to the most likely households in the region. The naive approach is to round the non-integer number to the close's integer. By doing this there is no guarantee that the number of predicted

EVs in the whole population of Melbourne and Geelong will be preserved. This issue is more apparent when predicted at SA1 level as EV predictions are small and thus more impacted by rounding.

The Truncate, Replicate and Sample (TRS) method was used to preserve the total number of EVs in the population. The method truncates the integer part of the prediction and then stores both the integer and fractional parts separately. Using the difference between the truncated total and the desired total, the discrepancy is probabilistically redistributed to the truncated predictions. Each truncated value has a small probability of having 1 added to it. This probability is a function of the fractional part removed from the prediction (Lovelace & Ballas, 2013).

Feature Selection

Feature Selection was used to assign the predicted number of EVs to the most likely households. The correlation between each household attribute and the number of EVs was conducted by the external project. A table consisting of each attribute, its correlation and the p-value was then used by this project to rank each household amongst the other households in its SA1 region. The ranking was based on the summation of a household's attributes and the associated correlation coefficient. To further explain this method Table 5.5, Table 5.6 and Table 5.7 will be used.

Table 5.5: Example Correlation Coefficients

Attribute	Correlation Coefficient
Income $4,000-$4,499	0.4
Income $3,000-$3,499	0.3
3 people in household	0.25
2 people in household	0.2

Table 5.6: Example Household 1

	Income = $4,500	Household Size = 2	Summation
Coefficients	0.4	0.2	0.6

Table 5.7: Example Household 2

	Income = $3,500	Household Size = 3	Summation
Coefficient	0.3	0.25	0.55

Based on each household attributes and the correlation coefficients. Household 1 would be ranked higher than household 2 because of its greater summation, and thus would be more likely to own an EV.

6 Results and Discussion

6.1 Evaluating Method 1: Absence of EV Data

6.1.1 Clustering Synthetic Population

To successfully find an EV cluster within the population several different feature spaces and different clustering algorithms were experimented with. Two feature spaces were constructed, one consisting of numeric and binary variables and the other consisted entirely of binary variables. The transformations used to create the first feature space is shown in Table 6.1.

Table 6.1: Continuous Feature Space

Feature	Initial Distribution	Transformation
Household Size	Integer	Winsorize outlier and scale to [0,1]
Total Vehicles	Integer	Winsorize outlier and scale to [0,1]
Household Income	Poisson distribution	Quantize and scale to [0,1]
Dwelling Type	Categorical	Convert to one-hot encoding
Dwelling Ownership	Categorical	Convert to one-hot encoding
Age Profile	Boolean	Person aged ≥ 30 and Number of Children ≤ 2

The second feature space built upon the first feature space. Conditional tests were used to transform continuous variables into binary variables, creating a binary feature space. The conditional tests are outlined in Table 6.2.

Table 6.2: Binary Feature Space

Feature	Condition for Boolean to be assigned 1
Household Size	Household Size ≥ 2
Total Vehicles	Total vehicles ≥ 2
Household Income	Household Income ≥ $3500
Dwelling Type	Separate House = 1
Dwelling Ownership	Dwelling Being Purchased or Fully Owned
Age Profile	At least person aged ≥30 and Number of Children ≤ 2

Selecting cluster size was a difficult process for several reasons. Firstly, there is no ground truth available to validate the results of the clustering. This is because clustering is performed on unlabelled data. Additionally, clustering size is subjective, depending heavily on the data being clustered. When clustering the synthetic population there is no clear way to determine how large each cluster should be. Moreover, when determining the number of clusters there is a balancing act between overfitting and underfitting the data. To address this issue a large degree of experimentation with both feature spaces, clustering algorithms, and the number of clusters was performed. The experimentation of clustering is shown in Figure 6.1.

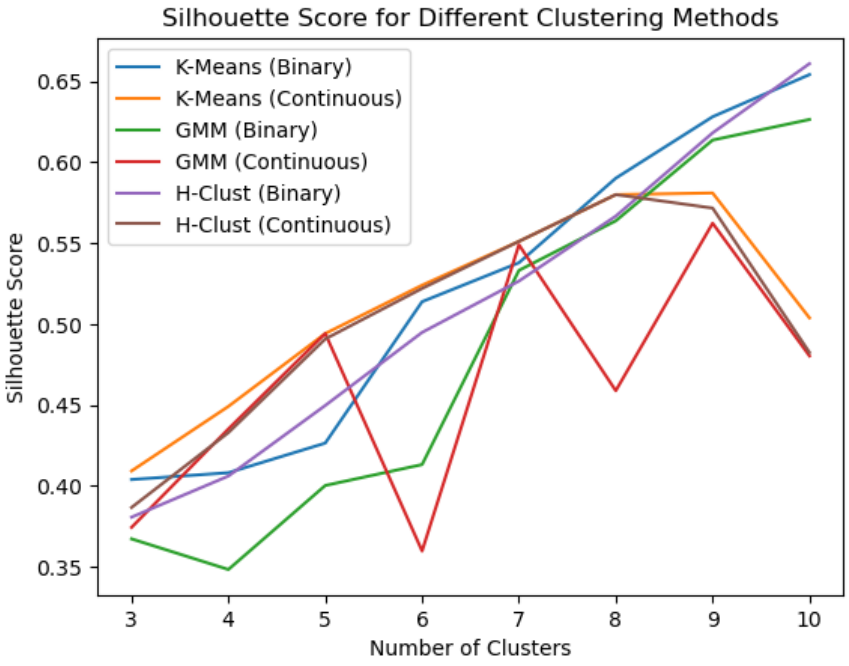


Figure 6.1: Silhouette Score

Both feature spaces were clustered by K-means, GMM and Hierarchical clustering at 3 to 10 clusters. This was done to narrow down the possible combinations of algorithms allowing for a more focused investigation of the most promising clustering algorithm. As seen in Figure 6.1 the Silhouette Score increased smoothly for the binary dataset from 3-10 clusters, however, the continuous datasets were unstable and showed a decrease in Silhouette Score at 10 clusters. This could be a result of the continuous data providing inadequate separation between clusters formed in the dataset (Nidheesh et al., 2020). The binary dataset gave the best Silhouette Score regardless of clustering algorithm. When comparing the three algorithms on the binary feature space, all produced similar results at 9 and 10 clusters. Because of the comparable outcomes, the K-means method was chosen as it is more computationally efficient resulting in faster clustering times. Furthermore, the binary dataset was chosen due to its consistency throughout the testing process.

1.1.1 Selecting Cluster

To test how well the binary K-means method identified a realistic EV cluster, several different cluster sizes were explored, with 10 clusters producing the best results. The quality of the result was based on if the clustering produced a cluster that fit the profile of an EV owner. The clustering of the binary dataset using 10 clusters and the K-means algorithm is shown in Table 6.3. The table displays the average value of each variable in a particular cluster. Cluster 8 was selected to be the EV owning cluster due to its significant representation of household's attributes associated with EV ownership. Specifically, the percentage of households having each attribute are as follows: two or more vehicles 100%, high income 100%, live in separate house 92.3%, are purchasing or fully own their dwelling 92%, fit the age profile 74.4% and have a household of two or more members 100%. Of the 1,676,258 households used in this model of Melbourne and Geelong, the model predicted 148,394 EVs which counts for approximately 8.85% of the population.

Table 6.3: Identified EV Cluster

Cluster	Vehicles	Income	Separate House	Purchasing/Fully Own Dwelling	Age Profile	Household Size	Cluster Size %
0	0.258781	0.015372	1.000000	0.158003	0.196541	0.828019	9.589813
1	1.000000	0.000000	0.905521	0.925067	0.833851	1.000000	39.884612
2	0.032210	0.000000	0.000000	0.000000	0.000000	0.000000	5.334024
3	0.064937	0.000000	0.625507	1.000000	0.000000	0.000000	14.602525
4	0.380653	0.094580	0.000000	0.000000	1.000000	1.000000	4.758396
5	0.000000	1.000000	0.742745	0.805235	0.937840	1.000000	1.447271
6	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000	7.861737
7	0.277819	0.057866	0.000000	0.000000	0.000000	1.000000	7.861737
8	**1.000000**	**1.000000**	**0.927564**	**0.920057**	**0.744127**	**1.000000**	**8.852695**
9	0.000000	0.000000	0.000000	1.000000	0.756161	1.000000	2.849263

1.1.2 Visualising Data

The results of the EV owning cluster are displayed geospatially using a map of Melbourne and Geelong. Figure 6.2 shows the total number of EVs predicted in each SA2 region. The highest number of EVs predicted are in East Melbourne and small parts of North Melbourne. The SA2 regions with the highest EV rates are shown in the Table 6.4 and the lowest in Table 6.5.

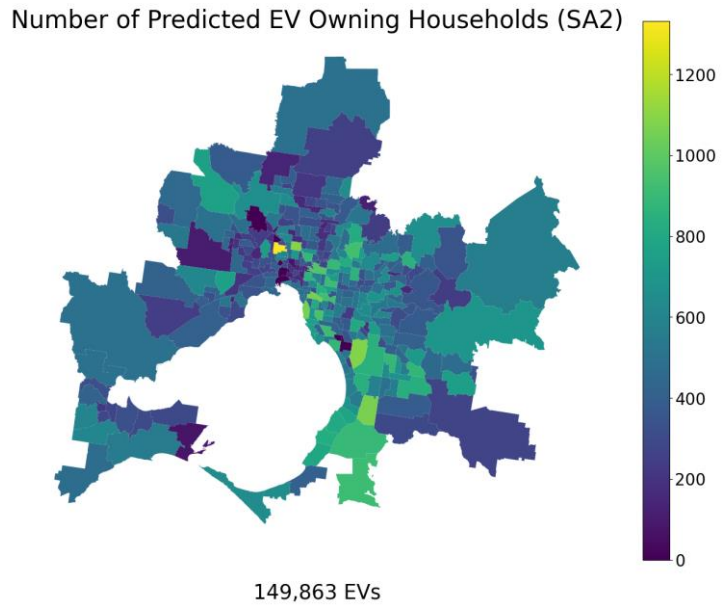


Figure 6.2: Predicted EV Owning Households

Table 6.4: Highest Number of EVs

SA2 Code	Name	Number of EVs	Number of Households
206031114	Essendon - Aberfeldie	1332	10316
206011108	Coburg	1088	9610
212041314	Keysborough	1086	7830

Table 6.5: Lowest Number of EVs

SA2 Code	Name	EV Count	Number of Households
203031052	Point Lonsdale - Queenscliff	73	1638
206041124	Parkville	75	1997
210031439	Gowanbrae	87	1022

These results are consistent with the initial perception of the EV distribution in Melbourne and Geelong based on income distribution. Melbourne is wealthier in the Eastern suburbs then the Western suburbs. Furthermore, the EV data distribution looks very similar to the high-income distribution map shown in Figure 6.3Figure 6.3: High Income Households. This is in part because the selection of the EV cluster was conditional on high income. However, it was also conditional on 5 other factors that needed to be met in some capacity, so it is surprising that the EV prediction is so similar the high-income distribution. Comparing the distribution more carefully the similarities become inconsistent in the inner city.

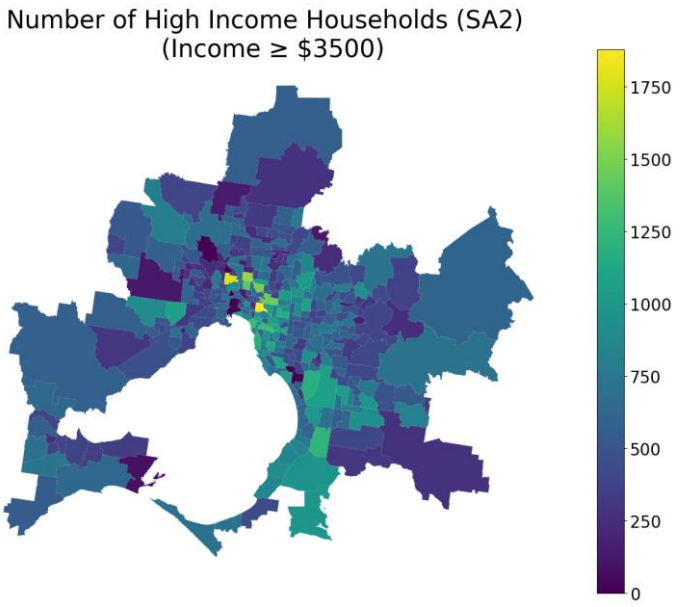


Figure 6.3: High Income Households

The differences in the inner city may be explained by the differences in dwelling type and transportation options. The inner-city is primarily composed of apartments and townhouses, both these dwelling types increase the difficulty of charging an EV at one's residency and therefore decreases the likelihood of the household owning an EV. Furthermore, the city has a greater density of workplaces, schools, and shops. This coupled with better public transport and reduced parking results in households in Melbourne's inner suburbs being less likely to own several cars. This is confirmed by Figure 6.4, that illustrates the average number of vehicles owned by households. The average number of vehicles increases as households become further away from the inner city.

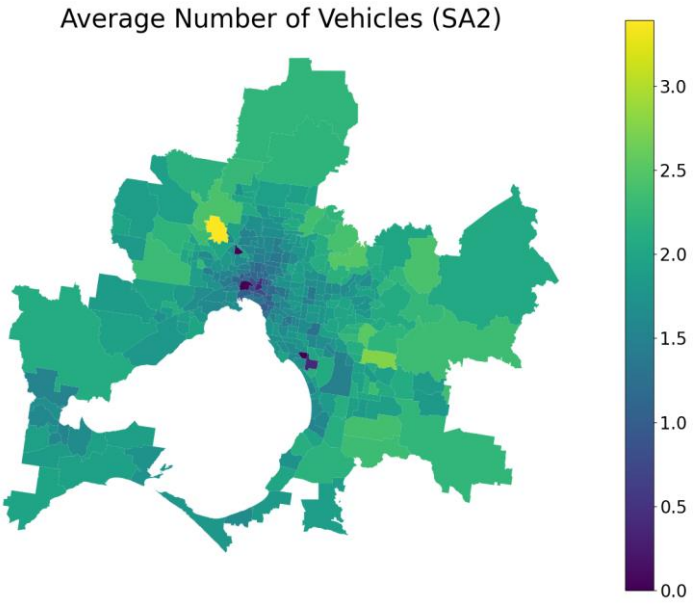


Figure 6.4: Average Number of Vehicles

Exploring EVs as a Percentage of Households

Visualising the number of households predicted to own EVs as a percentage of total households in each SA2 areas gives a significantly different picture as seen in Figure 6.5. By standardising the population of each SA2 area it allows for a better comparison between areas. When inspecting the percentage of households owning

EVs it becomes apparent that this does not reflect the real world. The model predicting 8.85% of vehicles being EVs is not a reflection of the real world as current EV sales in 2022 only accounted for 3.8% of new vehicle sales in Australia (Whitehead, 2023). One reason for the large over estimation may be that the model is predicting EV ownership solely based on the attributes of the household and cannot account for household preferences. For example, a household with a high income, who own their dwelling and has a detached home will have the means typically required to purchase an EV. However, they may still decide to buy an ICE vehicle out of personal preference. The model is limited to predicting which households have the means to own an EV rather than which households do own EVs. Although EV penetration in some SA2 areas being over 17% is currently unreasonable it is within the scope of possibility as evident by Norway's EV sales accounting for 72% of new car sales (Sievewright, 2022). Given enough time the households predicted to own EVs may purchase one.

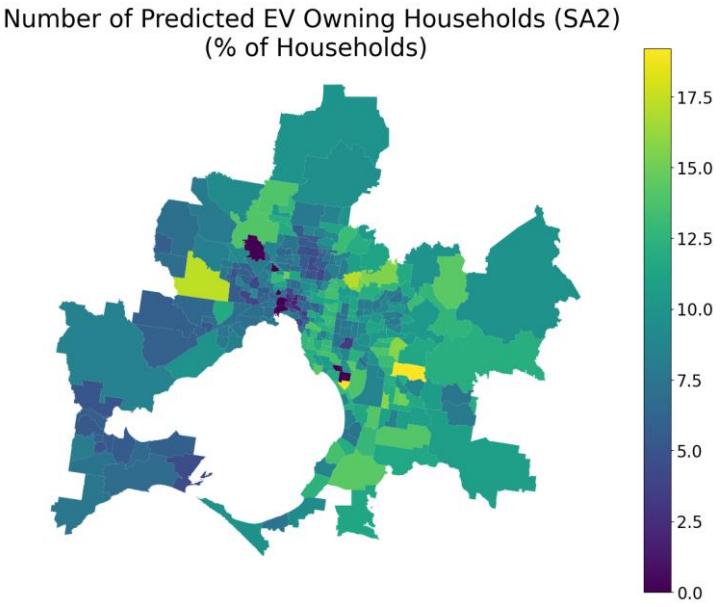


Figure 6.5: Percentage of EV Owning Households

1.1.3 Logistic Regression

Logistic regression was performed on the labelled dataset to test the significance and direction of each household attribute on EV ownership. Furthermore, understanding the relationship between household attributes and EV ownership within the synthetic population will be important when modelling the behaviour of EV households. The logistic regression was performed on the synthetic labelled data using the pre-transformed variable values. The statistical significance of the model has been reported in Table 6.6. All coefficients are positive except household size and age profile which were negative. Furthermore, all variables were significant at 0.05 significance level which was expected given the clustering and EV assignment was based on these variables.

Table 6.6: Logistic Regression all variables

	coef	std err	z	P>|z|	[0.025	0.975]
constant	-17.8837	0.052	-344.031	0	-17.986	-17.782
Income	0.0039	1.15E-05	336.733	0	0.004	0.004
Household Size	-0.1741	0.004	-45.368	0	-0.182	-0.167
Total Vehicles	0.9261	0.006	157.639	0	0.915	0.938
Dwelling Purchasing/Own	1.9869	0.019	106.84	0	1.95	2.023
Separate House	1.0887	0.017	64.02	0	1.055	1.122
Age profile	-0.8304	0.015	-57.076	0	-0.859	-0.802

The positive coefficients of the model indicate that for every $1 increase in weekly household income the household log odds of EV ownership will increase by 0.0039. For each additional vehicle owned, log odds of EV ownership increase by 0.9261. If the household fully owns or is purchasing the dwelling, the log odds of EV ownership increase by 1.9869 and owning a separate/detached house increases the log odds of EV ownership by 1.0887.

However, the coefficients of household size and age profile being negative was unexpected. This is because age profile and household size being 1 was a requirement when assigning the EV cluster so a strong positive relation should have been expected. This model indicated for each additional household member log odds of EV ownership decrease by 0.1741 and if age profile is 1, log odds of EV ownership decrease by 0.8304.

An additional logistic regression model was created using only household size and age profile to test their relationship on EV ownership in isolation. This regression gave positive coefficients for both as shown in Table 5.7. In the context of the other variables household size and age profile direction changes potentially indicating a weaker predictive power when identifying EV ownership.

Table 6.7: Logistic Regression Household Size and Age Profile

	coef	std err	z	P>|z|	[0.025	0.975]
Constant	-5.3675	0.012	-465.901	0	-5.39	-5.345
Household Size	0.6195	0.002	359.144	0	0.616	0.623
Age profile	1.2144	0.008	155.954	0	1.199	1.23

6.2 Evaluating Method 2: Presence of EV Data

6.2.1 Mapping the synthetic population

The area defined by the SUA of Melbourne and Geelong was successfully used to select households of interest, as illustrated in the shape of the two plots below. Furthermore, the mapping between SA1 to POA, as well as between SA1 to LGA are also represented. This mapping is vital for defining which household, recorded by SA1, belong to which POA, thus enabling the creation of the MLP's training and prediction dataset. The colours in Figure 6.6 represent the POA code assigned to each SA1, while Figure 6.7 represents the LGA code. There is a total of SA1 1,1947 SA1 regions, 285 POA regions and 40 LGA. A further break down of the population is show in Table 6.8: Melbourne and Geelong Population Table 6.8.

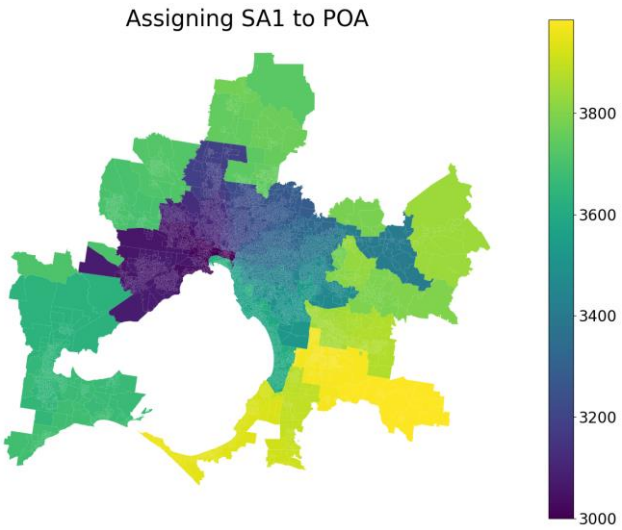


Figure 6.6: Mapping between SA1 and POA

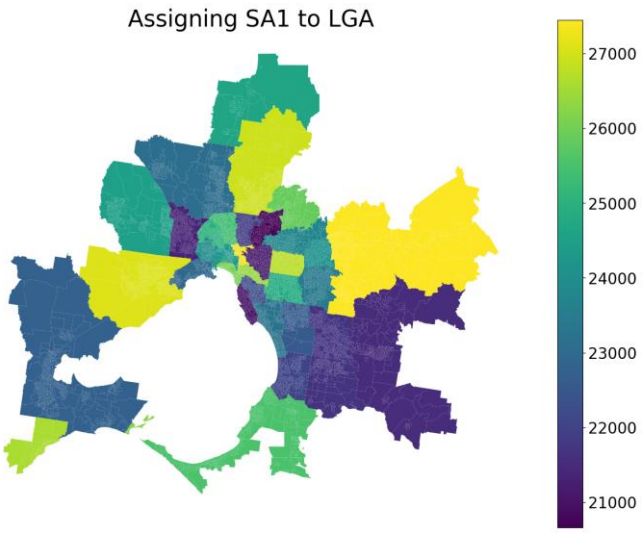


Figure 6.7: Mapping between SA1 and LGA

Table 6.8: Melbourne and Geelong Population

Population Statistics	
Number of Households	1,856,054
Number of People	4,931,785
Total number of household vehicles	3,157,081
Average Income	$1,823
Number of EVs	6,560

6.2.2 EV Data

EV data was obtained in the form of registration data from the Department of Transportation and Planning. The dataset required significant preparation before the number of EVs in each suburb could be obtained. The registration dataset contain vehicle make, registration year, total number of vehicles registered and engine type. Firstly, the engine type was used to filter out all none EVs. After doing so it was detected that some entries in the dataset register multiple EVs in some cases more than 100. This highlighted the issue that the dataset contains both EVs registered by households as well as EVs registered by companies. To solve this issue any registration that included more than 2 vehicles were removed from the dataset. Furthermore, to increase the relevance and to dilute the presence of EVs not registered by households, only registrations from the past 5 years were included. This decision was made based on the trend in vehicle sales that show that EVs sales as a proportion of total vehicle sales has increased by approximately 40 times in the past 5 years (Council, 2023) By only considering the past 5 years it increases the likelihood that the vehicle is a domestic vehicle.

Figure 6.8 represents the final EV dataset used by this project. Visually inspecting the graph, EVs appear to be mostly in Melbourne with several areas of higher EV counts, these regions are recorded in Table 6.9. Additionally, there was 27 POAs with no EVs.

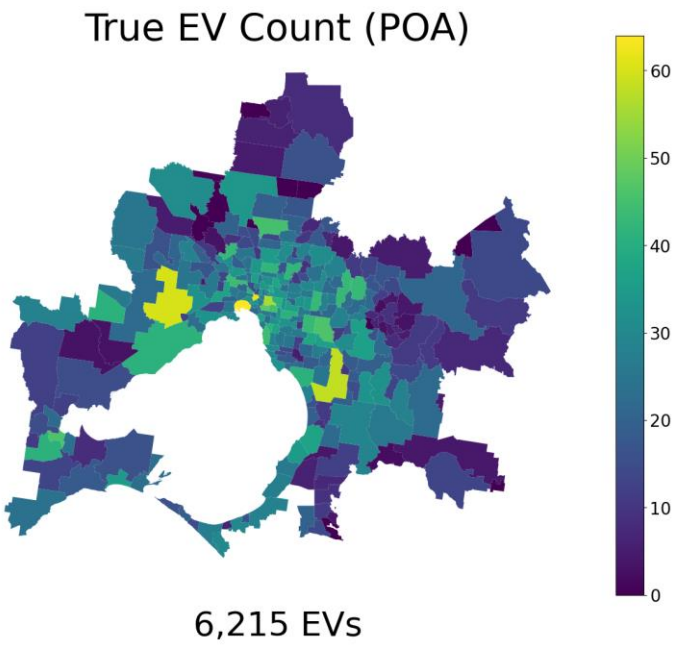


Figure 6.8: True EV Count

Table 6.9: Top 3 EV POA

POA Name	POA Code	EV Count
Port Melbourne	3207	64
Melbourne	3000	62
Truganina	3029	60

6.2.3 Neural Network

The final architecture of the MLP was determined through an iterative development process. The input layer has 42 nodes, one for each input feature. The final layer used one node to predict the number of EVs in the region and was capable of outputting a continuous value. The size of each hidden layer and number of hidden layers were fine-tuned by trial and error. To produce a model that can learn the intricacies of the feature space the model needed to be sufficiently deep to prioritise learning over memorization (Neto, 2018).

The performance of the model was evaluated by inspecting the training and validation loss. This evaluation aimed to determine if the model had sufficient capacity to learn. Furthermore, issues of underfitting and overfitting needed to be balance when trying to adjust the model's capacity. The following techniques were found to be useful in designing and training a well-performing model.

- Not learning: Increase layers heights or increase the number of layers.
- Overfitting: Decrease layers heights and or reduce the number of training epochs.
- Underfitting: Increase layer heights or add additional layers or increase training time.

The final MLP is shown below in Figure 6.9 and consists of 7 hidden layers. The training loss of the model is shown in Figure 6.10. The loss curves for both training and validation show a smooth decrease before stabilising at approximately 9.5 MAE (mean absolute error), indicating the model has finished learning. Furthermore, both the validation and training loss are similar indicates the model is well generalised. A well generalised model is then capable of being used to predict data outside of the training set. The model's ability to predict the number of EVs in the test set resulted in a MAE of 7.57. This error is similar is scale to both the training and validation loss, again providing evidence that the model is well generalised. The test error being smaller however, may indicate that the test set was easier. This could be a result of randomness present in

the small dataset and thus resulted in the easier test set. Given that there is only 281 POA in the population the training, validation and test sets were all small.

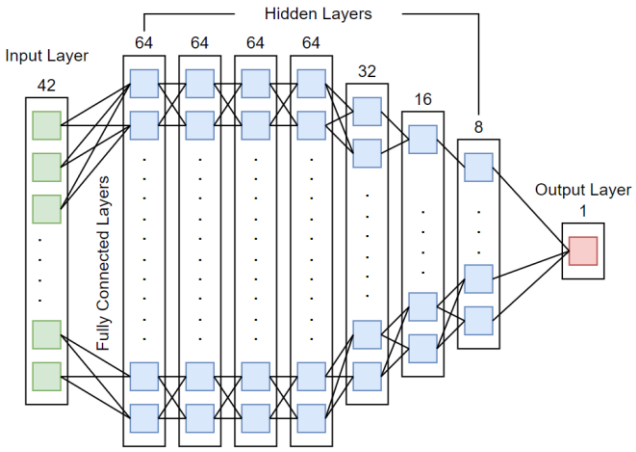


Figure 6.9: MLP Architecture

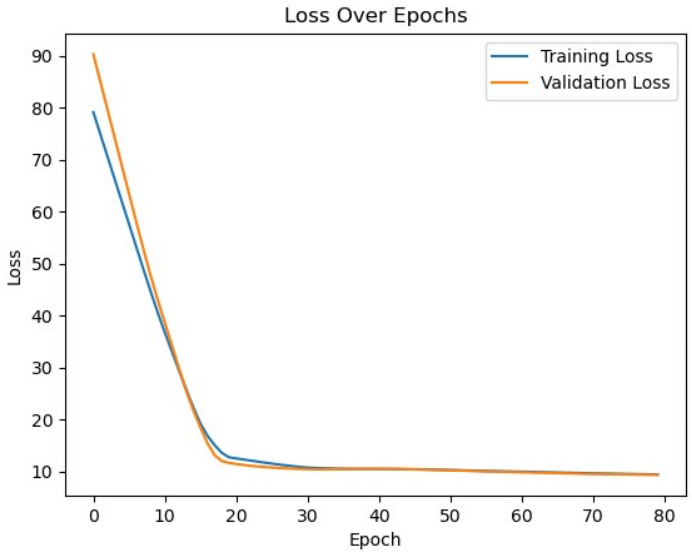


Figure 6.10: Training Loss

Throughout the training and model selection process, the majority of models produced a MAE of approximately 7 or above. Given the current features space used to train on, a MAE of 7 appears to be the minimum error achievable. To produce a more accurate model the incorporation of more household attributes into the synthetic population could be tried. This is because there could be additional attributes that are useful in predicting EV counts within a region such as households' political preference, education level and distance to existing EV charges. However, incorporating more features may not necessarily improve the model's ability to predict EV counts. The underlying randomness in the differences between households in different regions may be causing the remaining error in the model.

Figure 6.11 represents the true number of EVs in each POA. Figure 6.12 represents the predicted number of EVs as predicted on a feature space of household attributes aggregated at the POA level. The total number of predicted EVs was 5,607, whereas the true number of EVs is 6,215. The model under predicted by 608 EVs or approximately 9.8%. By visually inspecting the colour of the two plots the distributions are broadly similar. However, there is some key POAs that differ. For example, Port Melbourne (POA 3207) has 64 EVs however,

the model only predicted 32 EVs. This difference could be a result of the training feature spaces not encapsulating enough different household attributes to allow the model to identify the uniqueness of Port Melbourne.

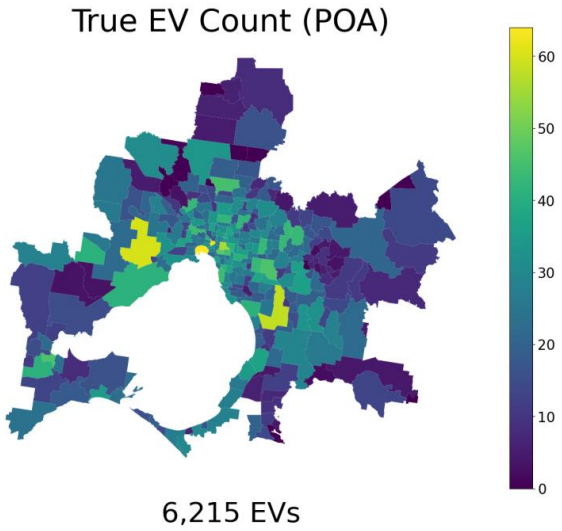


Figure 6.11: True EV Count (POA)

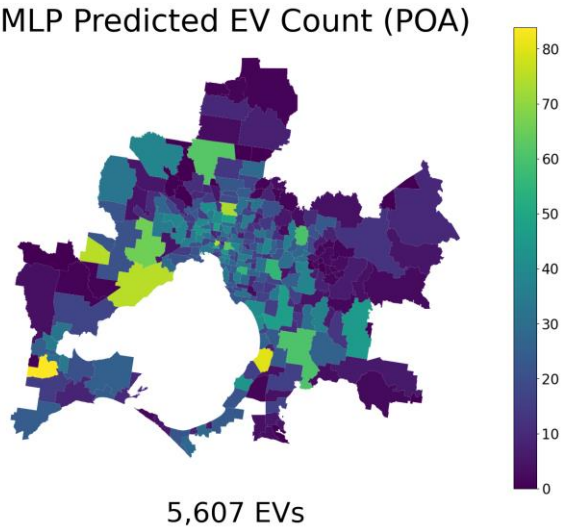


Figure 6.12: Predicted EV Count (POA)

Once the model was trained at the POA level it was then used to predict the number of EVs at the SA1 level. To predict at this level the model was given histograms of household attributes aggregated at the SA1 level. The resulting prediction is shown in Figure 6.13. The total number of EVs predicted at SA1 is 5,897. Unlike the prediction made at the POA level that under predicted by 10%, the prediction at SA1 level under predicted by only 317 resulting in an error of approximately 5.1%.

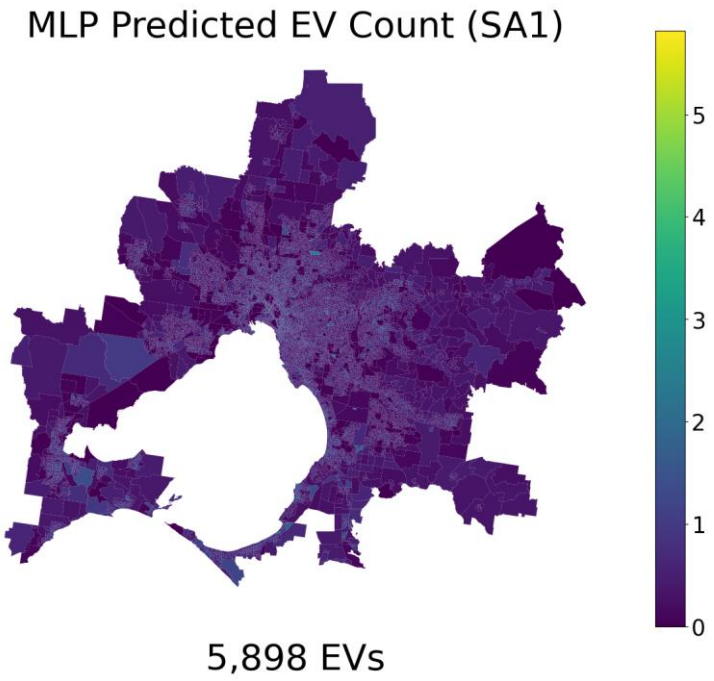


Figure 6.13: Predicted EV Count (SA1)

Each individual SA1 regions can be investigated further to test the validity of the model. As there is no EV data at the SA1 level, individually testing the regions with the most EVs is a simple way to refute the model's accuracy. If these SA1 regions do not align with the EV profile shown in

Table 6.11, then the model is likely to not process accurate predictions capabilities. The SA1 areas with the highest number of predicted EVs are shown in Table 6.10.

Table 6.10: SA1 with most EVs

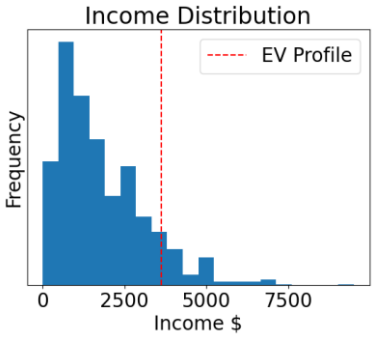
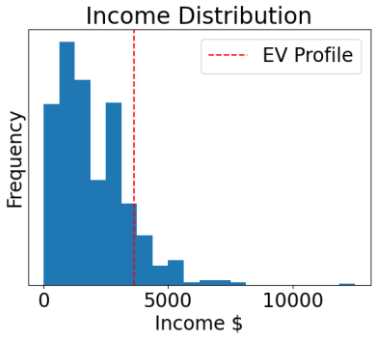
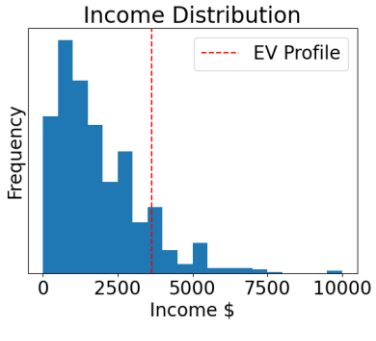
Name of SA2 containing SA1	SA1 Code	Count of EVs
Melbourne CBD - West	20604150504	5.82
Melbourne CBD - West	20604150518	3.36
Moonee Ponds	20603111634	3.26

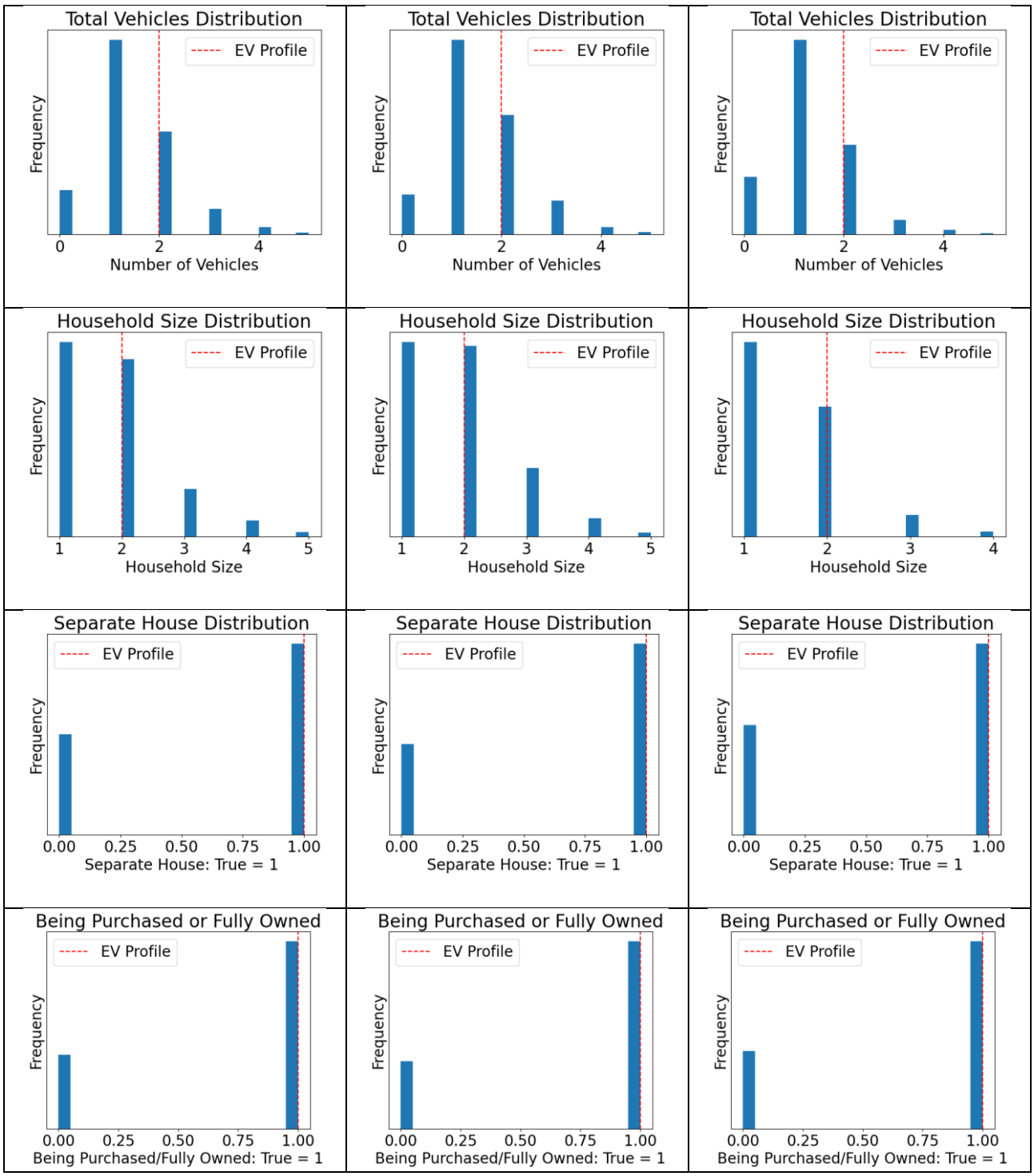
Table 6.11: EV Profile

EV Profile	Threshold
Household Income	≥ $3625
Total vehicles	≥ 2
Household Size	≥ 2
Separate House	True
Dwelling Being Purchased or Fully Owned	True

In this point of the EV prediction method there has been no EVs assigned to individual households. Therefore, the EV profile cannot to directly compared to any one household from the top EV SA1s. However, the distributions of attributes within each of these three SA1s can be investigated to determine if there is a significant distribution of households that could meet these requirements. Table 6.12 shows the frequency distribution of each attribute associated with EV ownership. Furthermore, a red line has been added to indicate the threshold value that is expected to be exceeded for the household to be likely to own an EV. All three SA1 regions population have a significant population above these thresholds. The total number of households that meet all criteria are 34, 20 and 16 for SA1 20604150504, 20604150518 and 20603111634 respectively. Therefore, it was not possible to refute the model's predictive power.

Table 6.12: Distribution top SA1 Attributes

20604150504	20604150518	20603111634
Number of households that met all criteria: 34	Number of households that met all criteria: 20	Number of households that met all criteria: 16
		



To further support the model's accuracy the three plots below can be compared. Figure 6.15 shows the total number of EVs predicted at POA level and Figure 6.16 shows the total number of EVs predicted at SA1 level, but then aggregated to POA level. Although these predictions were made at vastly different geospatial levels, Figure 6.15: Predicted at POA the total number of EVs and their geospatial distribution strongly resembles the true distribution shown in Figure 6.14. This alignment suggests that the model possesses robust predictive capabilities.

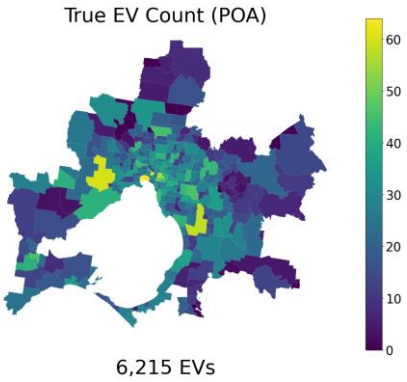


Figure 6.14: True EV Count

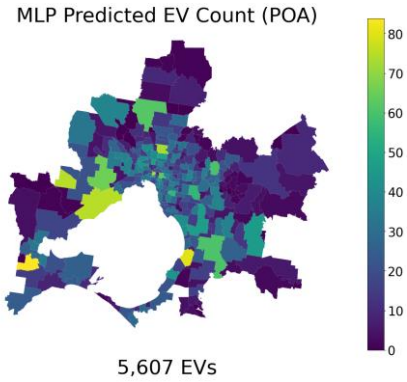


Figure 6.15: Predicted at POA

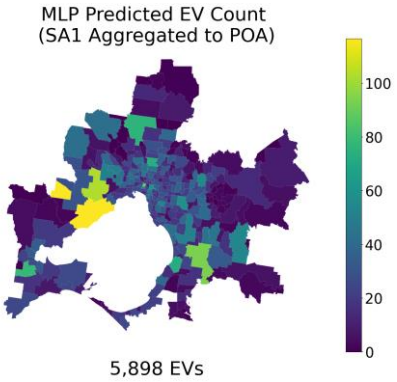


Figure 6.16: Prediction Aggregated to SA1

6.2.4 Assigning EVs

The output of the MLP does not predict EVs as whole numbers. This is problematic as households can only have an integer number of EVs. To solve this problem TRS was used to preserve the total number of EVs while converting all predictions to integers. After the TRS process was performed the total number of EVs went from 5,898.812 to 5,873 EVs. At this point the project goal of predicting the location of EV owners has been achieved. This is because the total number of EVs in each SA1, which is the smallest geospatial unit used in this project, has been predicted. However, there is still the goal of predicting which households will own EVs.

To achieve the goal of finding the household attributes of EV drivers, each EV in each SA1 was assigned to the most likely household. By doing so every household in the population would be labelled as owning an EV or not. The Feature Selection method outlined in the methodology was used to rank and assign the most likely EV owning households.

The external project provided correlation coefficients generated using Pearson, Spearman, and Kendall feature selection. The Pearson feature selection method was chosen because it works with numerical inputs (frequency of each attribute in the area), numerical outputs (number of EVs in each area) and the correlation coefficients are linear. It is important that correlation is linear because each coefficient was added together to rank each household. Table 6.13Table 6.13: Pearson Feature Selection shows the first 6 of 43 correlation coefficients.

Table 6.13: Pearson Feature Selection

Ranking	Attribute	Pearson Score	p Value	Confirmed by Literature
1	$4,000-$4,499	0.197699407	1.61E-07	Yes
2	$5,000-$5,999	0.17711065	2.81E-06	Yes
3	$4,500-$4,999	0.174804098	3.79E-06	Yes
4	Eight or more persons	0.173777433	4.33E-06	No
5	$6,000-$7,999	0.166762471	1.05E-05	Yes
6	$3,000-$3,499	0.157350518	3.26E-05	Yes

After assigning all EVs to the most likely household Figure 6.17 was produced and shows the number of EVs in each SA1 region. Furthermore, Table 6.14 shows the attributes of the EV population. Each attribute has been supported in full or partially by prior research. The average income of the EV population is $1,828, placing it in the top 40% quantile. While this signifies a relatively high-income, it is lower than initially expected. The average number of vehicles per household exceeded 2, aligning with the prior research.

Moreover, a majority of dwellings were separate houses 81.28% and 65.37% of EV owners are either purchasing or fully owned their dwelling, again aligning with the prior research.

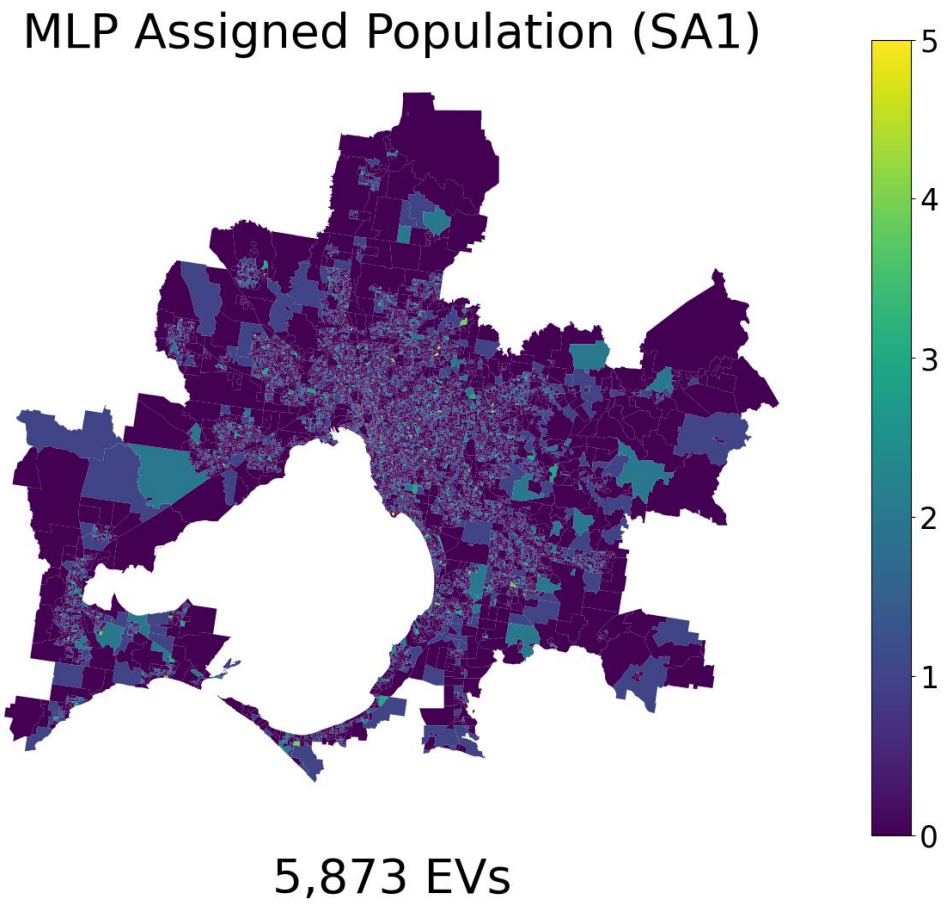


Figure 6.17: Final Population

Table 6.14: Final EV Population Statistics

	Mean	Supported by Literature
Income	$1,828	Partially
Total Vehicles	2.6	Yes
Household Size	6.15	Yes
Dwelling Type		Yes
Flat or Apartment	7.91%	
Separate House	81.28%	
Terrace/Townhouse	7.92%	
Other	2.90%	
Dwelling Ownership		Yes
Being Purchased	41.40%	
Being Rented	32.04%	
Fully Owned	24.97%	
Occupied Rent-Free	1.02%	
Something Else	0.56%	

To further explore the predicted EV population, it can be compared to the predicted non-EV population shown in Table 6.15. Using a Welch's t-test, each household attribute was compared between populations to test if the two populations were significantly different. All attributes were significantly different at the 5% significance level except for income, dwelling occupied rent-free and dwelling occupation something else. To address the insignificant difference between occupied rent-free and something else, this could be a result of these two groups being an insignificant sample of the population. As these categories accounted for approximately 1% or less, they likely didn't play a role in identifying EV ownership.

The income level between EV drivers and non-EV drivers not being significantly different is surprising given the prior research identified that EV ownership is associated with high income. This result challenges prior research and therefore raises the following questions. Firstly, is the model accurate, if not how can the model be improved. If it is accurate, this would suggest that the profile of EV drivers has changed in recent years or that the population of Melbourne and Geelong is unique. The insignificant difference between income could be explained by the increase in affordable EV options which in turns reduces the income barrier required to own one.

Table 6.15: Population Comparison

	EV Mean	Non-EV Mean	Welch t-Stat	p-value	Significant Level (5%)
Income	$1,828	$1,823	0.284	0.777	No
Total Vehicles	2.6	1.7	56.499	0.000	Yes
Household Size	6.15	2.65	200.631	0.000	Yes
Dwelling Type					
Flat or Apartment	7.91%	14.37%	-18.314	0.000	Yes
Separate House	81.28%	70.19%	21.721	0.000	Yes
Terrace/Townhouse	7.92%	10.37%	-6.920	0.000	Yes
Other	2.90%	5.07%	-9.898	0.000	Yes
Dwelling Ownership					
Being Purchased	41.40%	31.80%	14.913	0.000	Yes
Being Rented	32.04%	25.94%	10.002	0.000	Yes
Fully Owned	24.97%	40.87%	-28.084	0.000	Yes
Occupied Rent-Free	1.02%	0.91%	0.890	0.373	No
Something Else	0.56%	0.48%	0.826	0.409	No

6.3 Limitations and future work

6.3.1 Limitations

There were several limitations throughout the project with the main of which being having access to quality EV data. The absence of EV data played a pivotal role in shaping this project. For the entire first semester this project did not have EV data and therefore several initial aims needed to change. Furthermore, the EV data that was provided was not to a satisfactory standard and created several challenges. Firstly, the data did not specify if the vehicle registered was owned by a household or company. This was problematic because this project aimed to investigate which households own EVs and therefore needed to be able to distinguish EVs owned by households to ones owned by businesses. To continue developing a method that could utilise EV data, all vehicle registrations that included more than 2 EVs were removed, and the resulting dataset was assumed to only contained EVs registered by households.

An additional limitation of the dataset is that it only contained the make of vehicle and not the model as well. This limited the amount of exploration that could be conducted in the project. If the dataset contained the vehicle model it would be possible to make the prediction model incorporate price sensitivity into its

prediction. For example, some POAs might have more expensive EVs than other POAs, and therefore this information would give the model more information to learn and make predictions upon.

6.3.2 Future Work

Method 1: Absence of EV data
The primary point of focus for improving this method is decreasing the number of predicted EVs. Currently the model is over predicting by 24x. Future work could explore increasing the number of clusters. Potentially this could be done by increasing the complexity of the feature space thus providing more dimensions to find unique clusters in. The product of this would be smaller clusters and thus a smaller EV prediction.

Method 2: Presence of EV data
To further improve this method there are several changes that could be implemented. Firstly, the synthetic population could include more features. This would increase the learnable features available to the MLP thus improving its ability to differentiate between regions and find more associations between EV ownership and a regions population.

The EV allocation process could be improved. Currently it only relies on the correlation coefficients and not the p-values of these coefficients. The p-value could be used to modify the weight of each attribute's coefficient. If the p-value is small (indicating strong statistical significance) then the coefficient is weighted more and if it is large, then the coefficient is weighted less. Furthermore, the weight could incorporate some randomness based on the p-value breaking the rigid EV allocation based on ranking.

Lastly, research could be conducted to investigate ways of increasing the training dataset. Methods such as data augmentation could be implemented to artificially increase the training set and thus improved the overall performance of the model.

7 Conclusion

This project set out to explore EV ownership in Melbourne and Geelong, with the goal of creating a comprehensive dataset consisting of every household, their attributes, their location, and their EV ownership state. The project employed two distinct methods, each resulting in valuable findings.

Method 1 despite over predicting by approximately 24 times the true EV population of Melbourne and Geelong, has laid the groundworks for future research. As this method used household attributes alone to predict EV ownership, it has the potential to provide policy makers and researchers a way of predicting EV population without the need for extensive EV surveys. Furthermore, the 148,394 houses predicted by the model to own EVs all had attributes that align with the EV profile established in prior research. This therefore suggest that these households have the means to own EVs and could potentially own an EV in the future.

Method 2 was more successful in predicting an accurate number of EVs. Its success came down to utilising additional information that was not available to method 1. Method 2 predicted a total of 5,507 EVs when given household data aggregated to the POA level. Given the true number of EVs is 6,215 the models error was approximately 10%. When predicting EVs given household data aggregated to SA1 the model predicted 5,898 EVs and had an error of approximately 5%. The ability of method 2 to predict with similar levels of accuracy across vastly different geospatial levels helped to confirm the model's accuracy.

Overall, the project was successful in identifying the EV population of Melbourne and Geelong. This included identifying where their households were located, the individual household's attributes as well as the total number of EVs in each SA1 region. Method 2 is the preferred method for predicting EV ownership due to its greater accuracy and has been proven to work with geospatial data from POA down to SA1. The distribution of EVs found by this project can help guide future EV infrastructure and help design EV adoption policies. The

results from this project have the possibility to help smooth transition from ICE vehicles to EVs in Melbourne and Geelong.

8 Reflection on Project Management

8.1 Project Scope

The following are within the scope of the project:

1. Collecting vehicle ownership data associated with geographic location.
2. Cleaning vehicle ownership data to identify EVs.
3. Identify demographic attributes associated with EV drivers.
4. Assigning each household in a synthetic population EV ownership status.
5. Developing a model to assign EV ownership without EV data.
6. Developing a model to assign EV ownership that utilises EV data.

The following are outside the scope of the project:

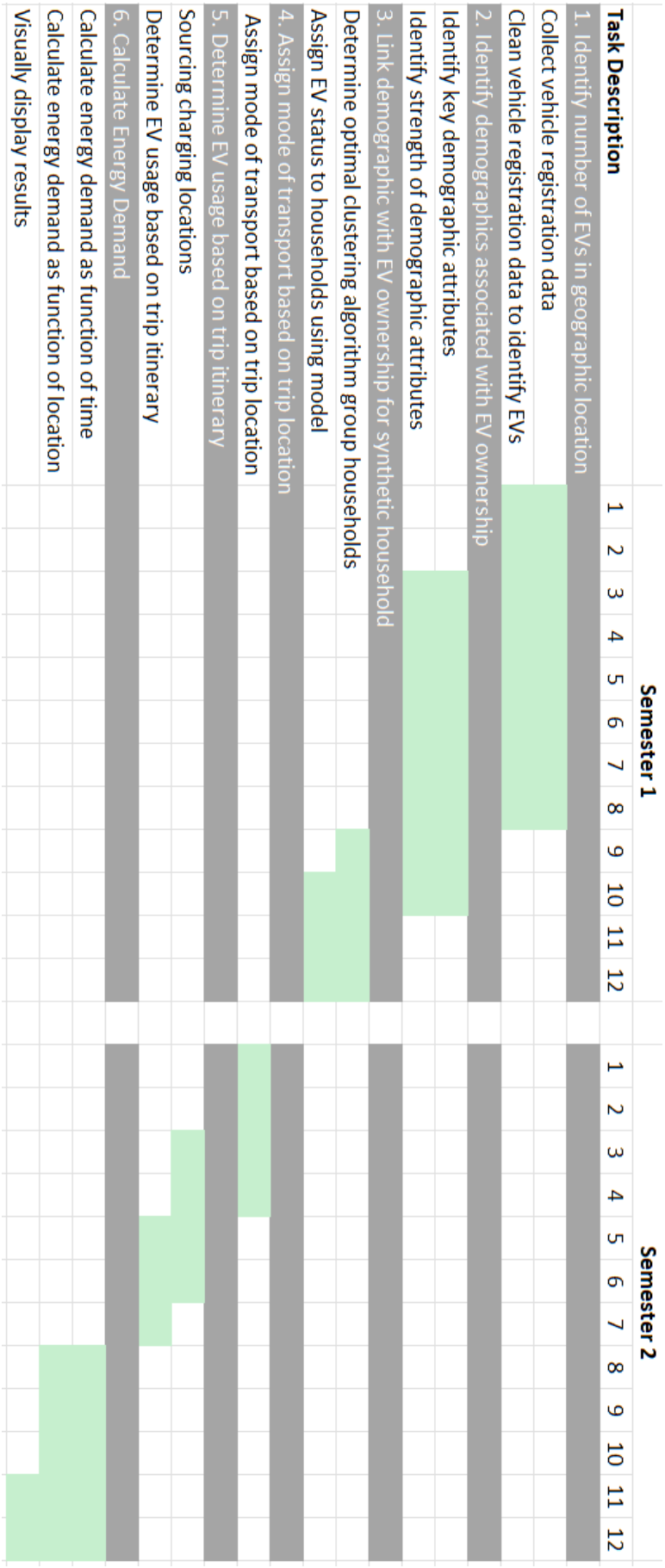
1. Creating a synthetic population for Melbourne and Geelong.
2. Produce correlation coefficient through Feature Selection.
3. Suggesting policies that would increase EV adoption.
4. Suggesting locations for EV charging infrastructure.
5. Modelling energy demand of EVs.

8.2 Project Plan & Timeline

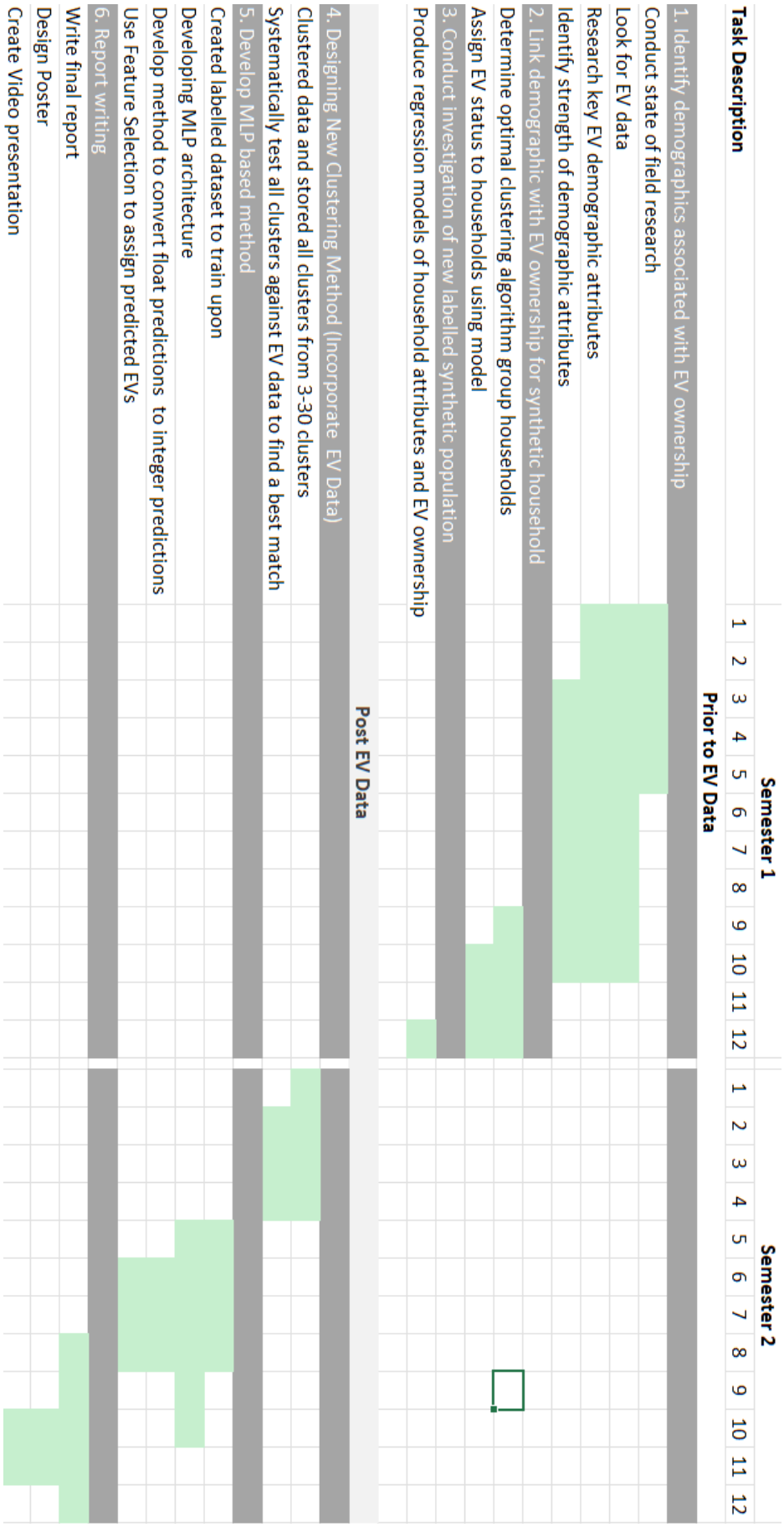
The timeline for this project has changed greatly with the delay of EV data. Originally the project aimed to firstly assign households in a synthetic population EV ownership then use this data to make predictions about EV energy demands. However, due to the EV data not arriving until after semester 1, the first part of the project was completed twice, once without EV data and once with EV data.

The original project timeline and the updated timeline of events which occurred are shown below. The first 12 weeks of the project were completed with compliance with the original timeline. Whereas the last 12 weeks of the project were completed with compliance to the second timeline. Other than data delays there were little setbacks throughout the project. This was achieved by completing difficult tasks in parallel with other tasks where possible. For instance, in semester 2 training a well-fitting MLP model was running over schedule. To prevent this part of the project delaying the rest of the project, the assignment of EVs to the most likely households and TRS processes were completed. The process of completing these tasks gave additional time to reflect and consult with my supervisors about the training issues. When revisiting the training process, it was complete effectively due to the additional preparation.

8.2.1 Original timeline:



8.2.2 Update timeline



8.3 Reflection on Project

This project was conducted individually with the help of my supervisors. As this was an individual project the success of the project was solely dependent on my own work. This had the benefit of allowing me to have greater control over the direction of the project and allowed for more flexibility in when work could be completed. However, there were several drawbacks such as an increase workload, decreased support in decision making and meant I was personally responsible for overcoming all challenges. Despite this I believed that I was successful in keeping myself accountable throughout the project.

The project changed considerably throughout the course of the year. Due to the limitation of EV data availability the second half of the project had to be reworked. This resulted in a complete rework of my goals for the second semester. Initially, I set out to also predict the energy demand of EVs, however this was no longer possible as the first stage of predicting EV ownership was delayed. The delays in data resulted in two separate methods for predicting EVs being developed. Despite the delay I ensured that the first semester was not wasted. I did this by creating an innovative method for predicting EVs without EV data. By doing so it allowed me to further familiarise myself with working with geospatial data as well as large datasets. This experience allowed me to be more productive and efficient in developing the second method once the data arrived.

If I were to complete a similar project in the future, there are several things I would change in my workflow. Firstly, I would create a clearer definition of the households included in my area of study. The initial synthetic population I used in this project was in SA2 and therefore I selected my households using this. However, in semester 2 the synthetic population provided was in SA1. Because of this the two populations between semesters varied slightly in which households were included. If I had initially used the area defined by SUA from the start, then there would have been smaller discrepancies. As this project relied heavily on programming and working with datasets, I believed I could have structured the data flow better from the start. Towards the end of the project, it became difficult to make small changes and to update my final results. However, knowing what was required is now a consequence of having hindsight. Despite the many challenges throughout the project, I was successful in completing my revised goal of creating two unique methods for predicting EV ownership.

9 References

Abdel-Rahman, A. A. (1998). On the emissions from internal-combustion engines: a review. *Int. J. Energy Res*, *22*(6), 483-513. https://doi.org/10.1002/(SICI)1099-114X(199805)22:6<483::AID-ER377>3.0.CO

2-Z

Albatayneh, A., Assaf, M. N., Alterman, D., & Jaradat, M. (2020). Comparison of the Overall Energy Efficiency for Internal Combustion Engine Vehicles and Electric Vehicles. *Environmental and Climate Technologies*, *24*(1), 669-680. https://doi.org/10.2478/rtuect-2020-0041

Bellemans, T., Kochan, B., Janssens, D., Wets, G., Arentze, T., & Timmermans, H. (2010). Implementation framework and development trajectory of FEATHERS activity-based simulation platform. *Transportation research record*, *2175*(2175), 111-119. https://doi.org/10.3141/2175-13

Bjerkan, K. Y., Nørbech, T. E., & Nordtømme, M. E. (2016). Incentives for promoting Battery Electric Vehicle (BEV) adoption in Norway. *Transportation research. Part D, Transport and environment*, *43*, 169-180. https://doi.org/10.1016/j.trd.2015.12.002

Bonaccorso, G. (2018). *Machine Learning Algorithms : Popular Algorithms for Data Science and Machine Learning, 2nd Edition* (2nd ed.). Birmingham : Packt Publishing, Limited.

Both, A., Singh, D., Jafari, A., Giles-Corti, B., & Gunn, L. (2021). An Activity-Based Model of Transport Demand for Greater Melbourne. In. Ithaca: Ithaca: Cornell University Library, arXiv.org.

Campbell, A. R., Ryley, T., & Thring, R. (2012). Identifying the early adopters of alternative fuel vehicles: A case study of Birmingham, United Kingdom. *Transportation research. Part A, Policy and practice*, *46*(8), 1318-1327. https://doi.org/10.1016/j.tra.2012.05.004

Council, E. V. (2023). *State of Electric Vehicles July 2023*.

Deb, S., Kalita, K., & Mahanta, P. (2017, 21-23 Dec. 2017). Review of impact of electric vehicle charging station on the power grid. 2017 International Conference on Technological Advancements in Power and Energy (TAP Energy),

Department of Climate Change, E., the Enviroment and Water. (2022). *Australia's Long-Term Emissions Reduction Plan*. DCCEEW. Retrieved 29 March 2022 from https://www.dcceew.gov.au/climate-change/publications/australias-long-term-emissions-reduction-plan

Department of Climate Change, E., the Environment and Water. (2023). *Renewables*. DCCEEW. Retrieved 21 May 2023 from https://www.energy.gov.au/data/renewables

Developers, G. (2023). *Prepare Data*. Google. Retrieved 22 May 2023 from https://developers.google.com/machine-learning/clustering/prepare-data

Hao, H., Qiao, Q., Liu, Z., & Zhao, F. (2017). Impact of recycling on energy consumption and greenhouse gas emissions from electric vehicle production: The China 2025 case. *Resources, conservation and recycling*, *122*, 114-125. https://doi.org/10.1016/j.resconrec.2017.02.005

Harland, K., Heppenstall, A., Smith, D., & Birkin, M. (2012). Creating Realistic Synthetic Populations at Varying Spatial Scales: A Comparative Critique of Population Synthesis Techniques. *Journal of artificial societies and social simulation*, *15*(1). https://doi.org/10.18564/jasss.1909

Hjorthol, R. (2013). Attitudes, ownership and use of Electric Vehicles - a review of literature. *Institute of Transport Economics*.

Kirill, M. (2011). Population synthesis for microsimulation.

Knapen, L., Kochan, B., Bellemans, T., Janssens, D., & Wets, G. (2012). Activity-Based Modeling to Predict Spatial and Temporal Power Demand of Electric Vehicles in Flanders, Belgium. *Transportation research record*, *2287*(1), 146-154. https://doi.org/10.3141/2287-18

LaMonaca, S., & Ryan, L. (2022). The state of play in electric vehicle charging services – A review of infrastructure provision, players, and policies. *Renewable & sustainable energy reviews*, *154*, 111733. https://doi.org/10.1016/j.rser.2021.111733

Lovelace, R., & Ballas, D. (2013). 'Truncate, replicate, sample': A method for creating integer weights for spatial microsimulation. *Computers, environment and urban systems*, *41*, 1-11. https://doi.org/10.1016/j.compenvurbsys.2013.03.004

Ma, J., & Ye, X. (2019). Modeling Household Vehicle Ownership in Emerging Economies. *Journal of the Indian Institute of Science*, *99*(4), 647-671. https://doi.org/10.1007/s41745-019-00133-9

Nations, U. (2023). *THE 17 GOALS | Sustainable Development Goals*. UN. Retrieved 16 May 2023 from https://sdgs.un.org/goals

Neto, E. C. (2018). Detecting Learning vs Memorization in Deep Neural Networks using Shared Structure Validation Sets. *arXiv.org*. https://doi.org/10.48550/arxiv.1802.07714

Nidheesh, N., Nazeer, K. A. A., & Ameer, P. M. (2020). A Hierarchical Clustering algorithm based on Silhouette Index for cancer subtype discovery from genomic data. *Neural computing & applications*, *32*(15), 11459-11476. https://doi.org/10.1007/s00521-019-04636-5

Noyce David, A. (2019). An Analysis of Attributes of Electric Vehicle Owners' Travel and Purchasing Behavior: The Case of Maryland. In (pp. 1-1). American Society of Civil Engineers (ASCE).

Ossama, E. R., & Virginia, P. S. (2019). A Critical Review on Population Synthesis for Activity- and Agent-Based Transportation Models. In L. Stefano De, P. Roberta Di, & D. Boban (Eds.), *Transportation Systems Analysis and Assessment* (pp. Ch. 1). IntechOpen. https://doi.org/10.5772/intechopen.86307

Patel, E., & Kushwaha, D. S. (2020). Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model. *Procedia computer science*, *171*, 158-167. https://doi.org/10.1016/j.procs.2020.04.017

Patel, V. R., & Mehta, R. G. (2011). Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm. *International journal of computer science issues*, *8*(5), 331.

Peters, A., & Dütschke, E. (2014). How do Consumers Perceive Electric Vehicles? A Comparison of German Consumer Groups. *Journal of environmental policy & planning*, *16*(3), 359-377. https://doi.org/10.1080/1523908X.2013.879037

Plötz, P., Schneider, U., Globisch, J., & Dütschke, E. (2014). Who will buy electric vehicles? Identifying early adopters in Germany. *Transportation research. Part A, Policy and practice*, *67*, 96-109. https://doi.org/10.1016/j.tra.2014.06.006

Qian, L., & Soopramanien, D. (2011). Heterogeneous consumer preferences for alternative fuel cars in China. *Transportation research. Part D, Transport and environment*, *16*(8), 607-613. https://doi.org/10.1016/j.trd.2011.08.005

Range of full electric vehicles. (2023).

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York : Free Press.

Secrist, E. S., & Fehring, T. K. (2023). Cobalt Mining in the Democratic Republic of the Congo for Orthopaedic Implants: A Complex Ethical Issue with No Simple Solutions. *J Bone Joint Surg Am*, *105*(2), 167-171. https://doi.org/10.2106/JBJS.21.01277

Sievewright, B. (2022). *State of Electric Vehicles – March 2022*.

van Vliet, O., Brouwer, A. S., Kuramochi, T., van den Broek, M. A., Faaij, A. P. C., Options for a sustainable energy, s., Sub Science, T., & Society, b. (2011). Energy use, cost and CO2 emissions of electric cars. *Journal of power sources*, *196*(4), 2298-2310. https://doi.org/10.1016/j.jpowsour.2010.09.119

Wang, K., Zhang, W., Mortveit, H., & Swarup, S. (2021). Improved Travel Demand Modeling with Synthetic Populations. In (pp. 94-105). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-66888-4_8

Whitehead, J. (2023). *2022 Australian Electric Vehicle Industry Recap*.

10 Appendices

10.1 Appendix A: Project Risk Assessment

46328	RISK DESCRIPTION	TREND	CURRENT	RESIDUAL
	ENG4701_CL_2023_S1_DanielLawson_Population synthesis and applications for urban planning		Negligible	Negligible

RISK TYPE
1. Activity or Task Based Risk Assessment

RISK OWNER	RISK IDENTIFIED ON	LAST REVIEWED ON	NEXT SCHEDULED REVIEW
DANIEL MAURICE LAWSON	14/03/2023	14/03/2023	14/03/2026

RISK FACTOR(S)	EXISTING CONTROL(S)	CURRENT	PROPOSED CONTROL(S)	TREATMENT OWNER	DUE DATE	RESIDUAL
Sitting for prolonged periods of time while programming may lead to: Back/neck injury Repetitive strain injuries	Control: 1. Avoid using computer for more than 2 hours at a time as well as taking regular breaks to perform stretching and walking. 2. Sitting with good posture. 3. Setting monitor height close to eye level. Control Effectiveness:	Negligible	No Control:	DANIEL MAURICE LAWSON		Negligible

10.2 Appendix B: Risk Management Plan

Project Risk	Risk	Likelihood	Consequence	Risk level	Mitigation	Residual Risk
Clean EV data set is significantly delayed or not delivered.	Unable to assign synthetic population EV ownership status. Will require cleaning the vehicle registration data set.	Possible	Major	High	Develop a small-scale API to link the uncleaned registration data set to an online car database to identify EVs. This will test the viability of this option. If the cleaned registration data is delayed but expected to be delivered, start working on the clustering of households to keep the project moving.	Will be unable to assign EV ownership status based on Melbourne data and will need to assign based on researched identified characteristics.
Not able to assign EV ownership-based EV data and household attributes.	Number of EVs in synthetic population will not accurately represent the real population	Possible	Major	High	Identify several ways to assign EV ownership and do not just rely on one method.	The total number of EVs does not match the real population and is not assigned logically.
Clustering Method doesn't not identify and EV population	The model will not predict any EVs	Unlikely	Major	Medium	Increase the number of clusters until there is an EV population. Or decrease the requirements for an EV population	The model will not be as accurate because the standard for EV prediction has been reduced.
MLP does not learning useful features from training set	The second method for prediction EVs will fail	Possible	Major	High	Increase the capacity of the model until it can learn. Or increase the proportion of data used in the training set.	By doing so the model risks overfitting the data and therefore holding no true predictive power.

10.3 Appendix C: Sustainability Plan

Sustainability is an important part of engineering because engineering is the foundation from which society is built. Therefore, for society to be able to progress toward its environmental and sustainability targets it is essential to focus on engineering projects that align with these targets. The United Nations has developed 17 interconnected goals that have been adopted by all United Nations Member States. Known as the United Nations Sustainable Development Goals (SDGs), which encompass issues such as poverty, health, education, gender equality, sanitation, economic growth, clean energy and many more. This research project most aligns with Goal 7: Affordable and Clean Energy, which focuses on ensuring access to affordable, reliable, and sustainable energy for all (Nations, 2023). By understanding the electricity demand of EVs and their impact on the energy sector this project contributes to the understanding of the role EVs play in achieving affordable and clean energy.

The SDGs are further broken down into targets that are to be achieved by 2030. Target 7.2 aims to substantially increase the share of renewable energy in the global mix. This goal's progress is indicated by measuring the share of renewable energy consumed. This project contributes to EV adoption and therefore will help decouple the world from fossil fuel dependant transport. Therefore, leading to a proportionally greater share of renewable energy in the energy sector regardless of increasing the total amount of renewable energy. Moreover, increasing the number of EVs will drive advancements in battery technology and environmental awareness regarding renewable energy.

Reducing greenhouse gas emissions is the primary driver in the switch to renewable energy. Society's dependency on fossil fuels in the transport industry is a key hurdle to overcome. An issue with the current state of EVs is they are essentially fossil fuel powered cars. Comparing the CO2 emissions from EV and petrol vehicle give similar result with one study finding that BPEV produce 136g/km of CO2 whereas a petrol vehicle produces 163g/km (van Vliet et al., 2011). These statistics are highly dependent on the energy production of the area. In 2021 Australia produced 29% of its electrical energy from renewable sources (Department of Climate Change, 2023) and aims to reach 100% by 2050 (Department of Climate Change, 2022). If this is achieved the CO2 emission of EVs will be reduce entirely to the production of the vehicle. EVs are also more energy efficient then ICE vehicles. When powered by solar, EVs can be 57% efficient and when powered by coal are 20% efficient which is still more efficient than gasoline ICE vehicles at 19% (Albatayneh et al., 2020). EV adoption is a path to increasing overall energy efficiency of our society.

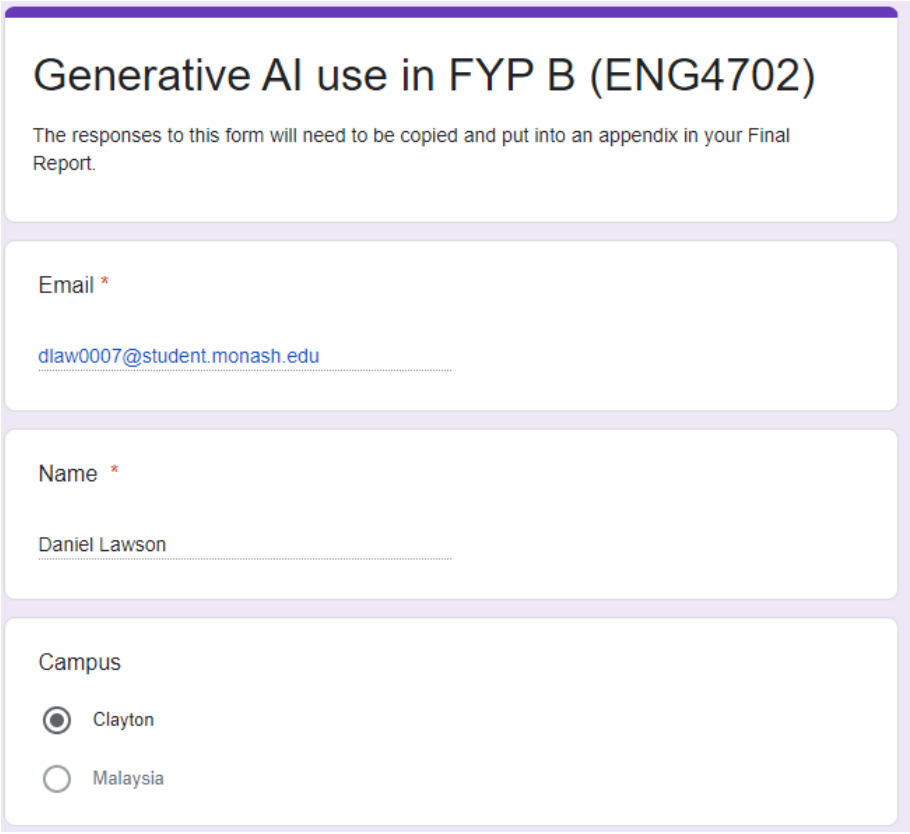
However, like any technology EVs come with downsides. The production of EVs is an energy and resource-intensive process particularly due to the requirement of rare earth minerals in batteries manufacturing. Mining is an essential and unavoidable part of extracting resources for the functioning of our society, however, when done irresponsibly, it has detrimental environmental and humanitarian implications. For example, cobalt is an essential resource in batteries, but the mining is known for human exploitation. Approximately 88% of cobalt comes from the Democratic Republic of the Congo (DRC), where around 1500,000 artisanal miners are exposed to high fatality rates. It has been estimated that over 80 people die each year. Moreover, the use of child labour is prevalent with more than 40,000 children as young at 7 working in the mines (Secrist & Fehring, 2023). For EVs to be a sustainable alternative to ICE vehicles they need to be sustainable in both an environmental and humanitarian sense. The UN will not be able to meet their sustainable energy goals, nor their humanitarian goals unless more attention is put toward addressing the consequences associated with the entire life cycle of EVs, from resource extraction to consumer use. This project plays a small but important role in furthering society's understanding of where precious resources can best be allocated to have a greater impact on EV adoption.

Another hurdle of transitioning to renewable energy is battery capacity. Renewable energy production is highly dependent on weather and therefore require significant energy storage. These batteries will degrade

over time and will require recycling. By tackling the EV battery recycling issue, it will contribute to battery recycling in general. Increased investment in battery recycle facilities is essential to prevent dangerous chemicals in the batteries from polluting landfill. Additionally, retrieving the rare earth minerals from the batteries will reduce the demand on mining. The greenhouse gas emissions within the EV production phase are much higher than conventional vehicles. However, by improving recycling, CO2 emission in the production phase can be reduced to 9.8t CO2eq from 14.9t (Hao et al., 2017). Understanding future demand of EVs will help predict future pressures on battery recycling.

Another target of goal 7 is 7.1 which aims to achieve universal access to affordable, reliable, and modern energy services. If EVs are to be universally adopted issues of price and charging need to be addressed. Currently there is a degree of inequality in the adoption of EVs which this project has partially used to assigning EVs in the modelling phase. The high initial cost of owning an EV is preventing a more universal adoption of EVs. EVs have a higher purchasing cost and require expensive charging units to be installed at home. To help address these issues the Australian government need to use incentives to support people in buying their first electric vehicle. This project can help identify critical areas for public EV chargers to be built to encourage EV adoption with the consideration of giving areas with lower means the opportunity to charge their EV, improving universal access to this sustainable technology.

10.4 Appendix D: Generative AI Statement



Generative AI use in FYP B (ENG4702)

The responses to this form will need to be copied and put into an appendix in your Final Report.

Email *

dlaw0007@student.monash.edu

Name *

Daniel Lawson

Campus

- (●) Clayton
- ( ) Malaysia

Host Department

- (●) Chemical and Biological Engineering
- ( ) Civil Engineering
- ( ) Electrical and Computer Systems Engineeing
- ( ) Materials Science Engineering
- ( ) Mechanical and Aerospace Engineering
- ( ) Software Engineering
- ( ) Robotics and Mechatronics Engineering

Supervisor

Le Hai Vu, Bob La

This project has been conducted using AI tools *

- (●) In this assessment, there will be no use of generative artificial intelligence (AI). All content in relation to the assessment task has been produced by the authors.
- ( ) In this assessment, the following generative AI will be used for the purposes nominated in part 2. (Please note: any use of generative AI must be appropriately acknowledged - see Learn HQ)
- ( ) In this assessment, AI writing assistants (e.g., Grammarly, Writesonic, Quillbot, Microsoft Editor) will be the only form of Generative AI used.
- ( ) This project involves the development or authoring of Unique Generative AI, Unique operation of commercially available Generative AI OR Unique non-generative AI (Machine Learning, Artificial Neural Network, Logistic Regression, etc.)

Permissions

The use of Generative AI has been discussed with and approved by my academic supervisor. *

- ( ) Yes
- (●) No

End

Thank you for completing this form - your responses will be emailed to you for your Progress Report