Department of Electrical and Computer Engineering

Machine Learning - ENCS5341

Assignment I

Prepared By

Ahmad Qaimari-1210190

Yazan AboEloun-1210145

Supervisor

Dr.Ismail Khater

Submission Date

Oct 31, 2024

Table of Contents:

Table of Figures:

Dataset Overview

The dataset captures information on Battery Electric Vehicles (BEVs) and Plug-in Hybrid Vehicles (PHEVs) registered in Washington State, including vehicle specifications (such as Make, Model, and Electric Range), registration details (State, City, and County), and legal identifiers (like the DOL Vehicle ID). This data is valuable for analyzing trends in electric and hybrid vehicle adoption, understanding regional registration patterns, and assessing the impact of regulations on vehicle ownership.

Data Cleaning and Feature Engineering

Cleaning Strategy

- The rows where all feature values are null will be removed.
- The rows where not all columns are null will use median imputation if it is of numeric type.
- Some columns have numeric values but are indeed encoded categorical values (such as Legislative district) Additionally, some are numeric but their median is analytically useless (E.g **Postal Code)**. These will use mode imputation.
- Categorical columns will be mode imputated

Documenting Missing Values

	Missing Values	Percentage
Legislative District	445	0.211738
Vehicle Location	10	0.004758
Electric Range	5	0.002379
Base MSRP	5	0.002379
County	4	0.001903
City	4	0.001903
Postal Code	4	0.001903
Electric Utility	4	0.001903
2020 Census Tract	4	0.001903

Figure 1 - Missing Values

These results show that most features have minimal missing values, except for the **Legislative District** feature which has 20% missing values.

Handling Missing Values

For dealing with missing values in features that are categorical or encoded as categorical, **mode imputation** method will be used. This means that missing values will be replaced with the most frequent value (**mode**) observed in the respective feature. Generally, Categorical features represent distinct categories or groups. Directly substituting missing values with a numerical

average (e.g., **mean or median**) wouldn't be appropriate as it wouldn't preserve the feature's categorical nature. Mode imputation, on the other hand, ensures that the substituted value aligns with the existing categories, maintaining the feature's integrity and distribution.

For instance, a feature like '**Postal Code**' is a numeric value that posses no meaning, computing the **average** would contribute nothing to the analysis. On the other hand, calculating the **mode** gives a sensible result. As for something like the base MSRP feature, imputing the values by the **median** doesn't cause wrong biasing of the analysis,which is why it was chosen over other methods.

Feature Encoding

The approach used involves encoding categorical features based on the number of unique categories(cardinality) they contain.

- **High Cardinality Features:** Features with a large number of unique categories (e.g., more than 3) are often best handled with label encoding. This assigns a unique numerical label to each category. While this introduces an artificial ordinal relationship between categories, it is often preferred for high-cardinality features because one-hot encoding would create an excessive number of new columns, potentially leading to the curse of dimensionality.

- **Low Cardinality Features:** Features with a small number of unique categories (e.g., less than or equal to 3) are better suited for one-hot encoding. This creates a new binary column for each category, where a 1 indicates the presence of that category and a 0 indicates its absence. One-hot encoding effectively avoids introducing artificial ordinal relationships and is often more appropriate for low-cardinality features.

The figures below show the results of applying both types of encoding to different features of the dataset.

County	City	State
87	595	44
87	524	44
169	61	44
87	64	44
85	546	44

Figure 2 - Label Encoding

Electric Vehicle Type_Battery Electric Vehicle (BEV)	Electric Vehicle Type_Plug-in Hybrid Electric Vehicle (PHEV)	Clean Alternative Fuel Vehicle (CAFV) Eligibility_Clean Alternative Fuel Vehicle Eligible	Clean Alternative Fuel Vehicle (CAFV) Eligibility_Eligibility unknown as battery range has not been researched	Clean Alternative Fuel Vehicle (CAFV) Eligibility_Not eligible due to low battery range
0.0	1.0	1.0	0.0	0.0
1.0	0.0	1.0	0.0	0.0
0.0	1.0	0.0	0.0	1.0
1.0	0.0	1.0	0.0	0.0
1.0	0.0	1.0	0.0	0.0

Figure 3 - One Hot Encoding

‘

Normalization

Normalization was performed using **z-score** standardization. This involves subtracting the **mean** and dividing by the **standard deviation** for each feature, transforming the data to have a **mean** of 0 and a **standard deviation** of 1.The figure below shows the normalization of the census feature.

	2020 Census Tract	2020 Census Tract new
0	5.303509e+10	0.035964
1	5.303509e+10	0.035964
2	5.306105e+10	0.052697
3	5.303508e+10	0.035957
4	5.303303e+10	0.034637

Figure 4 - Normalization

Descriptive Statistics

Non-numeric columns and things like **Postal Code** statistics are of no interest in this case. They are dropped to leave more space to the numeric columns so as to make the table less cluttered.

	Model Year	Electric Range	Base MSRP	2020 Census Tract
count	210165.000000	210165.000000	210165.000000	2.101650e+05
mean	2021.048657	50.601037	897.655533	5.297930e+10
std	2.988941	86.972525	7653.498812	1.551452e+09
min	1999.000000	0.000000	0.000000	1.001020e+09
25%	2019.000000	0.000000	0.000000	5.303301e+10
50%	2022.000000	0.000000	0.000000	5.303303e+10
75%	2023.000000	42.000000	0.000000	5.305307e+10
max	2025.000000	337.000000	845000.000000	5.602100e+10

Figure 5 - Descriptive Statistics for the Features in the Dataset

Spatial Distribution

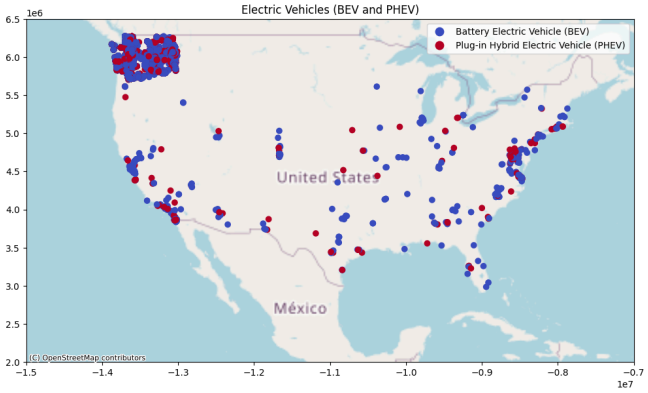

Figure 6 - Spatial Distribution for EVs in Washington State

The figure above shows a map of **Washington** state and the distribution of **EV sales** across it.The upper left corner where most **EV sales** are represents **Seattle**, which is the capital city.

Model Popularity

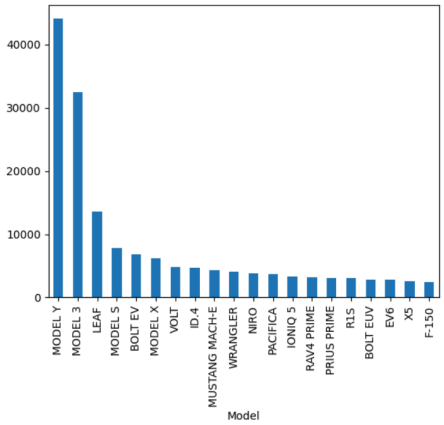

Figure 7 - Model Popularity Bar Plot

Figure 6 shows that Tesla clearly leads the **EV market** in terms of **model popularity**, with significant demand for a few other models. The remaining brands appear to have a small share each, indicating a potential opportunity for other manufacturers to increase market presence.

Correlation Matrix

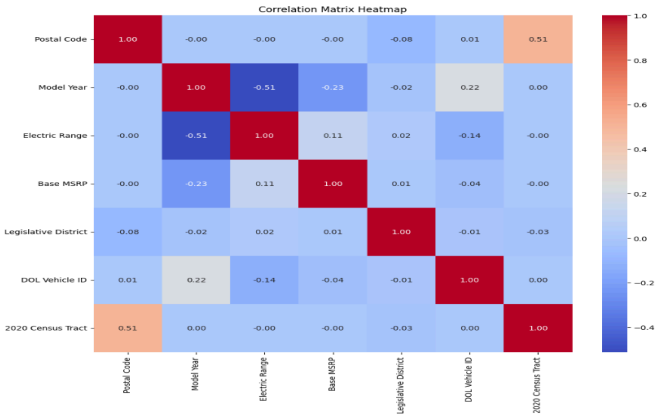

Figure 8 - Correlation Matrix

There are a few mild to moderate **correlations** in the dataset, notably between **Model Year** & **Electric Range** and **Postal Code** & **2020 Census Tract**. Most features do not exhibit strong **correlations**, indicating that they likely contribute unique information to the dataset without being highly interdependent.

Data Visualization

Data Exploration Visualizations

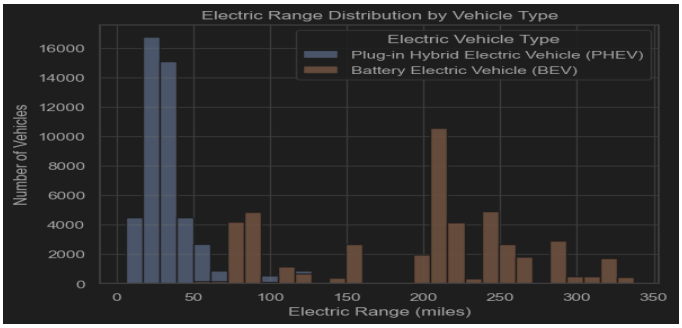

Figure 9 - Electric Range Distribution by Vehicle Type

The **histogram** shows **BEVs** generally have a higher **electric range** than the more common **PHEV**s, which rely partly on combustion engines. Most **BEVs** cluster around a 200-mile range, a standard EV benchmark.

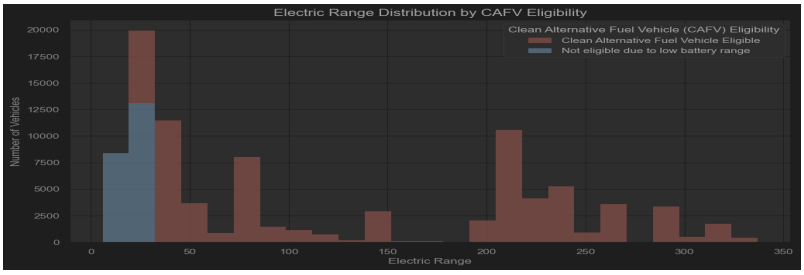

Figure 10- Electric Range Distribution by Vehicle Type

The **histogram** shows that vehicles with an electric range over 50 miles are more likely to qualify for the **Clean Alternative Fuel Vehicle** (CAFV) program, with peaks around 100, 200, and 300 miles. This suggests the program prioritizes **EVs** with longer ranges to support sustainable transportation.

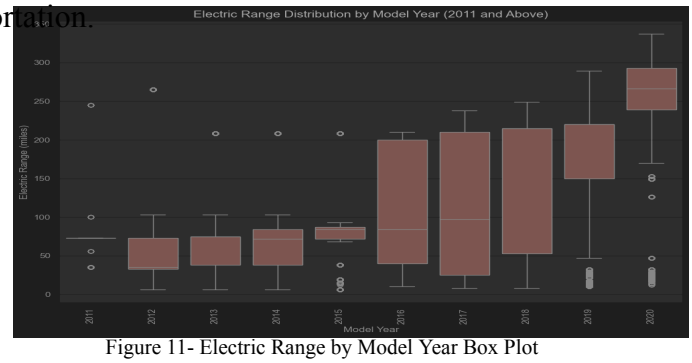

Figure 11- Electric Range by Model Year Box Plot

The **box plot** shows a sharp rise in **electric vehicle range** from 2016, with newer models often exceeding 200 miles, reflecting battery improvements. Recent years show increased range diversity, with some **EV**s surpassing 300 miles.

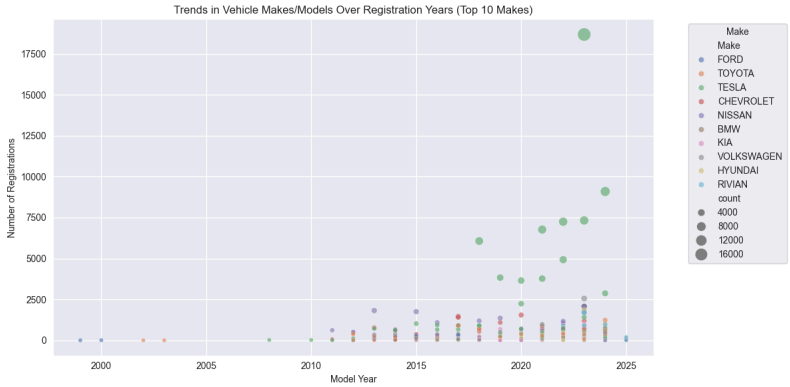

Figure 12- Trends in Vehicle Makes over Registration Years

The **scatter plot** illustrates the number of registrations over the years for the top 10 vehicle makes. We can observe that **Tesla, Ford, and Toyota** had the highest number of registrations in 2020, while **Rivian** had the lowest.

Comparative Visualization

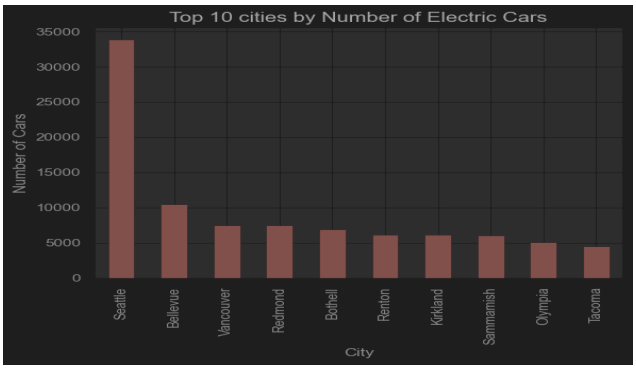

Figure 13 - Trends in Vehicle Makes over Registration Years

As seen in the **bar chart** above , **Seattle** significantly outpaces other cities in **electric vehicle** (EV) adoption, highlighting its strong local incentives and infrastructure support. While the remaining top 10 cities also show notable numbers, the gap indicates **Seattle's** clear leadership in promoting EVs.

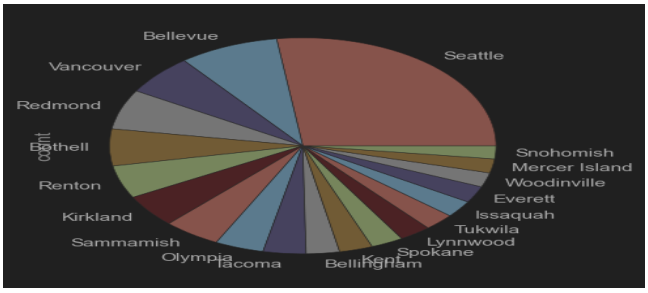

Figure 14 - Electric Range Pie Chart

The **pie chart** shows that **Seattle** dominates the electric range distribution, indicating that its residents generally own vehicles with the highest **electric ranges**. This may reflect the city's

economic prosperity and strong commitment to sustainability, promoting investment in long-range **electric vehicles.**

Temporal Analysis

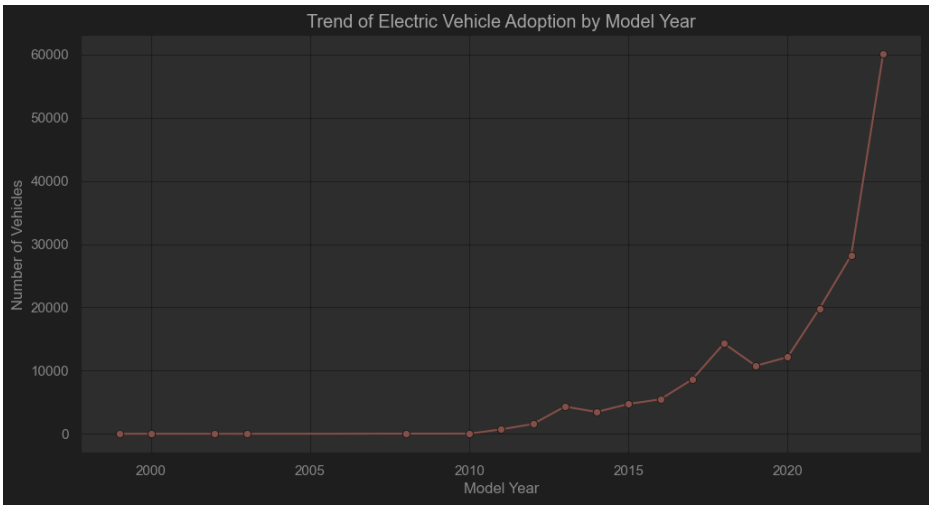

Figure 15 - Trend of Electric Vehicle Adoption

The trend of **electric vehicle (EV)** adoption shows significant exponential growth, driven by consumer awareness, technological advancements, and government incentives, highlighting the industry's shift toward sustainability.