Faculty of Engineering & Technology

Department of Electrical & Computer Engineering

ENCS5341 – Machine Learning and Data Since

First Semester 2024/2025

Assignment #1

Prepared by

Jalila Muadi – 1201611

Section No. 1

Instructor: Dr. Yazan Abu Farha

BIRZEIT

October 2024

Abstract

The primary objective of this report is to thoroughly explore the Electric Vehicle Population Data by conducting data cleaning, feature engineering, and exploratory data analysis (EDA) to uncover trends, insights, and potential patterns in the EV market within Washington State. The dataset, consisting of over 210,000 entries across 17 distinct features, provides a large and diverse sample suitable for detailed statistical and spatial analysis.

Table of Contents

List of Figures

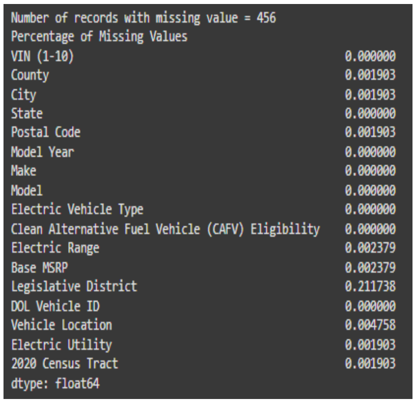
Introduction

The rapid growth in electric vehicle (EV) adoption reflects a significant shift towards sustainable and eco-friendly transportation. In support of this transition, the State of Washington has compiled a detailed dataset titled "Electric Vehicle Population Data," available through Data.gov. This dataset provides a comprehensive snapshot of battery electric vehicles (BEVs) and plug-in hybrid electric vehicles (PHEVs) registered in Washington State. The data includes crucial information about each vehicle's make, model, type, electric range, model year, and registration details, with additional geographic and demographic indicators such as county, city, postal code, and 2020 Census tract. This analysis addresses the following core areas:

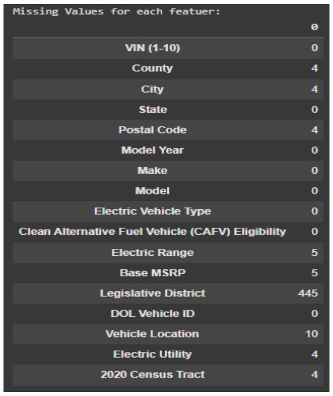
Data Cleaning and Feature Engineering:

In analyzing the Electric Vehicle Population dataset, I first examined the extent and distribution of missing values across all features. The dataset comprises 210,165 entries and 17 columns, providing detailed information about each vehicle's registration. Upon reviewing the data, I identified a few columns with minor missing values. For example, the County, City, Postal Code, Electric Range, Base MSRP, and Electric Utility columns all show minimal missing data, each with fewer than five missing values, accounting for approximately 0.002% of the dataset. Meanwhile, Legislative District has around 445 missing entries, representing 0.21% of the dataset, and the total missing records is about 456. Additionally, Vehicle Location shows a slightly higher number of missing values (10 entries) than most other columns but still represents only around 0.005% of the data. The dataset contains no empty records and no duplicate entries.

A bar chart was generated, showing the count of missing values for each feature. As shown in the figure below



```
Number of records with missing value = 456
Percentage of Missing Values
VIN (1-10)                                         0.000000
County                                             0.001903
City                                               0.001903
State                                              0.000000
Postal Code                                        0.001903
Model Year                                         0.000000
Make                                               0.000000
Model                                              0.000000
Electric Vehicle Type                              0.000000
Clean Alternative Fuel Vehicle (CAFV) Eligibility  0.000000
Electric Range                                     0.002379
Base MSRP                                          0.002379
Legislative District                               0.211738
DOL Vehicle ID                                     0.000000
Vehicle Location                                   0.004758
Electric Utility                                   0.001903
2020 Census Tract                                  0.001903
dtype: float64
```



Missing Values for each featuer:

	0
VIN (1-10)	0
County	4
City	4
State	0
Postal Code	4
Model Year	0
Make	0
Model	0
Electric Vehicle Type	0
Clean Alternative Fuel Vehicle (CAFV) Eligibility	0
Electric Range	5
Base MSRP	5
Legislative District	445
DOL Vehicle ID	0
Vehicle Location	10
Electric Utility	4
2020 Census Tract	4

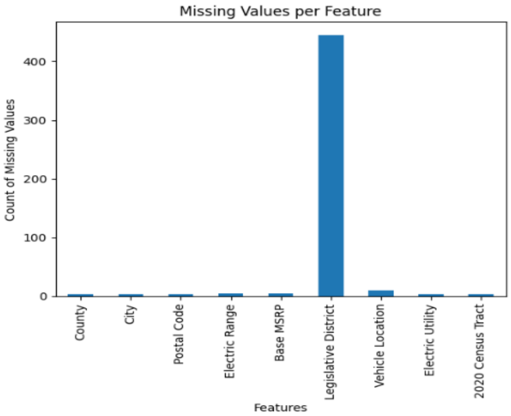
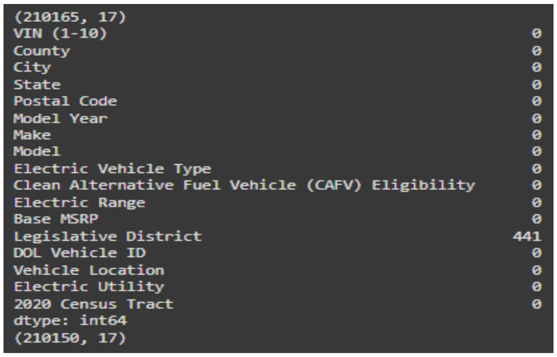


Figure 1: Summary of Missing Data

Note that most features have minimal or no missing values, with only a few columns showing notable gaps. The **Legislative District** feature has the highest count of missing values at around 445, while other columns like County, City…, each have fewer than five missing entries.

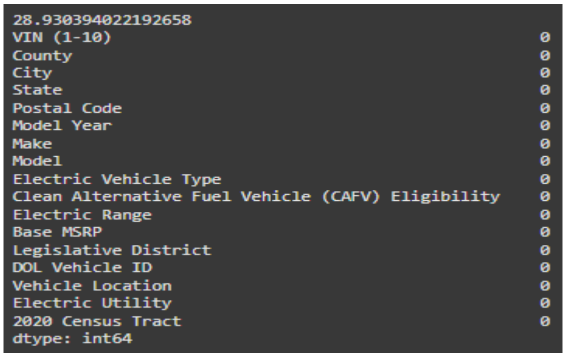
To address the missing data in the dataset, various techniques were applied based on the amount and distribution of missing values in each feature. First, features with a small number of missing values such as County, City …, were handled through row deletion. This step reduced the dataset size from 210,165 to 210,150 entries, effectively removing 15 records with minimal missing data across these columns. After this step, the only remaining feature with missing values was **Legislative District**, with 441 missing records.

Next, the **Legislative District** feature was evaluated to determine the most suitable imputation method. A bar chart of district frequencies showed an even distribution, suggesting balanced representation across legislative districts in the dataset. Additionally, a box plot analysis confirmed the absence of outliers in this feature, making mean imputation a viable option without risking distortion from extreme values. Consequently, the missing values in the **Legislative District** column were filled with the mean district value of approximately 29.



```
(210165, 17)
VIN (1-10)                                             0
County                                                 0
City                                                   0
State                                                  0
Postal Code                                            0
Model Year                                             0
Make                                                   0
Model                                                  0
Electric Vehicle Type                                  0
Clean Alternative Fuel Vehicle (CAFV) Eligibility      0
Electric Range                                         0
Base MSRP                                              0
Legislative District                                 441
DOL Vehicle ID                                         0
Vehicle Location                                       0
Electric Utility                                       0
2020 Census Tract                                      0
dtype: int64
(210150, 17)
```

Figure 2: After Clearing Minimal Missing Values



```
28.930394022192658
VIN (1-10)                                             0
County                                                 0
City                                                   0
State                                                  0
Postal Code                                            0
Model Year                                             0
Make                                                   0
Model                                                  0
Electric Vehicle Type                                  0
Clean Alternative Fuel Vehicle (CAFV) Eligibility      0
Electric Range                                         0
Base MSRP                                              0
Legislative District                                   0
DOL Vehicle ID                                         0
Vehicle Location                                       0
Electric Utility                                       0
2020 Census Tract                                      0
dtype: int64
```

Figure 3: After Handling Legislative District Missing Value

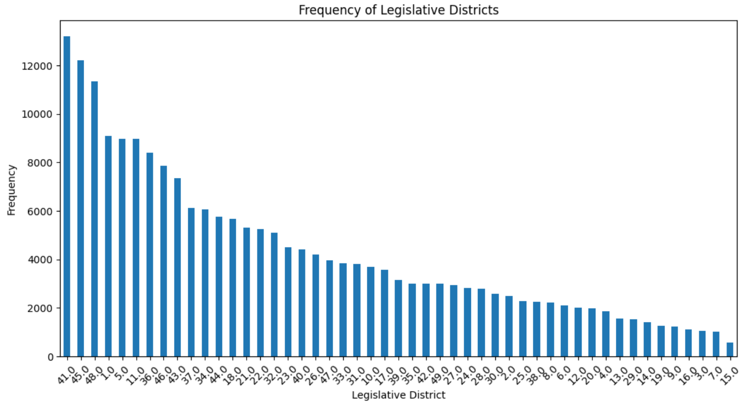


Figure 4: Frequency of Legislative Districts

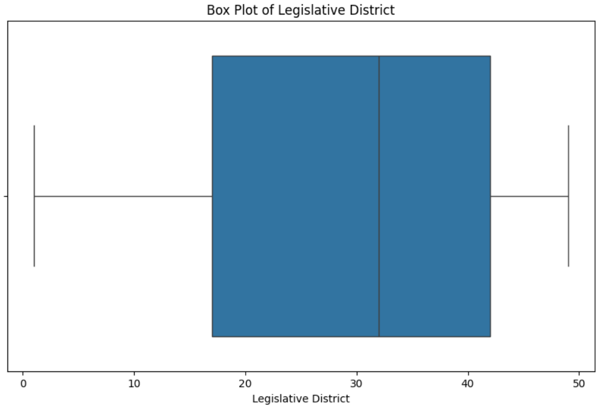
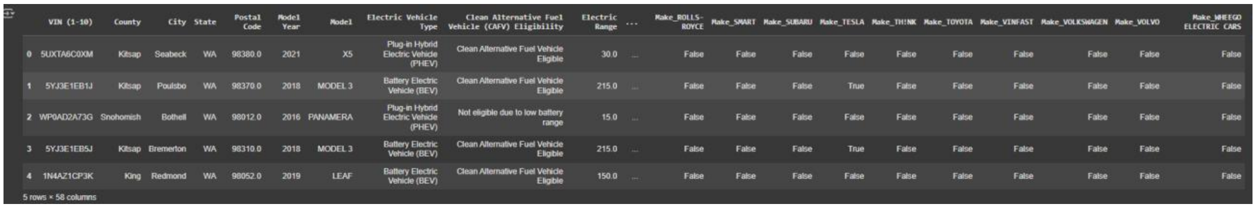


Figure 5: Box Plot of Legislative District

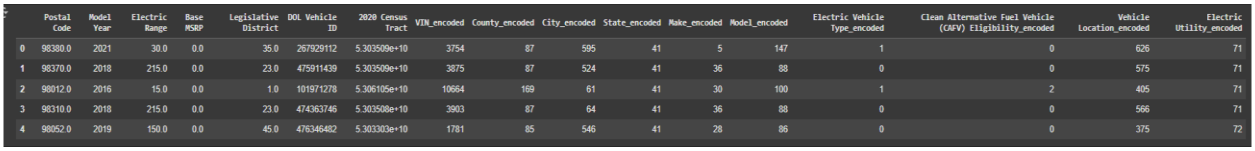
After handling the missing values, I applied feature encoding techniques to prepare the dataset for analysis. Initially, one-hot encoding was employed for categorical variables such as Make. However, I chose label encoding for features with high cardinality, to minimize the number of additional columns created. Label encoding assigns a unique integer to each category, making it particularly useful for categorical variables with an ordinal relationship. I applied label encoding to all categorical features to ensure a more streamlined dataset. The results of the feature encoding process are shown in the following figures.



	VIN (1-10)	County	City	State	Postal Code	Model Year	Model	Electric Vehicle Type	Clean Alternative Fuel Vehicle (CAFV) Eligibility	Electric Range	...	Make_ROLLS-ROYCE	Make_SMART	Make_SUBARU	Make_TESLA	Make_TH!NK	Make_TOYOTA	Make_VINFAST	Make_VOLKSWAGEN	Make_VOLVO	Make_WHEEGO ELECTRIC CARS
0	5UXTA6C0XM	Kitsap	Seabeck	WA	98380.0	2021	X5	Plug-in Hybrid Electric Vehicle (PHEV)	Clean Alternative Fuel Vehicle Eligible	30.0	...	False	False	False	False	False	False	False	False	False	False
1	5YJ3E1EB1J	Kitsap	Poulsbo	WA	98370.0	2018	MODEL 3	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	215.0	...	False	False	False	True	False	False	False	False	False	False
2	WP0AD2A73G	Snohomish	Bothell	WA	98012.0	2016	PANAMERA	Plug-in Hybrid Electric Vehicle (PHEV)	Not eligible due to low battery range	15.0	...	False	False	False	False	False	False	False	False	False	False
3	5YJ3E1EB5J	Kitsap	Bremerton	WA	98310.0	2018	MODEL 3	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	215.0	...	False	False	False	True	False	False	False	False	False	False
4	1N4AZ1CP3K	King	Redmond	WA	98052.0	2019	LEAF	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	150.0	...	False	False	False	False	False	False	False	False	False	False

5 rows × 58 columns

Figure 6: One Hot Encoding for Make Feature

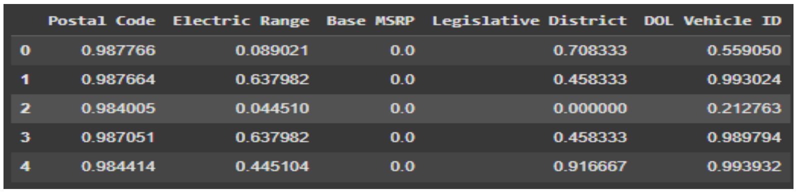


	Postal Code	Model Year	Electric Range	Base MSRP	Legislative District	DOL Vehicle ID	2020 Census Tract	VIN_encoded	County_encoded	City_encoded	State_encoded	Make_encoded	Model_encoded	Electric Vehicle Type_encoded	Clean Alternative Fuel Vehicle (CAFV) Eligibility_encoded	Vehicle Location_encoded	Electric Utility_encoded
0	98380.0	2021	30.0	0.0	35.0	267929112	5.303509e+10	3754	87	595	41	5	147	1	0	626	71
1	98370.0	2018	215.0	0.0	23.0	475911439	5.303509e+10	3875	87	524	41	36	88	0	0	575	71
2	98012.0	2016	15.0	0.0	1.0	101971278	5.306105e+10	10664	169	61	41	30	100	1	2	405	71
3	98310.0	2018	215.0	0.0	23.0	474363746	5.303508e+10	3903	87	64	41	36	88	0	0	566	71
4	98052.0	2019	150.0	0.0	45.0	476346482	5.303303e+10	1781	85	546	41	28	86	0	0	375	72

Figure 7: Label Encoding for All Categorical Features

Note that the dataset has significantly expanded from 17 columns to 58 columns as a result of implementing one-hot encoding. In contrast, label encoding preserved the original column structure.

The next step in the analysis was to normalize the numerical features to ensure they are on a comparable scale, which is particularly important for certain analytical methods. To achieve this, I utilized the Min-Max Scaler, which transforms the selected features to a range between 0 and 1. The features normalized in this process as shown below.



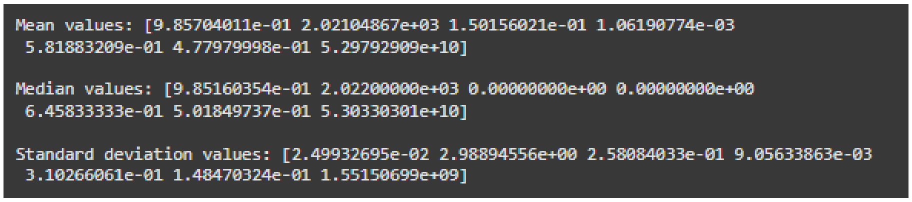
	Postal Code	Electric Range	Base MSRP	Legislative District	DOL Vehicle ID
0	0.987766	0.089021	0.0	0.708333	0.559050
1	0.987664	0.637982	0.0	0.458333	0.993024
2	0.984005	0.044510	0.0	0.000000	0.212763
3	0.987051	0.637982	0.0	0.458333	0.989794
4	0.984414	0.445104	0.0	0.916667	0.993932

Figure 8: Numerical Features After Normalization

Exploratory Data Analysis:

Descriptive Statistics

To understand the numerical features in the dataset, I calculated summary statistics, including the mean, median, and standard deviation for each numerical column. The selected numerical features included Postal Code, Model Year, Electric Range, Base MSRP, Legislative District, and DOL Vehicle ID.



```
Mean values: [9.85704011e-01 2.02104867e+03 1.50156021e-01 1.06190774e-03
 5.81883209e-01 4.77979998e-01 5.29792909e+10]

Median values: [9.85160354e-01 2.02200000e+03 0.00000000e+00 0.00000000e+00
 6.45833333e-01 5.01849737e-01 5.30330301e+10]

Standard deviation values: [2.49932695e-02 2.98894556e+00 2.58084033e-01 9.05633863e-03
 3.10266061e-01 1.48470324e-01 1.55150699e+09]
```

Figure 9: Summary Statistics for Each Numerical Feature

The resulting summary statistics provided insights into the central tendencies and variability of the features. To visualize these statistics, I plotted histograms for each numerical feature, overlaying the mean, median, and standard deviation values as shown below:

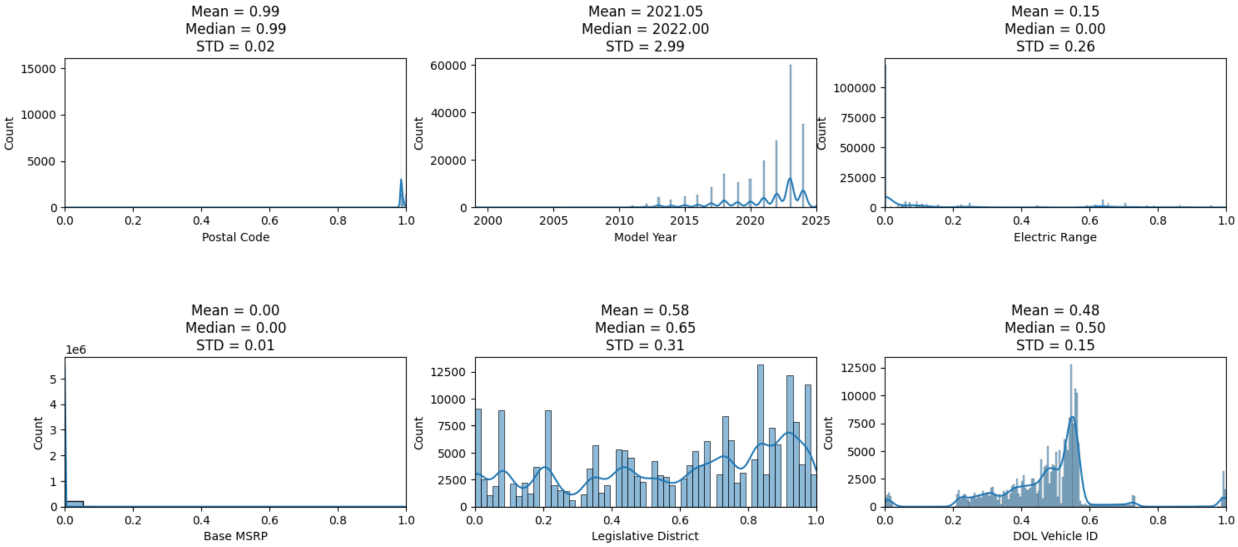


Figure 10: Histograms for Each Numerical Feature

These visualizations not only illustrate the distributions of each feature but also highlight the relationships between the mean, median, and standard deviation, providing a clearer understanding of the underlying data.

Spatial Distribution

To visualize the spatial distribution of electric vehicles across locations, I employed Folium to create an interactive map. This map gives a better understanding of where different makes and models of EVs are registered throughout Washington State. The process involved extracting longitude and latitude coordinates from the Vehicle Location column. To enhance performance, I sampled 5000 entries from the cleaned dataset.

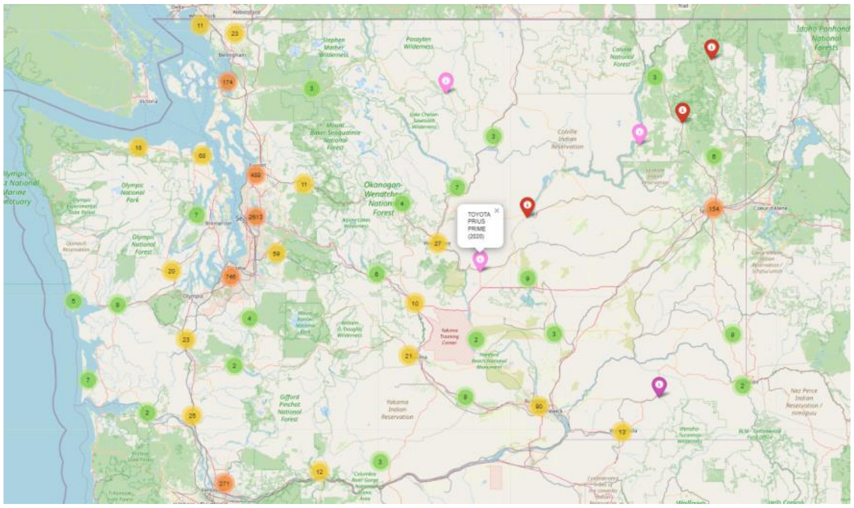


Figure 11: Spatial Distribution of EVs Across Locations

Model Popularity

To analyze the popularity of different EV models, I calculated the frequency of each model in the dataset. The results show that the Model Y and Model 3 from Tesla are the most common, with 44,037 and 32,519 registrations. This dominance of Tesla in the market is evident in the overall count of EVs, where Tesla's models collectively account for 91,376 registrations, significantly surpassing other manufacturers like Chevrolet and Nissan, which have 15,417 and 14,721 vehicles registered. I focused on the top 10 most popular models and makes to further refine the analysis, leading to the following findings:

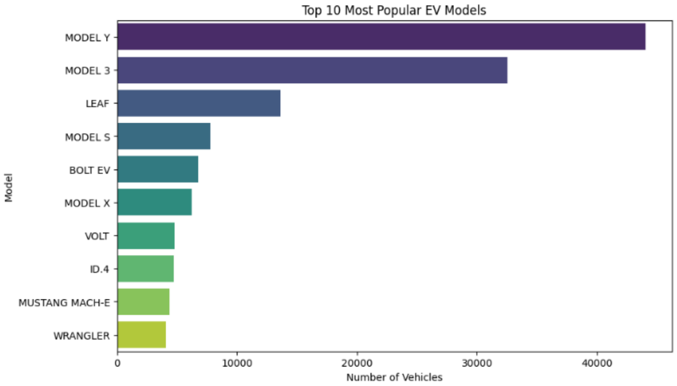


Figure 12: Top 10 Most Popular EV Models

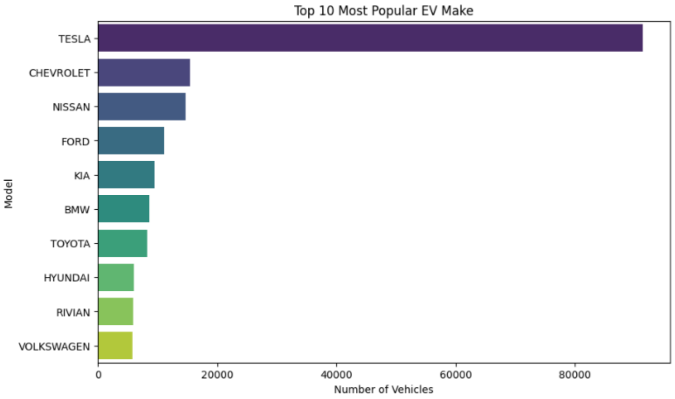


Figure 13: Top 10 Most Popular EV Make

To explore trends based on vehicle type, I filtered the dataset into Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs). The bar plots below illustrate the top 10 most popular makes for both BEV and PHEV categories:

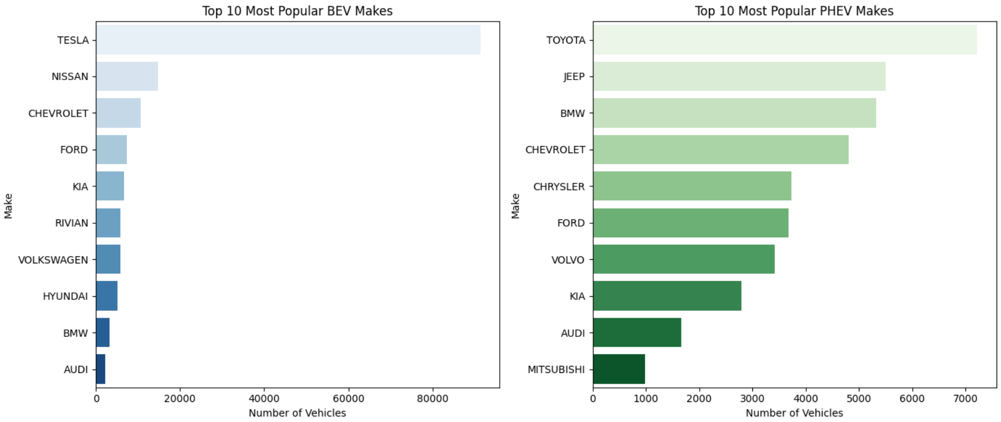


Figure 14: Top 10 Most Popular BEV and PHEV Categories

In the analysis of EV model popularity, Tesla dominates the BEV market, highlighting its strong brand presence. The PHEV market, however, is more diversified, with brands like Toyota, Jeep, and BMW competing closely, indicating a lack of a single leading brand. Some brands, such as Chevrolet and Ford, appear in both categories, reflecting a dual strategy to appeal to both fully electric and hybrid consumers. These trends indicate that the BEV market is dominated by established brands like Tesla. On other hand, the PHEV market is more varied and competitive, appealing to consumers interested in the flexibility of hybrid vehicles.

Correlation Analysis

To understand the relationships between numerical features in the dataset, I found the correlation matrix shown in the figure below:

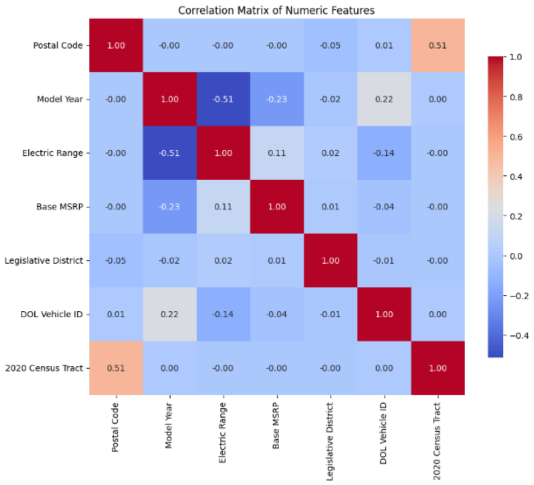


Figure 15: Correlation Matrix of Numeric Features

The correlation analysis reveals a moderate positive correlation (0.51) between *Postal Code* and *2020 Census Tract*, suggesting a geographical association, as both features serve as location identifiers. There is a moderately negative correlation (-0.51) between *Model Year* and *Electric*

Range, implying that newer EV models might have slightly shorter electric ranges, which could reflect design trends or efficiency improvements in specific models. A weak negative correlation (-0.23) between *Model Year* and *Base MSRP* suggests that newer models tend to have slightly lower prices, potentially due to advancements in technology or increased manufacturing efficiencies in recent years. Similarly, a weak positive correlation (0.11) between *Electric Range* and *Base MSRP* suggests that vehicles with higher ranges may have marginally higher base prices, likely due to the costs associated with larger or more advanced batteries. The *DOL Vehicle ID* shows weak associations with other features, such as *Model Year* (0.22) and *Electric Range* (-0.14), indicating that it operates mostly independently of other variables. Most other feature pairs exhibit near-zero correlations, suggesting minimal linear relationships and likely independence or weak relationships between those features.

Visualization

Data Exploration Visualizations

This collection of charts in figure below shows a data exploration of electric vehicles (EVs), examining attributes like electric range, MSRP (Manufacturer's Suggested Retail Price), model year, make, and city distribution.

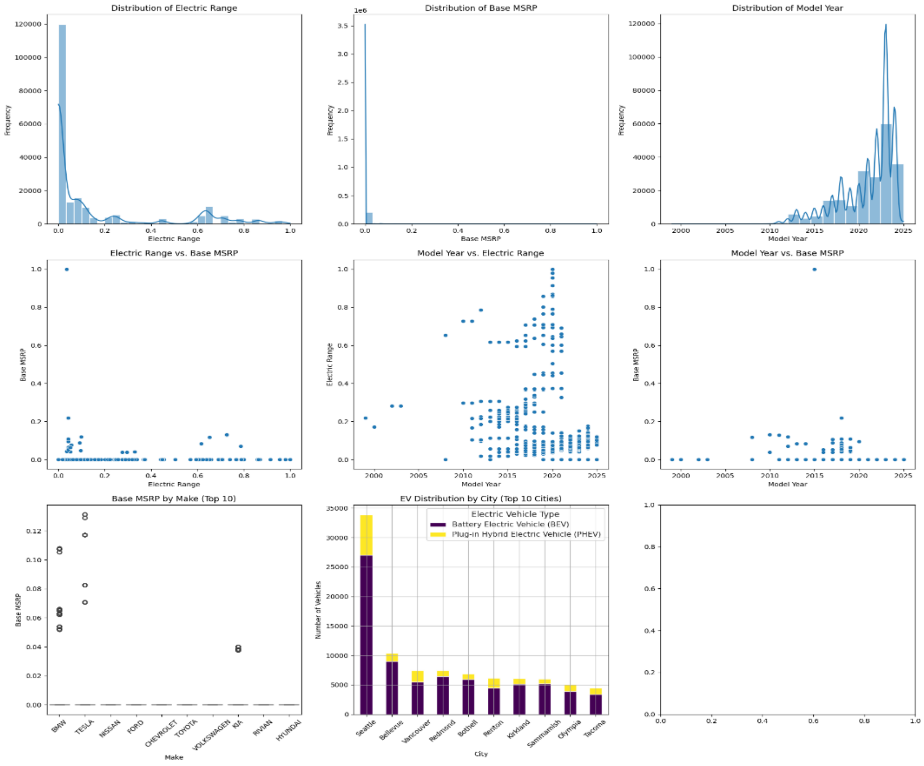


Figure 16: Data Exploration of Electric Vehicles

The distribution of electric range reveals that most EVs have shorter ranges, while a few outliers achieve higher ranges. MSRP distribution is highly skewed toward lower prices, indicating affordability for many models. EV model years show a steady increase in adoption since the early 2010s, peaking around 2020-2023. In the electric range versus MSRP plot, higher ranges are available across a wide price spectrum, with more affordable models generally having shorter ranges. The model year versus electric range plot highlights a trend of improved range over time, while MSRP does not show a significant upward trend with newer models, despite increased diversity in price points. Comparing MSRP by manufacturer, brands like BMW and Tesla show higher median MSRPs among the top 10 brands. The distribution of EVs by city indicates Seattle leads in EV adoption, followed by cities like Los Angeles and San Francisco, with Battery Electric Vehicles (BEVs) outnumbering Plug-in Hybrid Electric Vehicles (PHEVs) in most areas.

Comparative Visualization

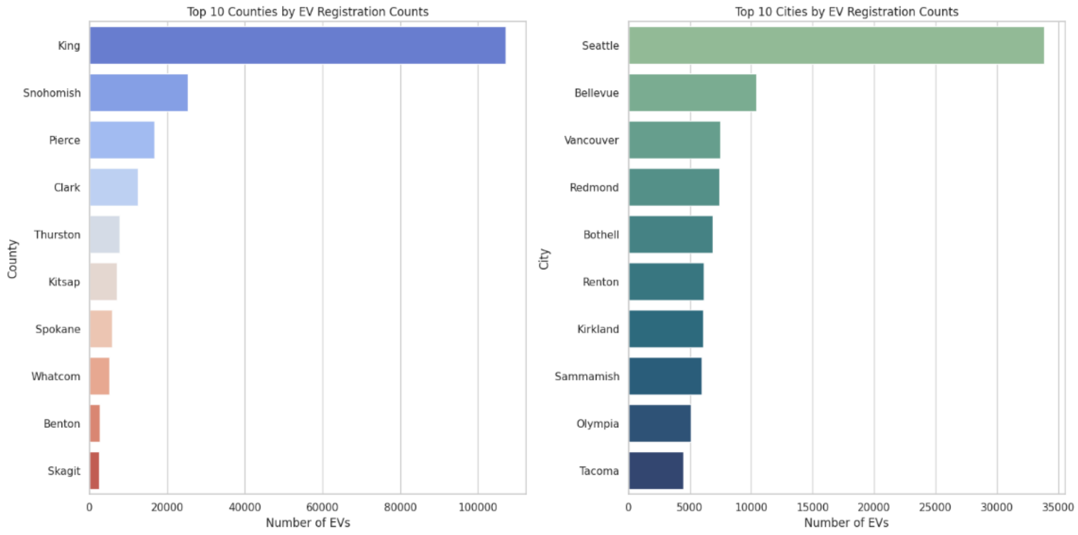


Figure 17: EV registration distribution across the top counties and cities

The bar chart on the left shows that King County has the highest number of EV registrations by a large margin, followed by Snohomish and Pierce counties. This suggests a significant concentration of EVs in these areas, potentially due to urban density.

The bar chart on the right indicates that Seattle leads in EV registrations among cities, followed by Bellevue, Vancouver, and Redmond. This pattern aligns with the county distribution, as King County includes many of these high-registration cities.

These comparative charts illustrate that EV adoption is heavily centered around major urban areas, with the highest adoption rates in densely populated cities and counties.

Conclusion

This assignment looks at the "Electric Vehicle Population Data" for Washington State, showing how more people are adopting electric vehicles (EVs). The data gives useful insights into the state's EV situation. The main findings highlight the importance of having good quality data. The dataset had very few missing values, making it easier to clean and prepare for analysis. We used methods like deleting rows and filling in missing values with averages to ensure the data was strong for our study.

By normalizing the data using Min-Max scaling and applying the right coding techniques, we made the dataset much easier to work with. This preparation was important for making accurate comparisons and analyses. Our exploration of the data shows interesting trends, showing that Tesla is the most popular brand for Battery Electric Vehicles (BEVs), while the Plug-in Hybrid Electric Vehicle (PHEV) market has a wider variety of popular brands.

We also looked at where EV registrations are happening and found a lot of them are in urban areas, especially in King County and Seattle. This suggests that living in a city, with better access to charging stations and more environmental awareness, encourages people to adopt EVs. Our analysis also found interesting links between different features, like how newer models might have better electric ranges and still be competitively priced. These insights can help predict future trends and guide manufacturers in meeting customer needs.