



Faculty of Engineering and Technology
Department of Electrical and Computer Engineering
Machine Learning and Data Science - ENCS5341

Assignment #1

Prepared by:

Ro'a Nafi 1201959

Zainab Jaradat 1201766

Instructor:

Dr. Ismail Khater

Date: October 30, 2024

Contents

1. Dataset Description	3
2. Data Cleaning and Feature Engineering	3
2.1. Document Missing Values	3
2.2. Missing Value Strategies	3
2.3. Feature Encoding	4
2.4. Normalization	4
3. Exploratory Data Analysis (EDA)	5
3.1. Descriptive Statistics	5
3.2. Spatial Distribution Analysis	5
3.3. Model Popularity Analysis	6
3.4. Correlation Analysis	7
4. Visualization	8
4.1. Data Exploration Visualizations	8
4.2. Comparative Visualization	10
5. Additional Analysis	10
6 . Conclusion	11

1. Dataset Description

This dataset shows the Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) that are currently registered through the Washington State Department of Licensing (DOL). Key features include VIN, county, city, make, model, electric type, and electric range, offering insights into the distribution and characteristics of EVs across the state.

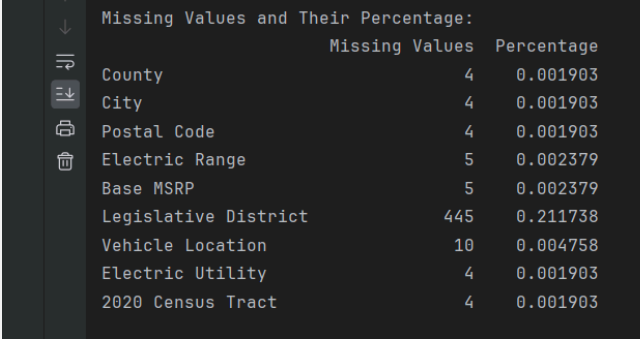
The dataset contains 210165 rows and 17 columns, combining categorical and numerical features. Data types include `float64` (5), `int64` (2), and `object` (10).

- **int64 Columns:** Model Year, DOL Vehicle ID.
- **object Columns:** VIN (1-10), County, City, State, Make, Model, Electric Vehicle Type, Clean Alternative Fuel Vehicle (CAFV) Eligibility, Vehicle Location, Electric Utility.
- **float64 Columns:** Postal Code, Electric Range, Base MSRP, Legislative District, 2020 Census Tract.

2. Data Cleaning and Feature Engineering

2.1. Document Missing Values

We checked for missing values in the dataset and found that a few columns have missing entries. Most columns have very low percentages of missing data (around 0.002% to 0.005%). Only the "Legislative District" column has a slightly higher missing rate at 0.21%.



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

2.2. Missing Value Strategies

We tested several methods to handle missing values, each with its strengths and weaknesses. **Mean and median imputation** worked well for numeric columns, preserving the dataset size and leaving only **22** missing values in non-numeric columns. Median imputation is especially useful for skewed data, as it's less affected by outliers. **Dropping rows with missing values** completely removed all gaps, but at the cost of reducing the dataset by **456** rows, which could impact the overall analysis. **Mode imputation**, which fills non-numeric columns with the most common (or frequent) value, addressed some missing values but still left **463** gaps. This approach is effective when certain categories are more frequent, but it may not be enough when

there's a high variety of data values. These results were obtained by using Python to calculate the remaining missing values after each method.

<pre> Missing values per column: 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 (CAEV) Eligibility 0 Electric Range 5 Base MSRP 5 Legislative District 445 DOI Vehicle ID 0 Vehicle Location 10 Electric Utility 4 2020 Census Tract 4 dtype: int64 Missing values per column after numerical imputation: VIN (1-10) 0 County 4 City 4 State 0 Postal Code 0 Model Year 0 Make 0 Model 0 Electric Vehicle Type 0 Clean Alternative Fuel Vehicle (CAEV) Eligibility 0 Electric Range 0 Base MSRP 0 Legislative District 0 DOI Vehicle ID 0 Vehicle Location 10 Electric Utility 4 2020 Census Tract 0 dtype: int64 Total missing values after numerical imputation: 22 </pre>	<pre> Missing values per column: 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 (CAEV) Eligibility 0 Electric Range 5 Base MSRP 5 Legislative District 445 DOI Vehicle ID 0 Vehicle Location 10 Electric Utility 4 2020 Census Tract 4 dtype: int64 Missing values per column after non-numerical imputation: VIN (1-10) 0 County 0 City 0 State 0 Postal Code 4 Model Year 0 Make 0 Model 0 Electric Vehicle Type 0 Clean Alternative Fuel Vehicle (CAEV) Eligibility 0 Electric Range 5 Base MSRP 5 Legislative District 445 DOI Vehicle ID 0 Vehicle Location 0 Electric Utility 0 2020 Census Tract 4 dtype: int64 Total missing values after non numerical imputation: 463 </pre>
---	--

2.3. Feature Encoding

We used one-hot encoding on the columns Make, Model, County, and City because they are important categorical features in our dataset. The number of columns increased from 17 to 1170. One-hot encoding changes categorical data into numbers, which makes it easier to use for machine learning. The dataset still has 210165 rows, but now it has many new columns for each unique category in the original features. This step is important to prepare the data for more analysis and predictive modeling.

```

Original Data Shape: (210165, 17)
Encoded Data Shape: (210165, 1170)

```

	VIN (1-10)	State	Postal Code	...	City_Yelm	City_Yorktown	City_Zillah
0	5UXTA6C0XM	WA	98380.0	...	False	False	False
1	5YJ3E1EB1J	WA	98370.0	...	False	False	False
2	WP0AD2A73G	WA	98012.0	...	False	False	False
3	5YJ3E1EB5J	WA	98310.0	...	False	False	False
4	1N4AZ1CP3K	WA	98052.0	...	False	False	False

```

[5 rows x 1170 columns]
Process finished with exit code 0

```

One-hot encoding

2.4. Normalization

Normalization is a technique that scales numerical features to a common range, typically [0, 1] or [-1, 1]. We normalize all the numerical features in our dataset using Min-Max Scaling. The dataset shape remains the same, but the values within these columns are scaled to a range between 0 and 1, ensuring that all features contribute equally to the analysis. This photo shows the Head of the data :

```

PS C:\Users\miditech\OneDrive\Desktop> python machine.py
Original Data Shape: (210165, 17)
Normalized Data Shape: (210165, 17)

```

	VIN (1-10)	County	City	State	Postal Code	...	Legislative District	DOL	Vehicle ID	Vehicle Location	Electric Utility	2020 Census Tract
0	5UXTAGC00M	Kitsap	Seabeck	WA	0.987766	...	0.708333	267929112	POINT (-122.8728334 47.5798304)	PUGET SOUND ENERGY INC	0.945730	
1	5YJ3E1EB1J	Kitsap	Poulsbo	WA	0.987664	...	0.458333	475911439	POINT (-122.6368884 47.7469547)	PUGET SOUND ENERGY INC	0.945730	
2	WP0AD2A73G	Snohomish	Bothell	WA	0.984005	...	0.000000	101971278	POINT (-122.206146 47.839957)	PUGET SOUND ENERGY INC	0.946202	
3	5YJ3E1EB5J	Kitsap	Bremerton	WA	0.987051	...	0.458333	474363746	POINT (-122.6231895 47.5930874)	PUGET SOUND ENERGY INC	0.945730	
4	1N4AZ1CP3K	King	Redmond	WA	0.984414	...	0.916667	476346482	POINT (-122.13158 47.67858)	PUGET SOUND ENERGY INC CITY OF TACOMA - (WA)	0.945693	

[5 rows x 17 columns]

Normalization numerical features

3. Exploratory Data Analysis (EDA)

3.1. Descriptive Statistics

The dataset's numerical features show some interesting patterns. Postal Code and 2020 Census Tract have high mean and median values, meaning they are not very spread out. Model Year has a median slightly above the mean, suggesting a small left skew. Electric Range and Base MSRP have many low or zero values, making their distributions right-skewed. Legislative District and DOL Vehicle ID have moderate variation, while 2020 Census Tract has low variation, showing consistency in values. This summary helps to understand the basic distribution and spread of each feature.

Summary statistics for numerical features:

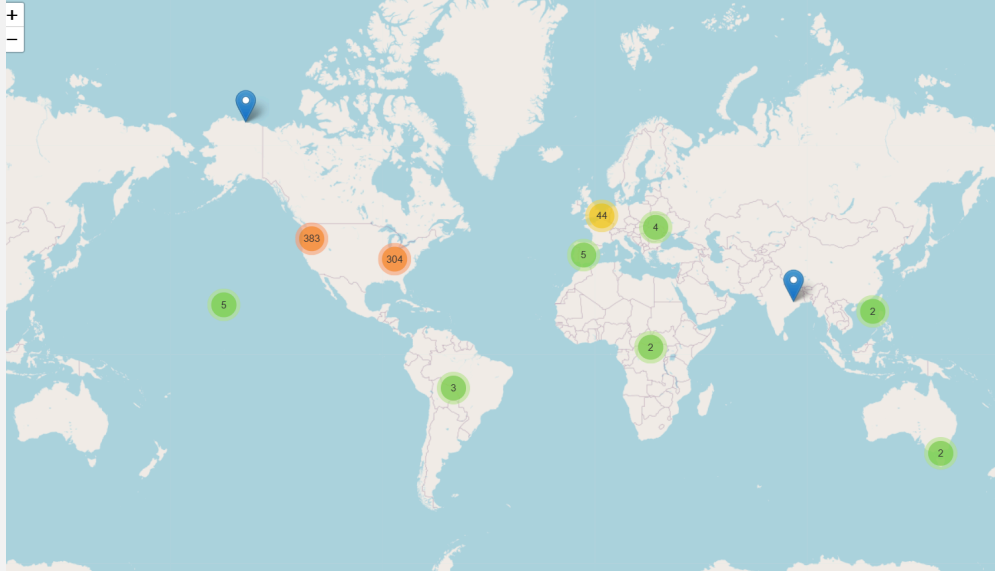
	mean	median	std
Postal Code	0.985704	0.985160	0.024992
Model Year	0.848025	0.884615	0.114959
Electric Range	0.150151	0.000000	0.258079
Base MSRP	0.001062	0.000000	0.009057
Legislative District	0.581874	0.645833	0.310263
DOL Vehicle ID	0.477982	0.501850	0.148472
2020 Census Tract	0.944716	0.945693	0.028198

Descriptive Statistics

3.2. Spatial Distribution Analysis

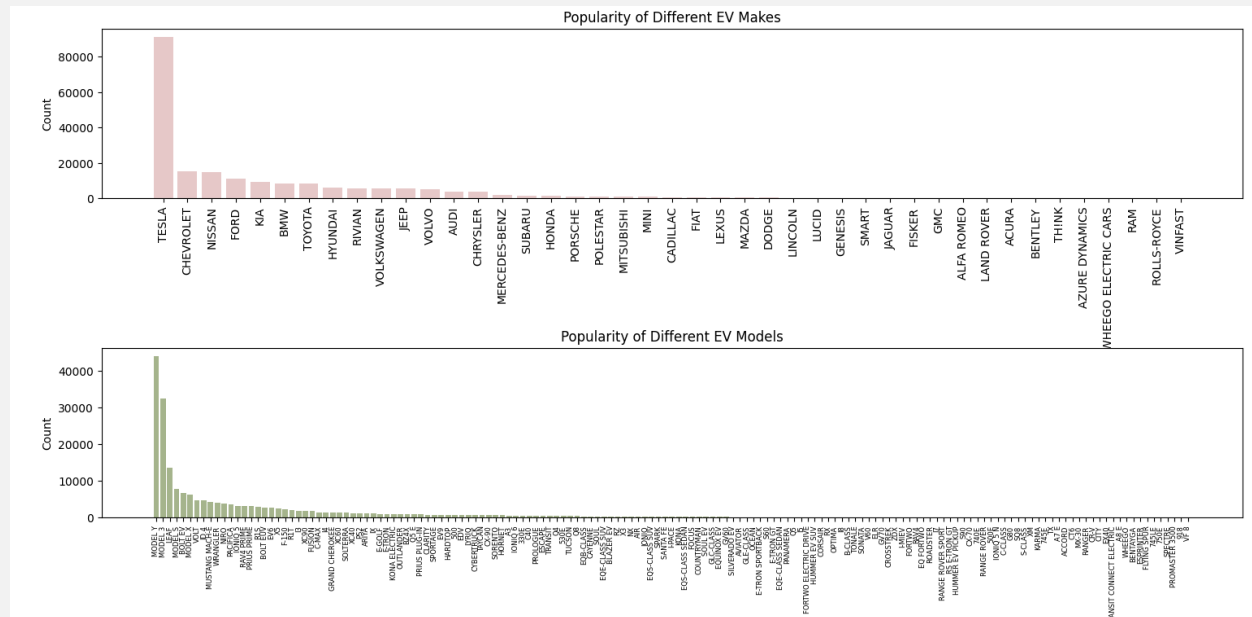
We created a map to show the distribution of electric vehicles (EVs) across different cities. First, we loaded a cleaned dataset with EV counts for various cities. We identified the city columns and extracted the city names and their respective EV counts. Then, we used a geocoding service to get the latitude and longitude for each city. After filtering out cities with missing coordinates, we created a base map centered on the United States. We added markers for each city, showing the EV count, and clustered them for better visualization. Finally, we saved the map as an HTML file for easy viewing. (Please check the map in this [link](#)).

We find that EVs are primarily concentrated in Washington State, especially around Seattle. Globally, there are notable clusters in North America and Europe, reflecting higher EV adoption in these regions.



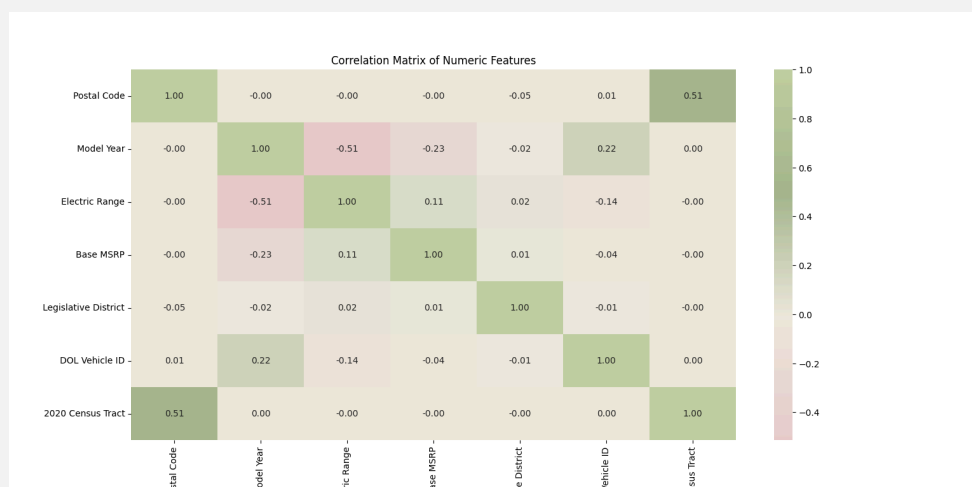
3.3. Model Popularity Analysis

In this analysis, we explored the popularity of EV makes and models using a dataset of EV registrations. We consolidated data from multiple columns into single 'Make_' and 'Model_' columns for clarity. After grouping and counting the data by EV make, we plotted the results to visualize the popularity. The results show Tesla as the leading make, followed by Chevrolet, Nissan, and Ford. Model analysis further highlights popular vehicles, with top models aligning closely with the leading makes. This provides a clear view of consumer preferences and market trends, showcasing Tesla's strong presence in the EV market.



3.4. Correlation Analysis

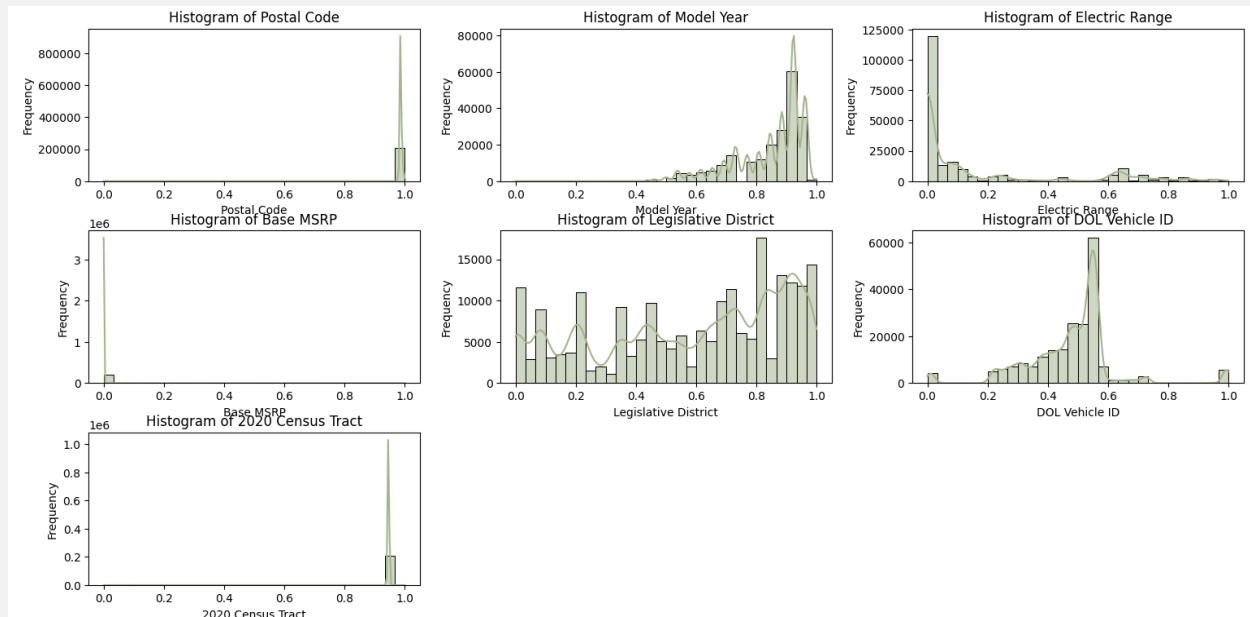
This analysis examined the correlations between numeric features in an electric vehicle dataset. Using Python libraries—Pandas for data manipulation, Matplotlib for plotting, and Seaborn for advanced visualizations—we computed and visualized a correlation matrix. This matrix, displayed as a heatmap, shows how pairs of numeric features relate to each other, with correlation coefficients ranging from -1 to 1. Notably, there is a strong negative correlation (-0.51) between 'Model Year' and 'Electric Range,' indicating that newer models tend to have longer electric ranges. Conversely, 'Postal Code' and '2020 Census Tract' show a positive correlation (0.51), suggesting a geographical clustering of the data.



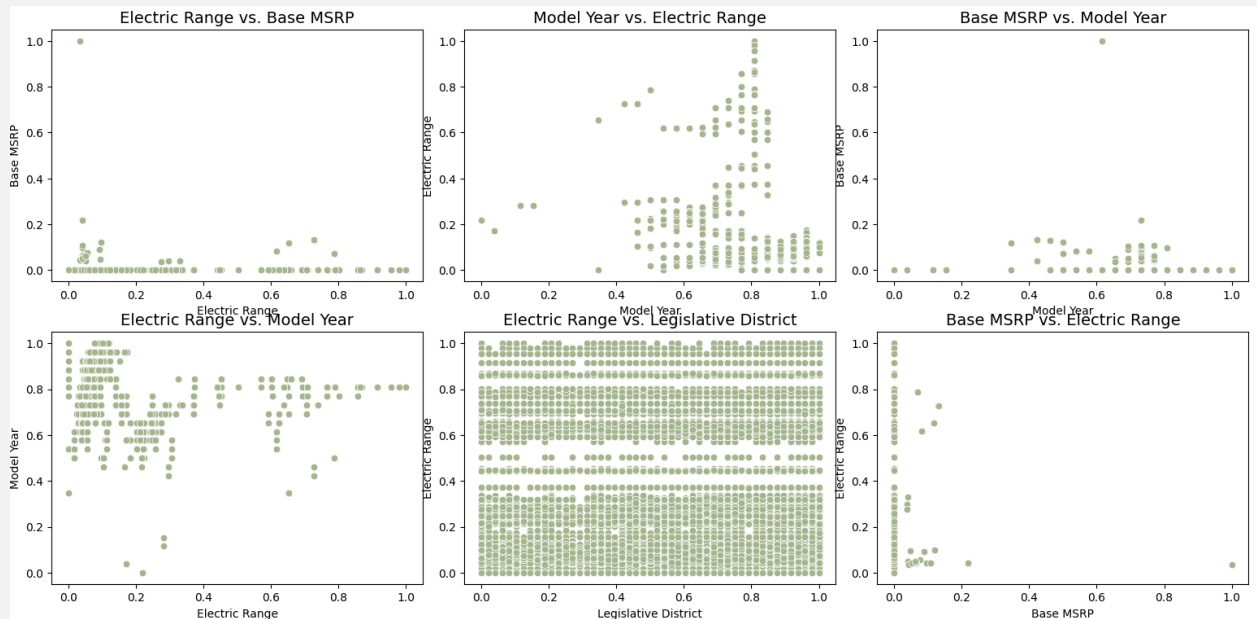
4. Visualization

4.1. Data Exploration Visualizations

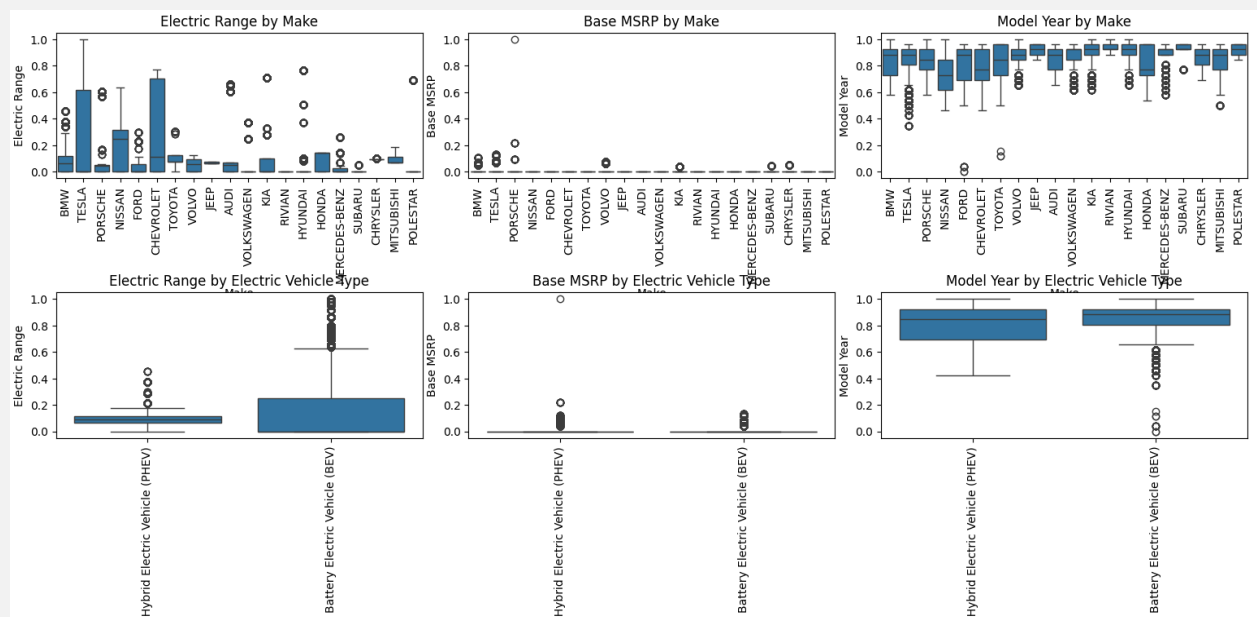
The histograms show that many features have values concentrated in specific areas. For example, Electric Range and Base MSRP are mostly on the lower end, with a few high-value outliers. Model Year has an upward trend, suggesting an increase in recent electric vehicle registrations. The concentration in Postal Code and 2020 Census Tract might be due to the way data is recorded or encoded.



The scatter plots reveal that most EVs are clustered at lower ranges and prices, with a few high-range, high-price models. Newer models generally show improved ranges, hinting at technological advancements, while prices have remained mostly stable across years, aside from a few recent high-cost outliers. Range distribution across legislative districts shows no clear regional preference, suggesting range and price are influenced more by model advancements than location.

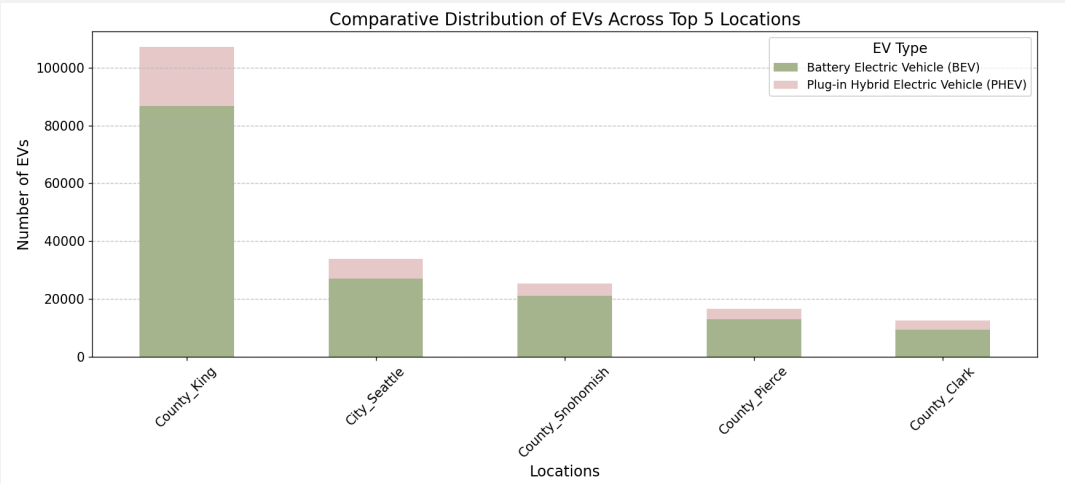


The boxplots reveal that Tesla leads in the electric range, with the highest values, while Chevrolet offers moderate-range options with some variability. BEVs generally have higher ranges than PHEVs, and model years are mostly recent across all types and makes. Prices (MSRP) are generally low for most brands, with a few high-priced outliers, especially among BEVs.



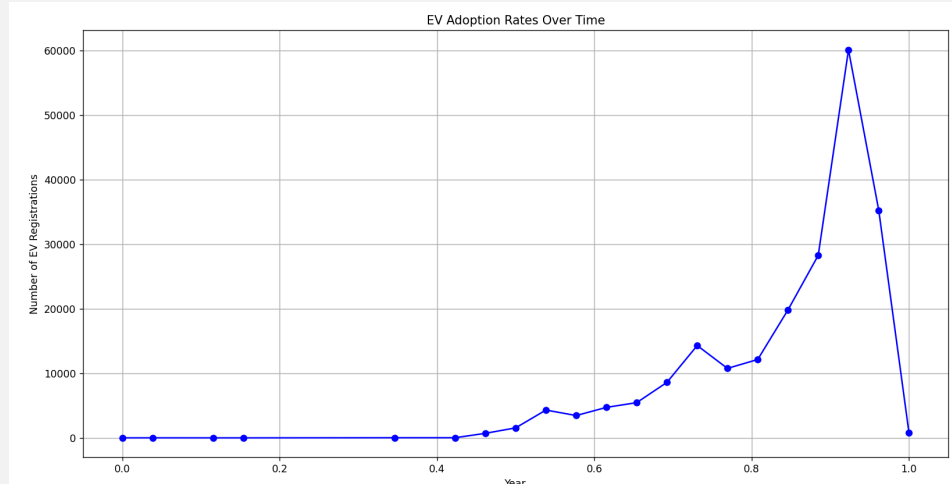
4.2. Comparative Visualization

The bar chart illustrates the number of electric vehicles, segmented into Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs), across five key regions including three counties (King, Snohomish, Pierce) and one city (Seattle). Notably, King County showcases the highest number of EVs, predominantly BEVs, indicating a strong preference for fully electric models possibly due to robust infrastructure and incentives. Seattle, while smaller in total numbers, maintains a balanced mix of both EV types, reflecting urban adoption trends. The other counties, although contributing less to the totals, demonstrate varying preferences for EV types. This distribution highlights regional disparities in EV adoption which may be driven by local policies, economic factors, and the availability of charging infrastructure. This could inform enhancements in charging stations, tax incentives, and educational campaigns to boost EV adoption, especially in underperforming regions.



5. Additional Analysis

In this part we explored the temporal trends in electric vehicle (EV) adoption by plotting the annual registration data, revealing a significant overall increase in EV registrations over the years, with a notable spike towards the end of the timeline, indicating a surge in EV adoption. This increase might reflect growing environmental awareness, improvements in EV infrastructure, or favorable policies. We also planned to examine the popularity of specific EV models over time; however, this part of the analysis is contingent on the availability of a 'Model' column in the dataset to track how preferences for different models have evolved, identifying trends that might inform future EV marketing and development strategies.



6 . Conclusion

In this project, we analyzed electric vehicle (EV) data from Washington State, focusing on data cleaning, feature engineering, and exploratory data analysis. We documented missing values, addressing them through imputation (using mean, median, and mode) and row removal, then encoded categorical features with one-hot encoding and normalized numerical data. EDA revealed Tesla as the leader in electric range, with BEVs generally offering higher ranges than PHEVs. Descriptive statistics, spatial analysis, and visualizations (histograms, scatter plots, boxplots, and bar charts) helped identify key patterns, including the prominence of recent, moderately priced models, though some high-cost outliers exist. King County shows strong EV adoption and temporal analysis reflects the growing interest in EVs, likely due to improved infrastructure and regional support for clean energy initiatives.

The results identified a notable rise in EV registrations over time, indicating growing interest in EVs, possibly due to favorable policies and an increase in charging stations. This analysis provides valuable insights into the trends and characteristics of EVs in Washington State, highlighting the need for continued support for EV infrastructure and incentives to further encourage adoption.