Machine Learning and Data Science - ENCS5341

Assigment #1

Prepared By: Ali Halayqa - 1201769

Section : 1

Instructer : Dr.Yazan Abu Farha

Mohammad shrateh – 1201369

Section : 3

Instructer : Ismail khater

Date: 10/30/2023

Abstract

The "Electric Vehicle Population Data" from Washington State is examined in this study, with a focus on important stages in data pretreatment, exploratory data analysis (EDA), and visualization. In order to get ready for analysis, we use imputation techniques to fill in missing data and encode categorical variables. Geographic and model popularity representations highlight important patterns and variances in numerical data, while descriptive statistics highlight spatial trends and preferences in EV adoption. The relationship between features is also examined via correlation analysis, which may show trends in EV range and other areas. Histograms, scatter plots, and comparative bar charts are used to effectively convey patterns and give a comprehensive picture of the electric vehicle market in Washington State.

1. **Document Missing Values**: Check for missing values and document their frequency and distribution across features.

Discussion result:

The dataset offers details on electric vehicles (EVs) registered in the state of Washington and consists of 210,165 entries and 17 attributes. The following are crucial details: electric vehicle type (PEV or battery), electric range, model year, make, model, county, city, postal code, vehicle identifying numbers (VINs), and eligibility for Clean Alternative Fuel Vehicle (CAFV) subsidies.

The missing values report indicates that only a very small percentage of the data is missing:

Legislative District: 445 missing values, or 0.21%

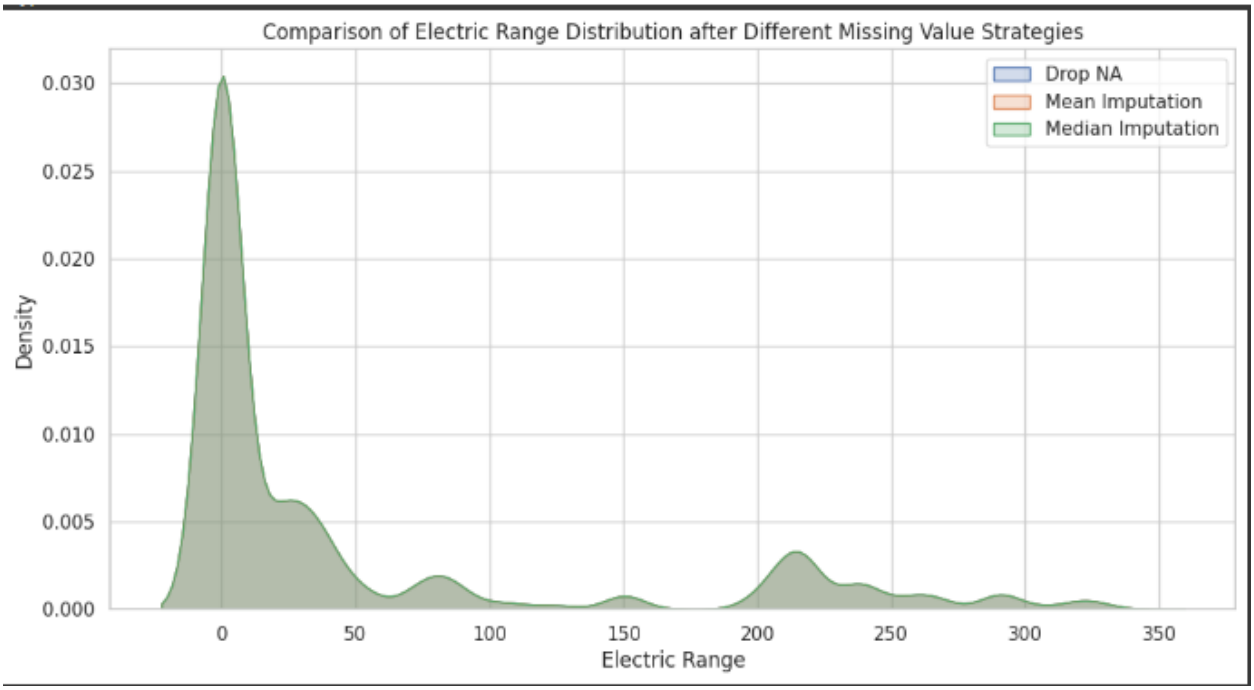
Vehicle location: 10 missing values (0.0048%)

Electric Utility, Electric Range, Base MSRP, County, City, and Postal Code, 2020 Census Tract: missing in four or five rows, accounting for 0.0024% to 0.0019% of the dataset.

The low percentage of missing data indicates that the dataset is incredibly

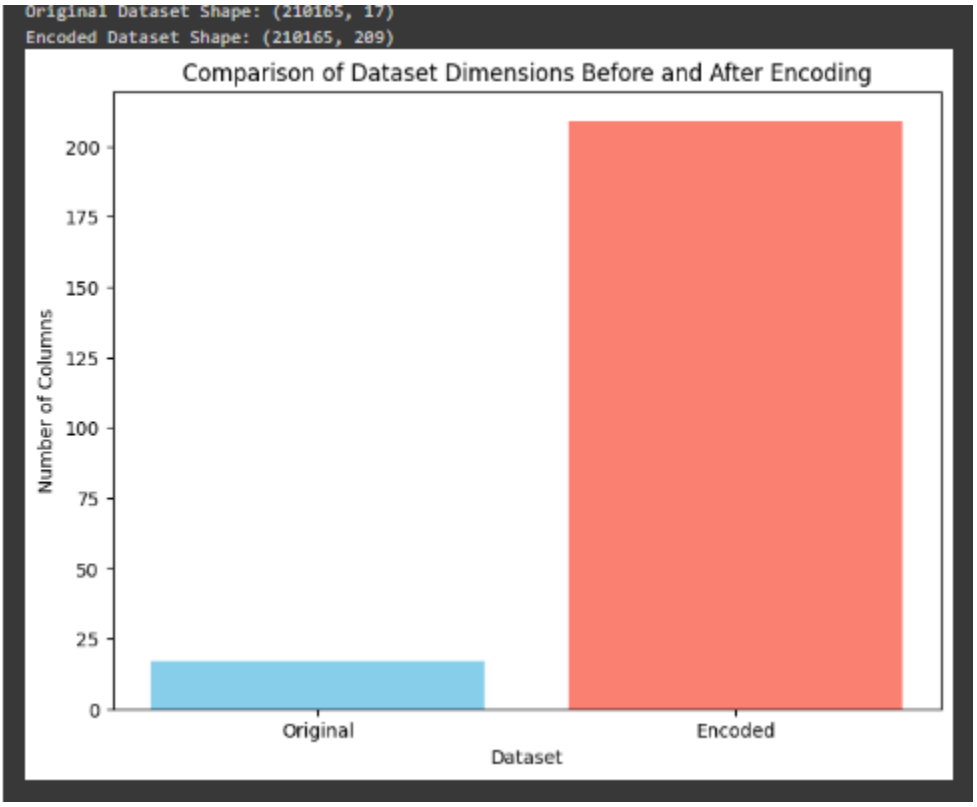
thorough. (all less than 1%), and data imputation or row exclusion for analysis should minimally impact results. Overall, the dataset provides a comprehensive basis for exploring EV trends, incentives, and regional distributions across Washington State.

2. **Missing Value Strategies**: If missing values are present, apply multiple strategies (e.g., mean/median imputation, dropping rows) and compare their impact on the analysis.



The plot shows that mean imputation, median imputation, and deleting missing data result in distributions that are quite comparable for the Electric Range feature. Since it implies that the missing data are either insignificant or not extremely skewed, any of these methods can be used without significantly changing the analysis. Both mean and median imputation are reliable options for handling missing values in this dataset since they maintain the original distribution structure.

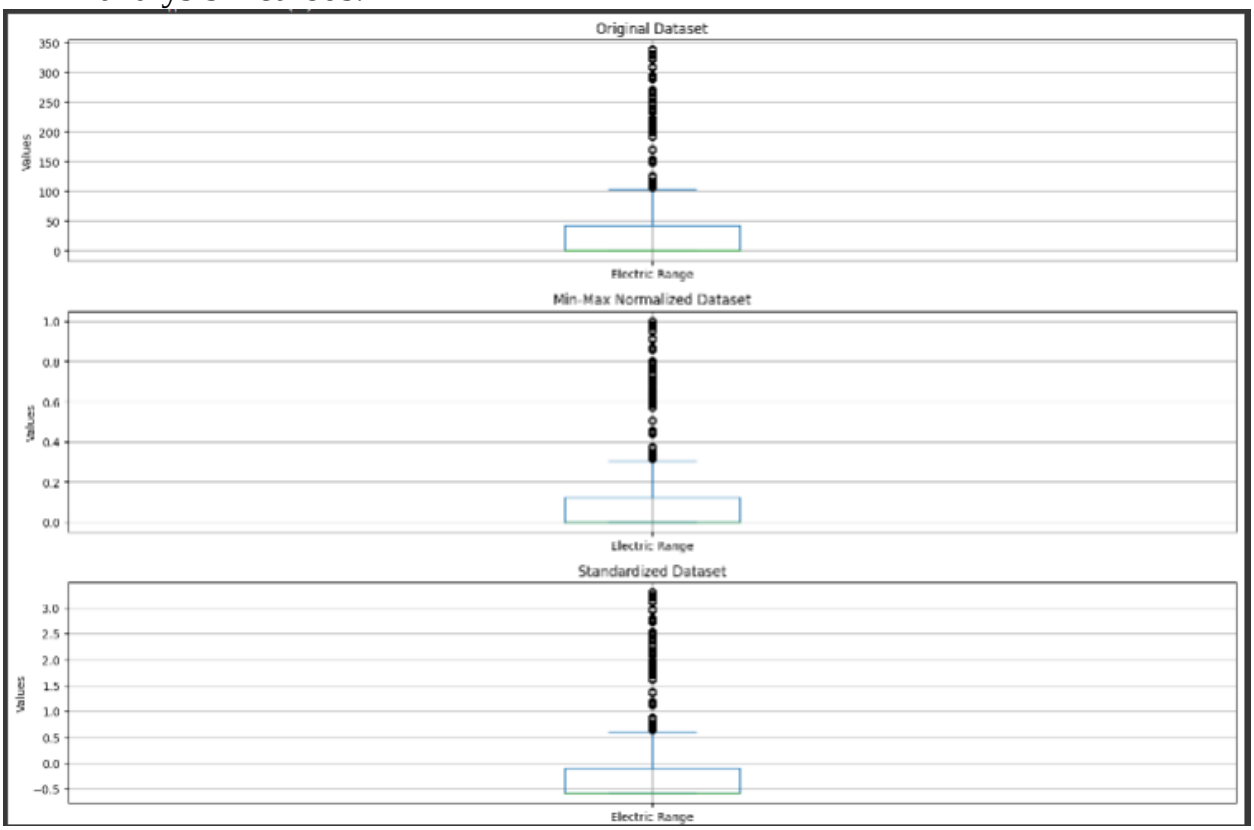
3. **Feature Encoding**: Encode categorical features (e.g., Make, Model) using techniques like one-hot encoding.



The bar chart shows the impact of categorical feature encoding on the dimensions of the dataset. After encoding, the dataset's original 17 columns grew to 209 columns. Techniques like one-hot encoding, which converts each distinct category in categorical columns (such Make and Model) into a unique binary column, are to blame for this notable increase. This encoding significantly expands the number of features in the data, makes it machine-readable, and aids computers in comprehending categorical information. Problems like increased memory utilization and computational complexity could result from this. On the other hand, it might improve the model's capacity to distinguish between several categories in

high-dimensional models. In order to guarantee effective model performance and prevent possible overfitting, this expansion must be balanced.

4. **Normalization:** Normalize numerical features if necessary for chosen analysis methods.



the Original Dataset: The first boxplot displays a highly skewed distribution of "Electric Range" with multiple upper outliers, suggesting that some vehicles have extremely high values and the majority have lower ranges.

Min-Max Normalized Dataset: Scaled from 0 to 1, the second boxplot shows the Min-Max normalized data. It is appropriate for distance-based approaches since it

lessens the impact of outliers while maintaining the distribution structure.

Standardized Dataset: The third boxplot displays standardized data with a standard deviation of one and a center of zero. For algorithms that assume normalcy, this transformation produces a distribution that is more Gaussian-like.

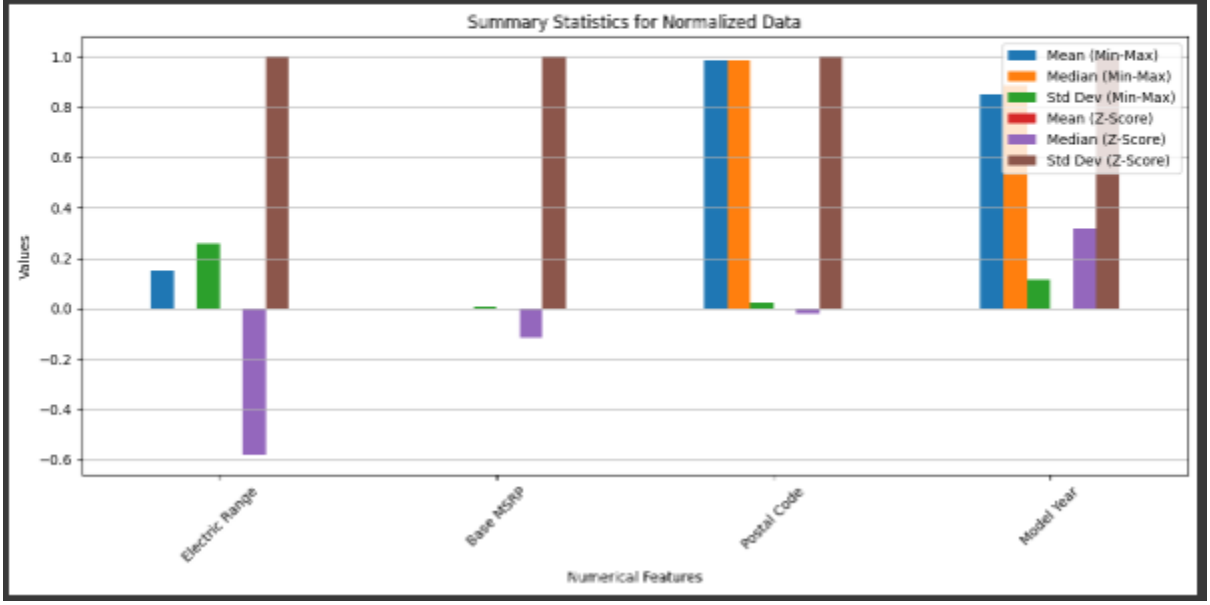
Lastly, the analytical applications of various transformations vary. While min-max normalization aids in data scaling, standardization aids in operations that are sensitive to variance. Selecting the best techniques for enhanced model performance is made simple by analyzing these modifications.

5. **Descriptive Statistics**: Calculate summary statistics (mean, median, standard deviation) for numerical features.

Summary Statistics for Normalized Data:

	Mean (Min-Max)	Median (Min-Max)	Std Dev (Min-Max)	\
Electric Range	0.150155	0.000000	0.258881	
Base MSRP	0.001062	0.000000	0.009058	
Postal Code	0.985704	0.985160	0.024993	
Model Year	0.848025	0.884615	0.114959	

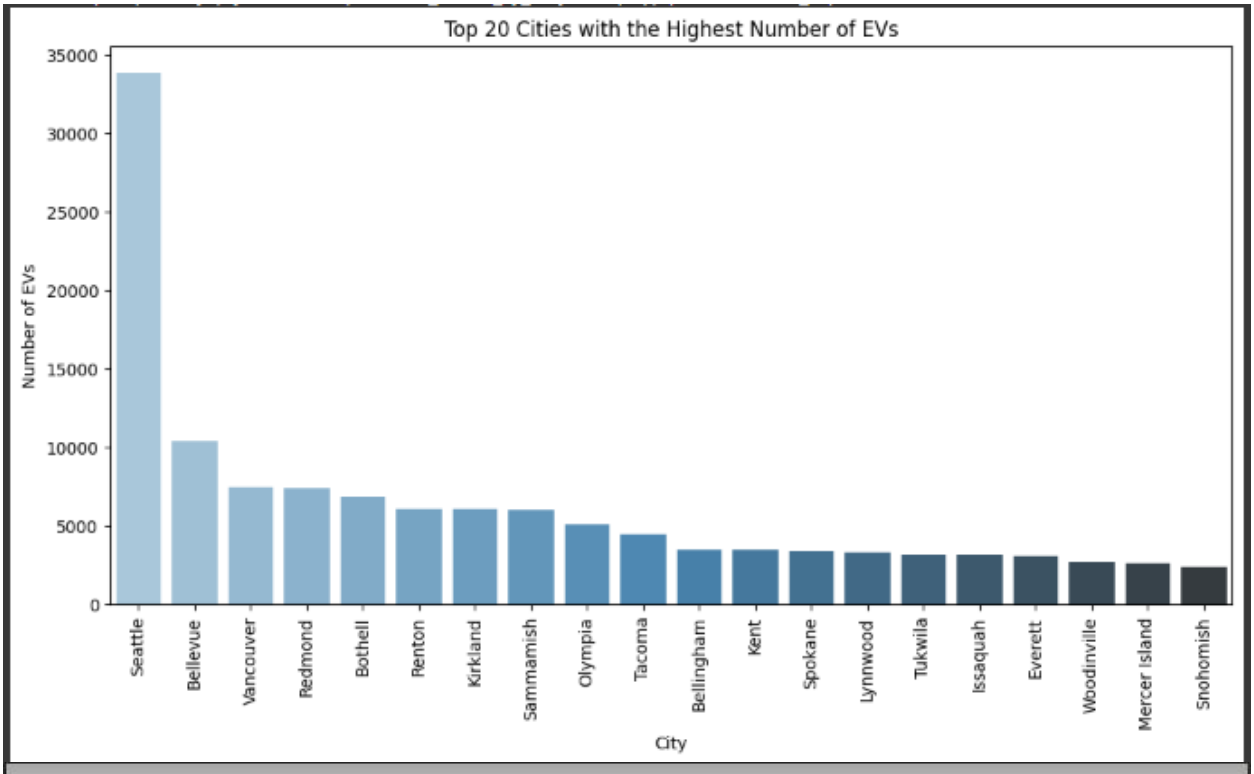
	Mean (Z-Score)	Median (Z-Score)	Std Dev (Z-Score)
Electric Range	2.840007e-18	-0.581816	1.000002
Base MSRP	8.520021e-18	-0.117289	1.000002
Postal Code	1.844643e-15	-0.021759	1.000002
Model Year	-1.736501e-14	0.318288	1.000002

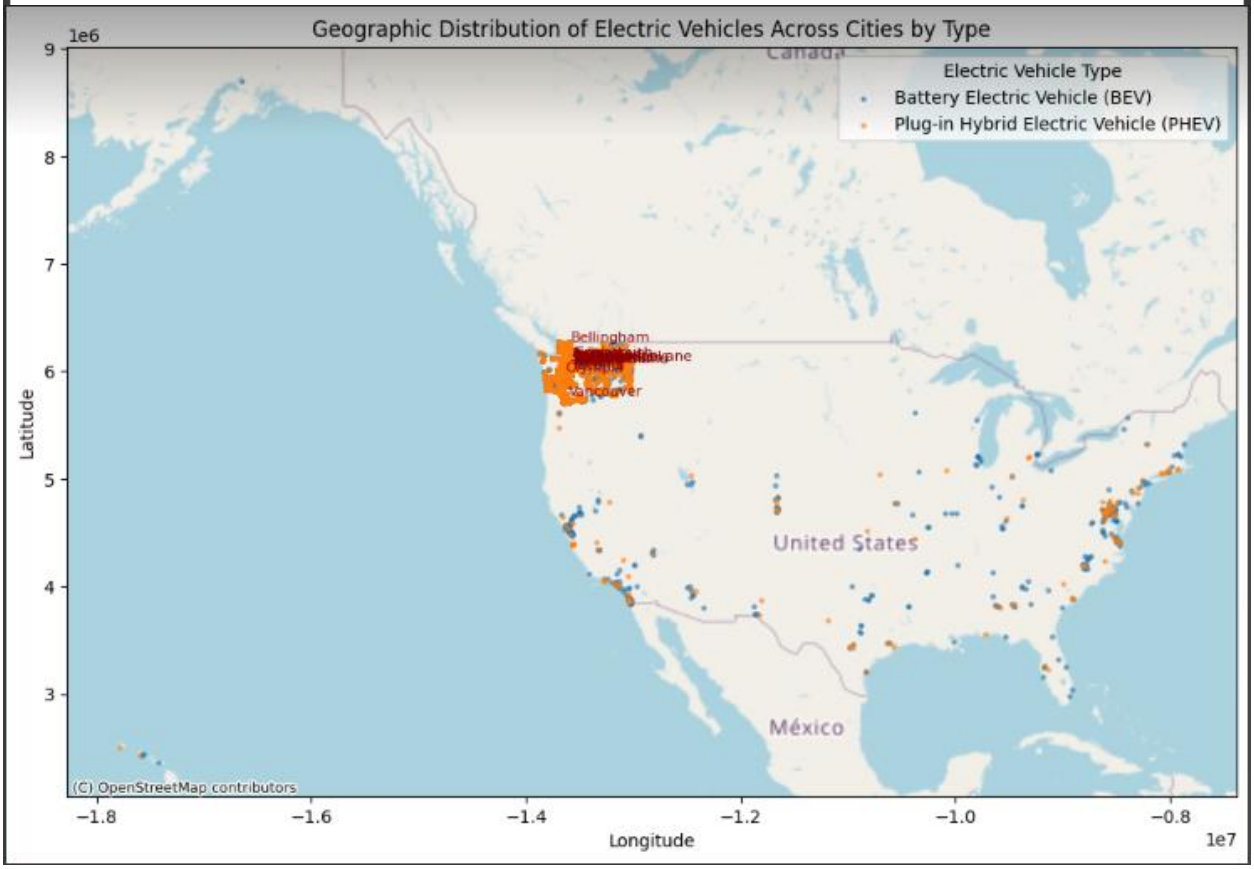


The table and bar plot display summary statistics for the normalized values of the following parameters: "Electric Range," "Base MSRP," "Postal Code," and "Model Year." Among the significant discoveries are:

With the exception of "Postal Code," which has a mean of almost 1, the distribution is skewed, as would be expected.
The median (Min-Max): A median of roughly 0 indicates symmetry, while "Postal Code" represents an oddity.
Min-Max The standard deviation Low values indicate clustering around the mean, whereas "Electric Range" displays greater spread.
Average (Z-Score): Consistent feature scaling is indicated by values near 0.
Z-Score The standard deviation and median: Standard deviations around 1 and medians near 0 validate effective Z-score normalization.
The bar plot draws attention to these variations, which displays clear distribution patterns among.

6. **Spatial Distribution:** Visualize the spatial distribution of EVs across locations (e.g., maps).


Top 20 Cities with the Highest Number of EVs
Number of EVs
35000
30000
25000
20000
15000
10000
5000
0
Seattle
Bellevue
Vancouver
Redmond
Bothell
Renton
Kirkland
Sammamish
Olympia
Tacoma
Bellingham
Kent
Spokane
Lynnwood
Tukwila
Issaquah
Everett
Woodinville
Mercer Island
Snohomish
City


Geographic Distribution of Electric Vehicles Across Cities by Type
1e6
Latitude
Longitude
1e7
Electric Vehicle Type
Battery Electric Vehicle (BEV)
Plug-in Hybrid Electric Vehicle (PHEV)
Bellingham
Vancouver
United States
México
(C) OpenStreetMap contributors

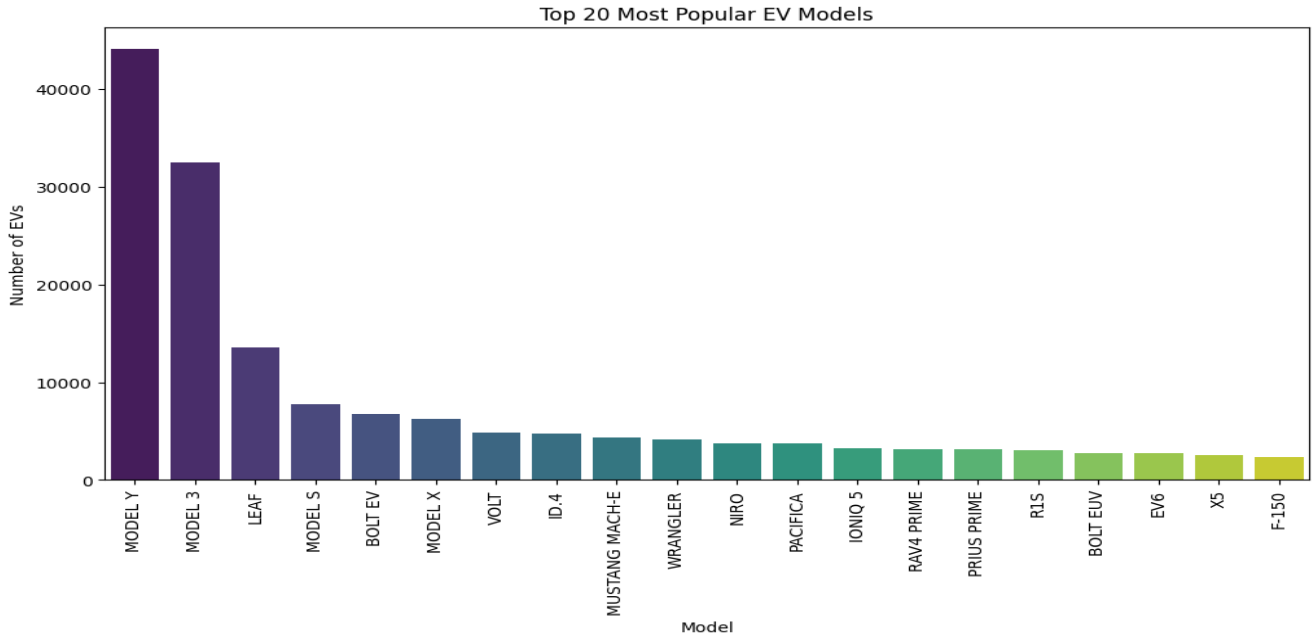
The 20 Cities with the Most Electric Vehicles:

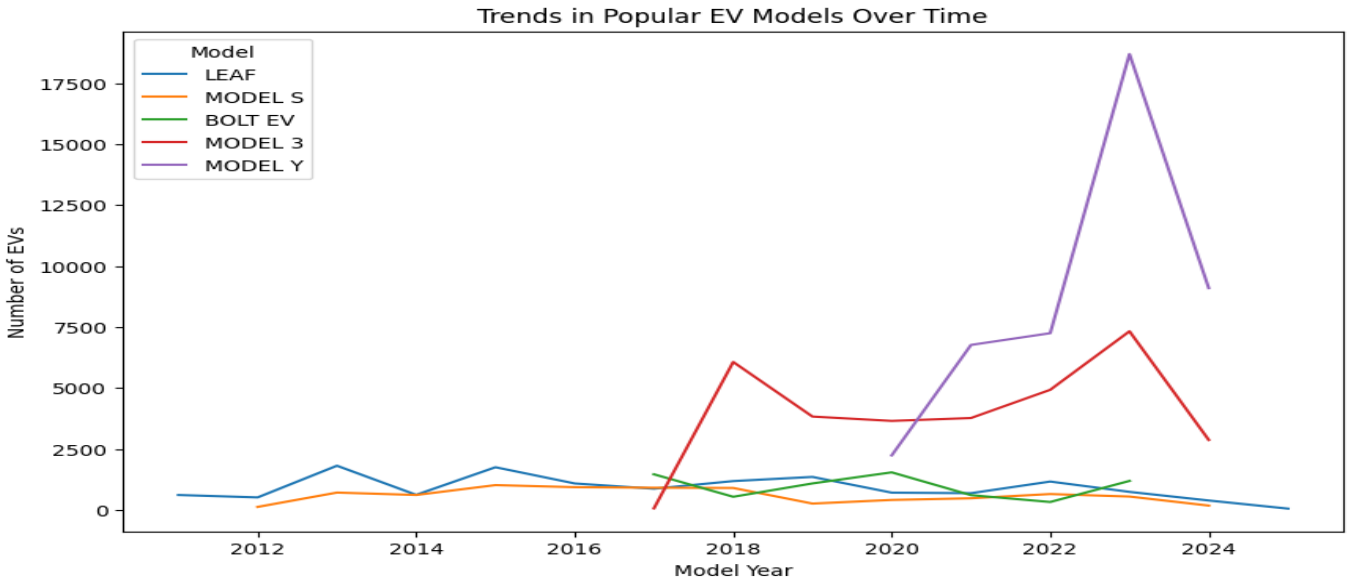
Seattle leads a bar chart of the top 20 cities for EV adoption, suggesting a robust local uptake that is probably supported by infrastructure support and incentives. Richmond, Vancouver, and Bellevue also have noteworthy counts, albeit they are far lower than Seattle's.

Distribution by EV Type and Geography:

Battery electric vehicles (BEVs) and plug-in hybrid electric vehicles (PHEVs) are the two types of EVs that are most prevalent in the Northwest, particularly in Washington State, according to a map representation. On the East Coast, there are other urban clusters where usage is less common. These images allow for a detailed analysis of the concentration and distribution of EV use in the US.

7. **Model Popularity**: Analyze the popularity of different EV models (categorical data) and identify any trends.



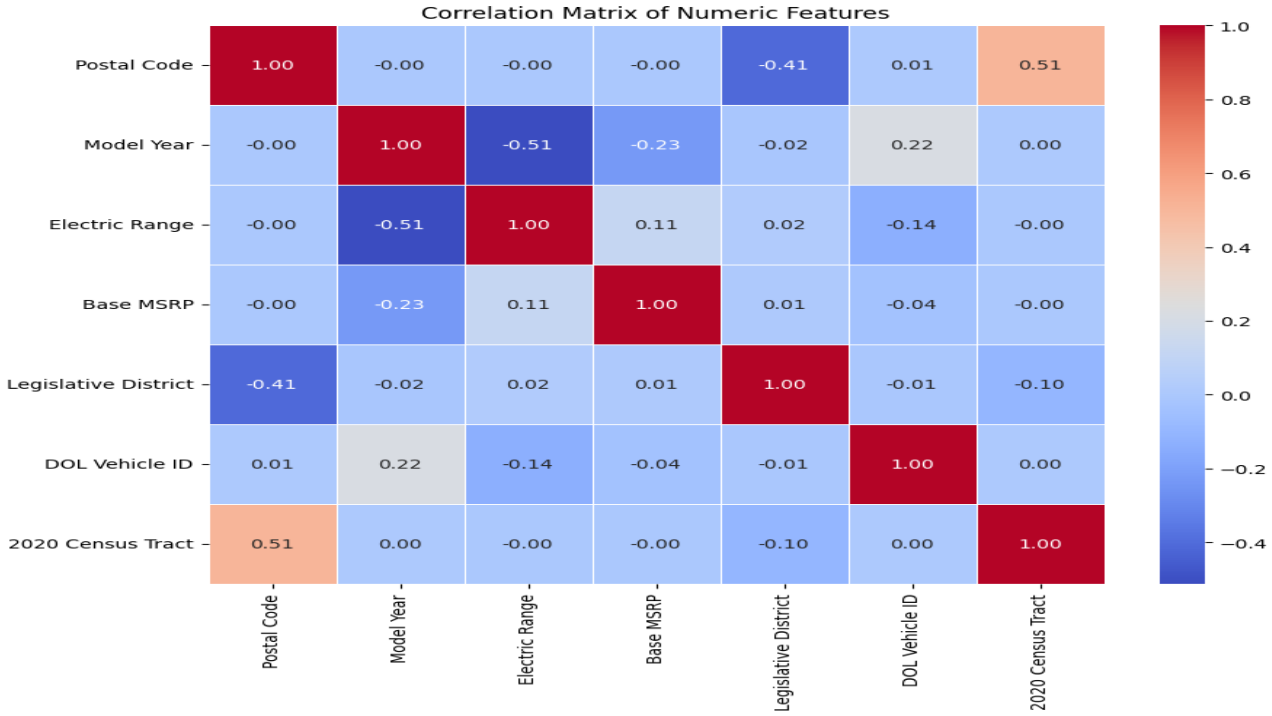


Top EV Models: The bar plot displaying the top 20 EV models reveals that the Tesla Model Y and Model 3 are the most popular electric vehicles in Washington, with significant adoption compared to other models like the Nissan Leaf and Chevrolet Bolt EV. This suggests a high market share for Tesla in the state's EV population, possibly due to its brand recognition, extensive range, and advanced features.

Trends Over Time: The line plot tracking trends in EV model popularity by model year shows that Tesla's Model Y and Model 3 have seen a sharp increase in popularity over recent years, peaking in the early 2020s. This trend underscores

Tesla's recent dominance in the EV market, particularly since its production scale-up around 2020. Other models, such as the Nissan Leaf, show steadier but less dramatic growth over time, indicating stable yet limited market penetration.

8. Investigate the relationship between every pair of numeric features. Are there any correlations? Explain the results



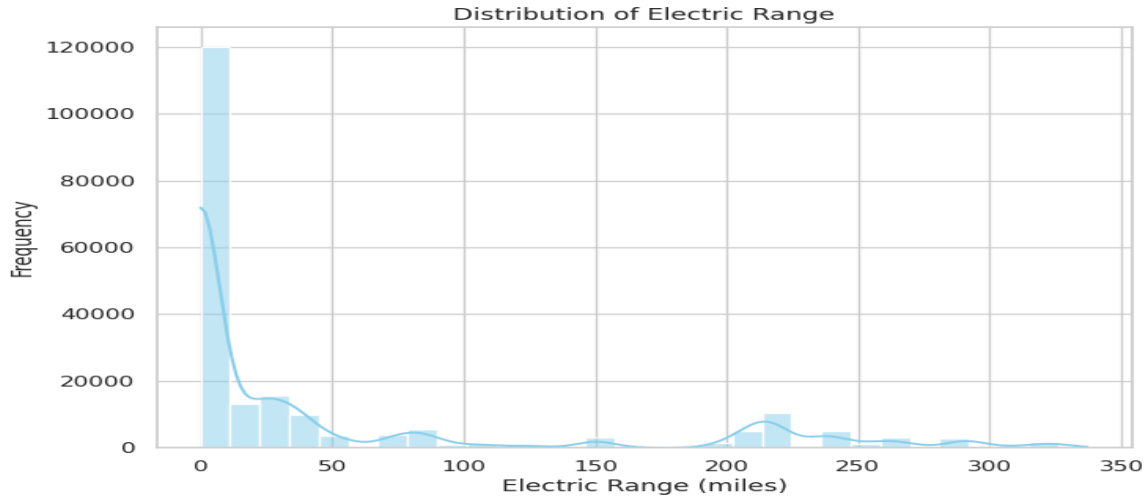
Electric Range and Model Year: There is a moderate negative correlation (-0.51) between model year and electric range, which could suggest that newer models tend to have higher electric ranges, possibly due to improvements in battery technology.

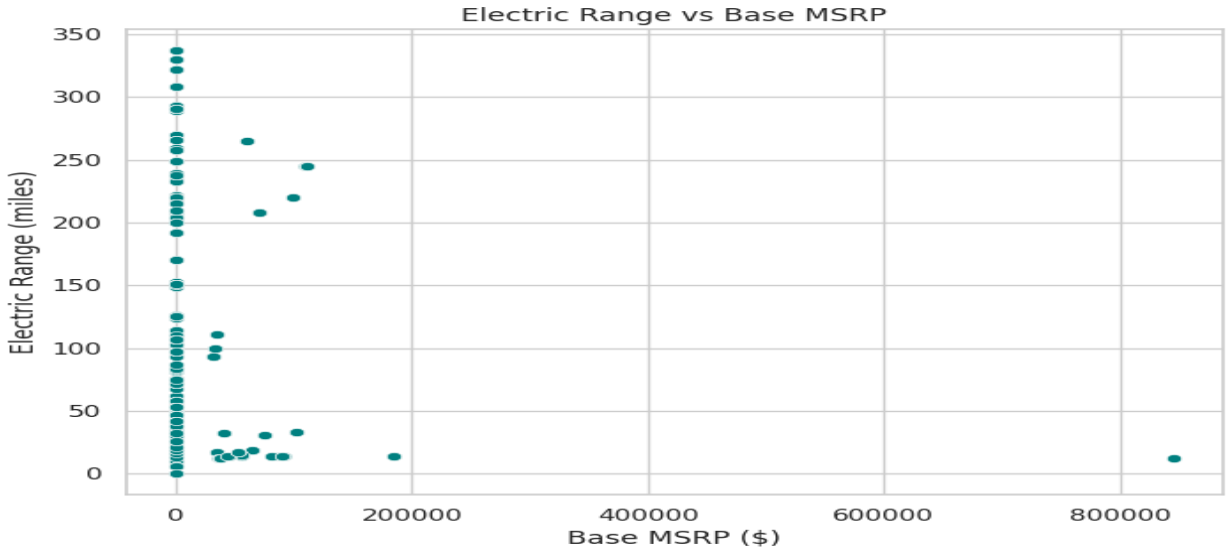
Base MSRP and Electric Range: A small positive correlation (0.11) exists between base MSRP and electric range, indicating that higher-priced vehicles may offer longer electric ranges, aligning with the trend of premium EVs offering greater capabilities.

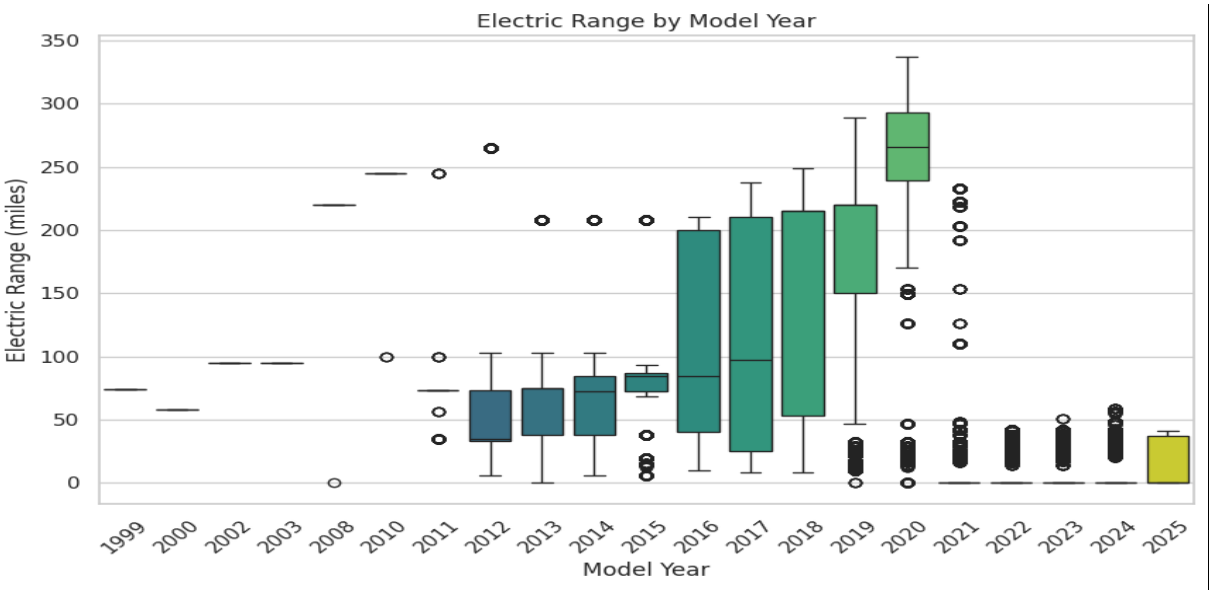
Postal Code and 2020 Census Tract: A moderate positive correlation (0.51) between postal code and census tract suggests that the dataset captures EV registrations in specific areas that may be closely associated with these geographic identifiers.

Model Year and Base MSRP: The slight negative correlation (-0.23) might imply that older EV models tend to have lower original prices, while newer models, especially as EV technology advances, are priced higher.

9. **Data Exploration Visualizations**: Create various visualizations (e.g., histograms, scatter plots, boxplots) to explore the relationships between features.







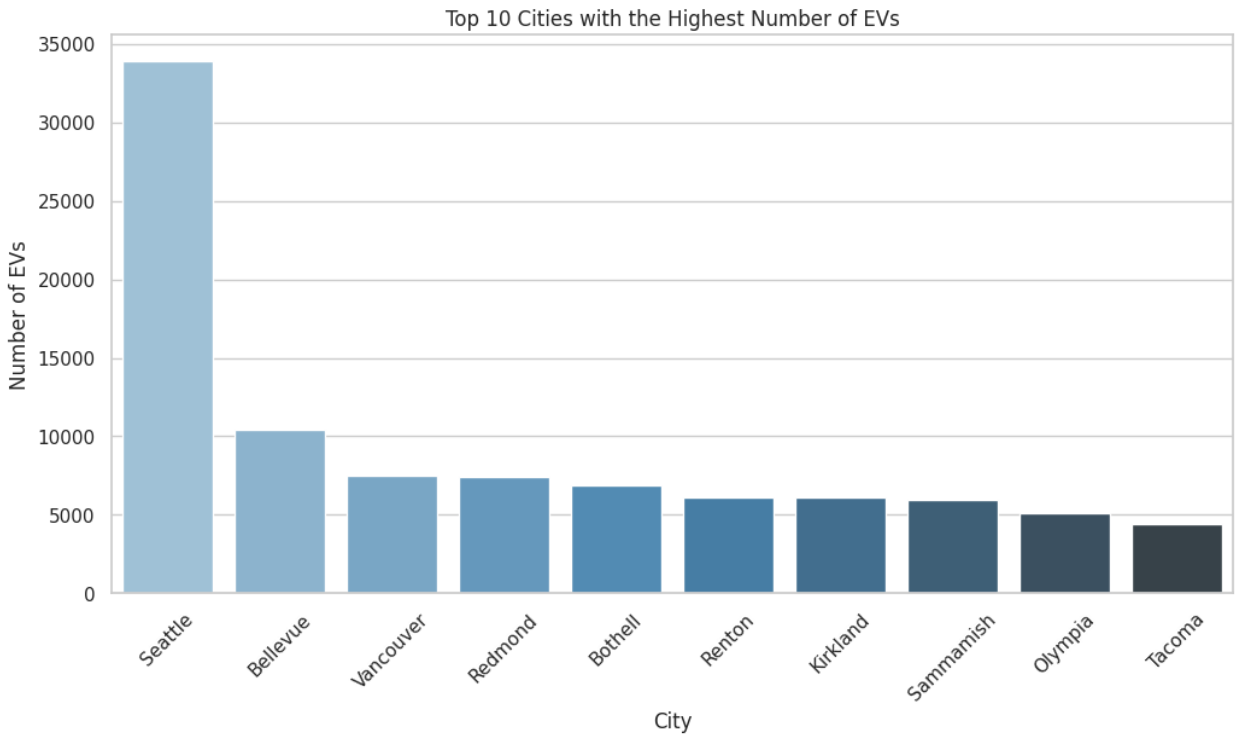
Distribution of Electric Range: The histogram with KDE (Kernel Density Estimation) overlay shows the distribution of electric ranges across the dataset. A large portion of EVs has a range under 50 miles, with fewer models offering longer ranges. This suggests that most EVs may be suitable for short-distance commuting rather than long-distance travel.

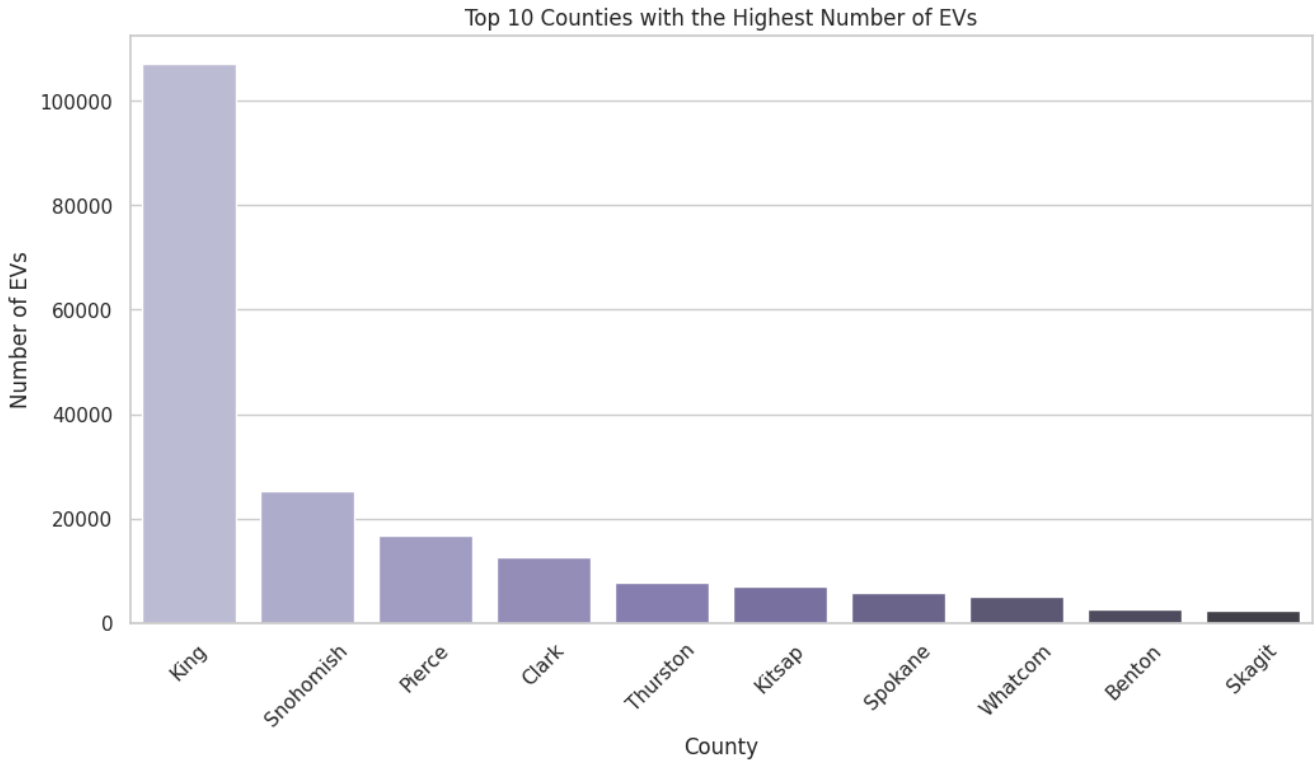
Electric Range vs Base MSRP: The scatter plot displays the relationship between the electric range and base MSRP of EVs. The concentration of points along the y-axis suggests that many EVs with lower prices tend to offer shorter ranges. This

could indicate that higher range is often a premium feature associated with more expensive models.

The Electric Range by Model Year boxplot shows a general increase in electric vehicle range from around 2015 to 2019, reflecting advancements in battery technology and design. Between 2017 and 2019, the range distribution broadens, indicating variability to meet different consumer needs or market segments. However, from 2021 to 2025, there is a noticeable decline in both median and maximum ranges, possibly due to more short-range, budget-friendly models or limited data on recent models. Outliers, particularly from 2015 to 2020, reveal some premium models achieving ranges over 250 miles.

10. **Comparative Visualization**: Compare the distribution of EVs across different locations (cities, counties) using bar charts or stacked bar charts.





Top 10 Cities with the Highest Number of EVs:

Seattle leads significantly, with a much higher count of registered EVs compared to other cities, followed by Bellevue, Vancouver, and Redmond.

Insights: The high EV numbers in Seattle suggest a strong adoption of electric vehicles in larger urban areas, likely due to higher environmental awareness, better charging infrastructure, and supportive local policies.

Top 10 Counties with the Highest Number of EVs:

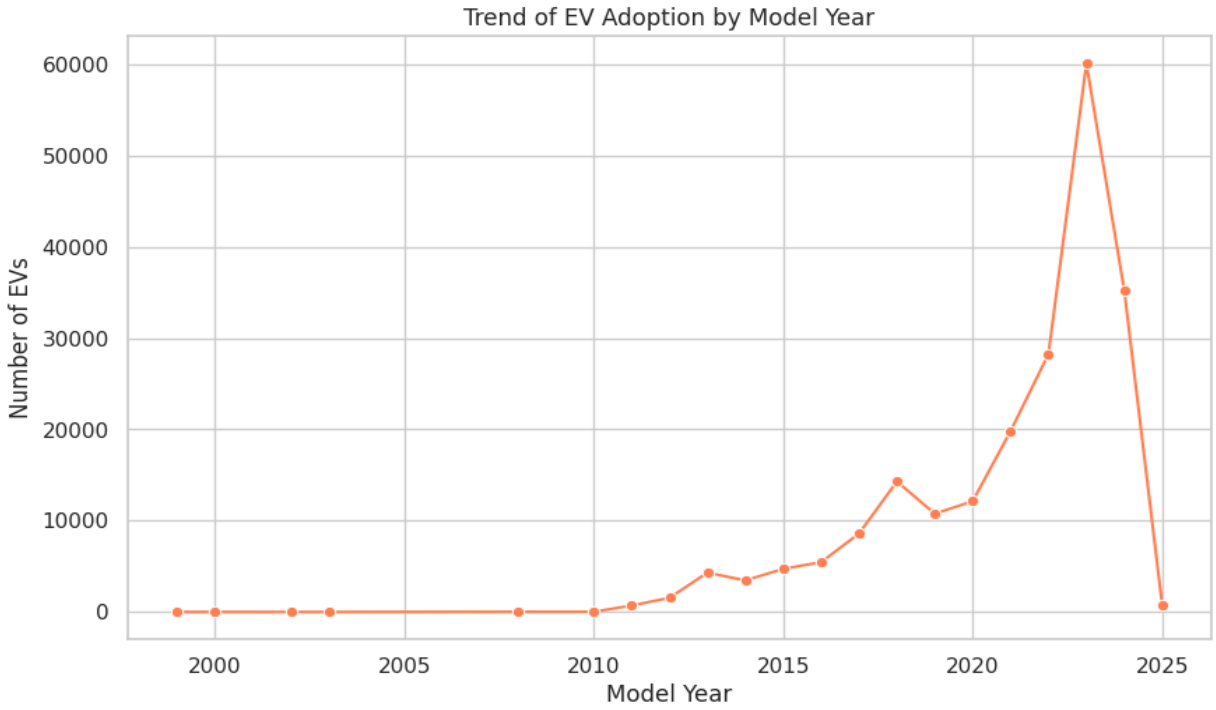
King County, where Seattle is located, has by far the highest number of EV registrations, dwarfing the numbers in other counties like Snohomish and Pierce.

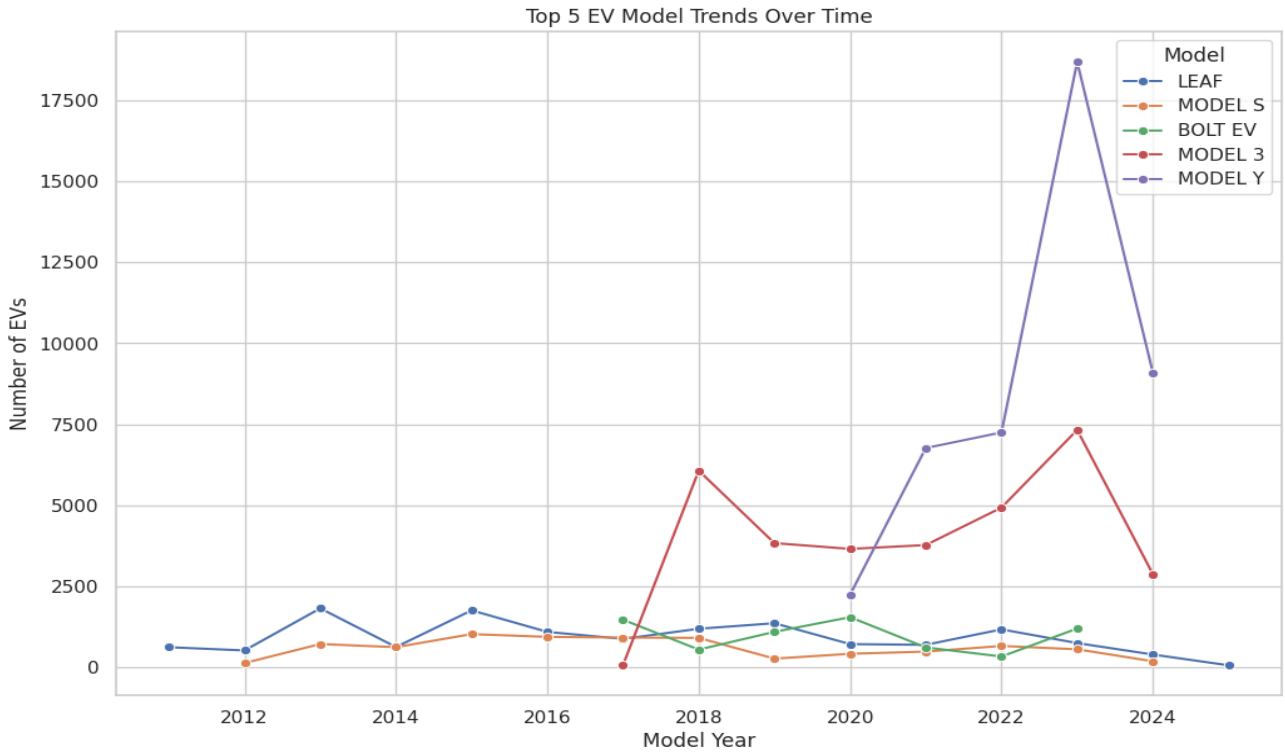
Insights: King County's dominance in EV registrations highlights a regional concentration of EV adoption, which may be influenced by socio-economic

factors, population density, and infrastructure availability. Other counties, while smaller in numbers, still show notable adoption, indicating EVs' appeal beyond just the most urbanized areas.

The concentration of EV registrations in Seattle and King County indicates that urban, affluent areas lead in EV adoption, likely due to better charging infrastructure and incentives. This trend suggests that expanding EV infrastructure should prioritize high-growth areas while also addressing underserved regions where adoption is increasing.

11. **Temporal Analysis**: If the dataset includes data across multiple time points, analyze the temporal trends in EV adoption rates and model popularity.





EV Adoption Trend by Model Year:

The line plot for the number of EVs by model year provides a visual representation of the growth in EV adoption over time. The code sorts the Model Year values and plots the frequency for each year, which clearly shows the increasing trend of EV adoption, with a notable spike in recent years.

Top EV Model Trends Over Time:

This section identifies the top 5 EV models by their total count over all years, then filters the data to display only those models. A line plot with each model's count per year shows how the popularity of each model has changed over time. It provides insight into how certain models have surged in popularity compared to others.