Faculty of Engineering & Technology
Electrical & Computer Engineering Department

Machine Learning and Data Science

ENCS5341

Assignment #1

"Exploring Electric Vehicle Trends in Washington State"

Prepared by: Noor Hamayel — **ID Number**: 1202853
Rivan Jardat — 1200081

Instructor: Dr.Yazan Abu Farha

Section: 1

Oct-2024

Abstract

Using Google Colab for data processing and analysis, this report analyzes Washington State's electric vehicle (EV) registration data to identify trends in EV adoption. We analyze trends in plug-in hybrid electric cars (PHEVs) and battery electric vehicles (BEVs) using data from Data.gov. Data cleaning, feature engineering, and visual inspection of EV distribution across regions, popular models, and adoption growth over time are all included in the analysis. An overview of EV trends and the variables influencing their ascent in Washington State is given via key insights.

Procedure and discussion:

Data Cleaning and Feature Engineering:

1. Document Missing Values:

After analyzing the data set, missing values were identified and documented to understand the quality of the data and the extent of imputation required.The data 210165 entries , with a total of 456 missing values across All Feature ,Below is a summary of each feature with missing values:

- County :4
- City:4
- Postal Code:4
- Electric Range:5
- Base MSRP5Legislative Legislative District: 445
- Vehicle Location:10
- Electric Utility :4
- 2020 Census Tract: 4

2. Missing Value Strategies:

Since the dataset contains missing values, two imputation strategies were applied: dropping rows with missing values and mean imputation. To analyze the impact of each, find the method that best preserves data quality and reduces bias. The mean imputation replaces missing values in a numeric column with the mean of that column. This method helps maintain the size of the data set; however, it can hide variations in the data and introduce bias, and it cannot be applied to non-numeric columns. The dropping method removes any rows that contain missing values. This maintains the integrity of the remaining data because it removes potential inaccuracies caused by assumed values. However, it can result in a significant reduction in the size of the data set, which can lead to the loss of valuable information.

3. Feature Encoding:

Since the dataset contains categorical features, it must be encoded so that machine learning algorithms can interpret it as numerical data. We applied one-hot encoding to the EV type and state. Each feature was transformed into a separate binary column (0 or 1), indicating the presence of each category in the original feature.
Example: The feature State includes categories like “WA,” “OR,” and “MI.” With one-hot encoding, a new column is created for each make, where a row value is 1 if that category is present and 0 otherwise. but it ensures that the model can interpret categorical data effectively without assuming any ordinal relationship between categories.

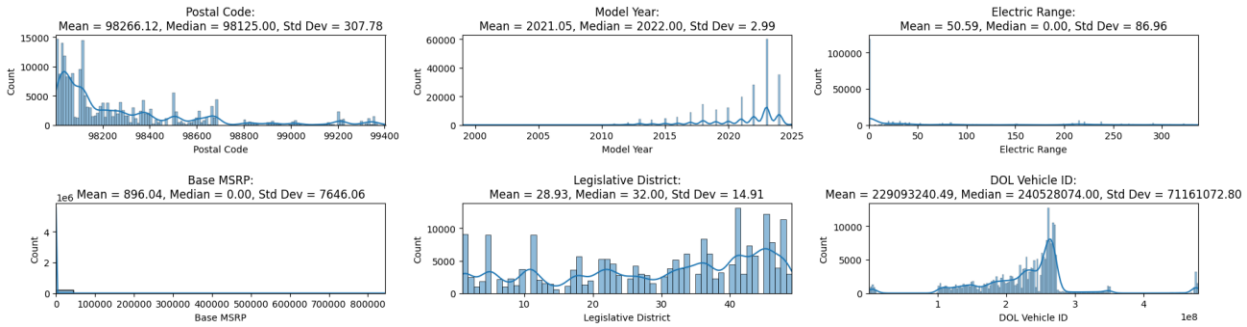
4. Normalization:

Normalization is used on numerical features to prepare them for analysis, ensuring that all features have comparable scales, especially for machine learning algorithms that are sensitive to feature size. The Min-Max method was used to transform each numerical feature to a scale between 0 and 1. This method was chosen because it preserves the relationships between values while bringing all features into a uniform range.

Exploratory Data Analysis

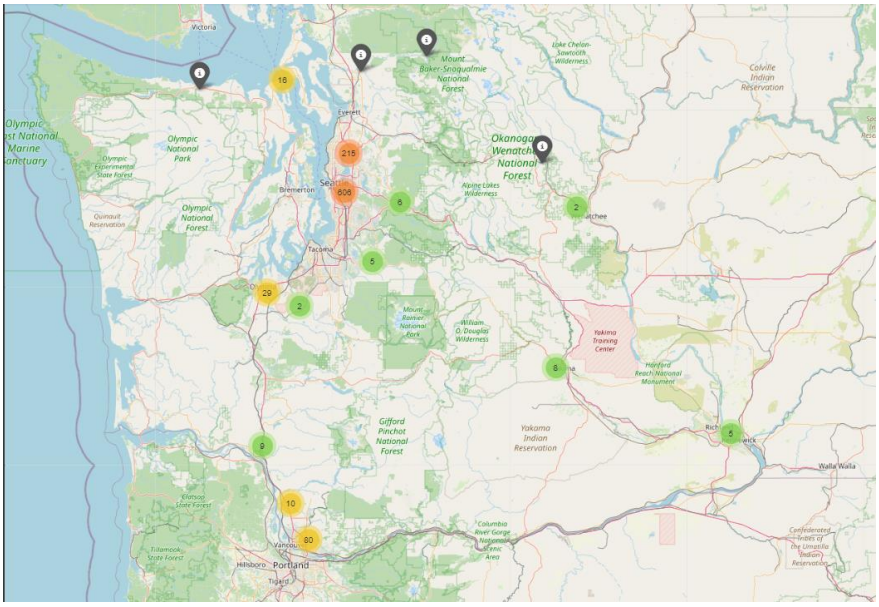
5. Descriptive Statistics

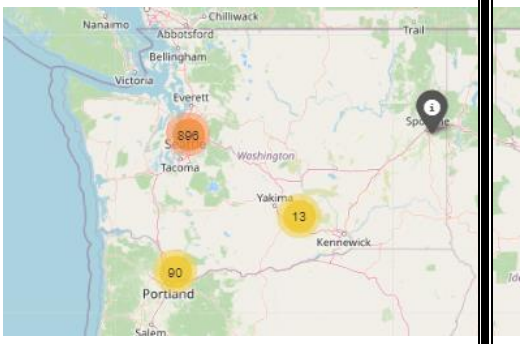
Central patterns across several variables are shown by the **mean** and **median** values of the dataset's numerical features. The **Model Year,** for example, shows a recent emphasis on EV registrations with a mean of 2021 and a median of 2022. With a median of 0 and a mean of roughly 50, the electric range indicates that many cars have little to no range when using only electricity. This is probably because plug-in hybrids are now available. Likewise, low **mean** and **median** values in the **Base MSRP** may suggest that many items lack pricing data. The dataset's variability is further demonstrated by the standard deviation values: **Model Year** has a reduced standard deviation, clustering around recent years (2021–2022), while **Electric Range** and **Base MSRP** exhibit considerable dispersion, demonstrating a variety of EV types and pricing.



The histograms with KDE curves clearly show the distribution of each numerical feature. Newer EV models are heavily concentrated in recent years, as indicated by the **Model Year** histogram's peak. Due to a combination of plug-in hybrids and less expensive models, the **Base MSRP** and **Electric Range** are significantly skewed to the right, with many cars having low price points or little electric range. With a preponderance of newer models and a variety of range and pricing characteristics, this image demonstrates the dataset's heterogeneity, especially in the age, range, and cost of EVs.

6. Spatial Distribution

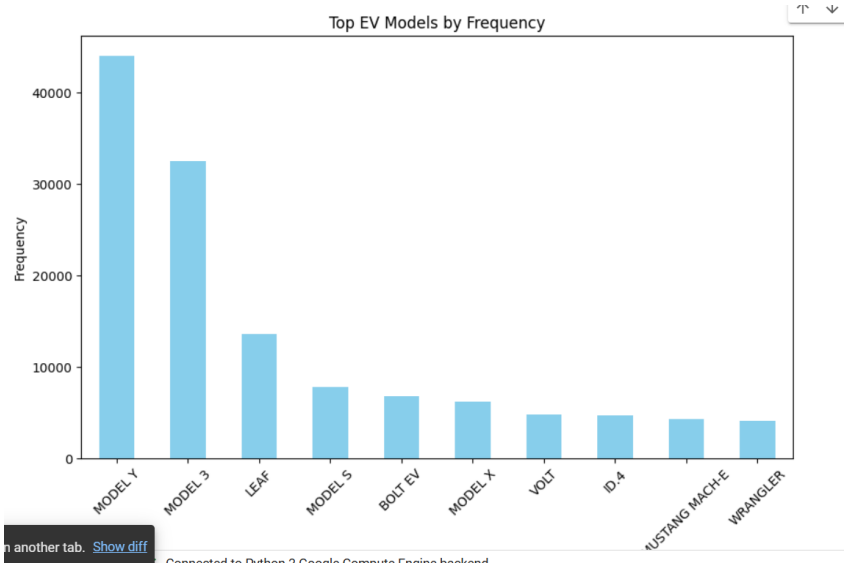




With clusters denoting regions of significant EV density, particularly in nearby areas of major cities like Seattle and Tacoma, the map displays the spatial distribution of about 1,000 EVs throughout Washington State. Utilizing geographical coordinates and clustering, it illustrates how infrastructure and population density cause EV adoption to be significantly higher in populated areas and lower in rural ones.

7. Model Popularity:

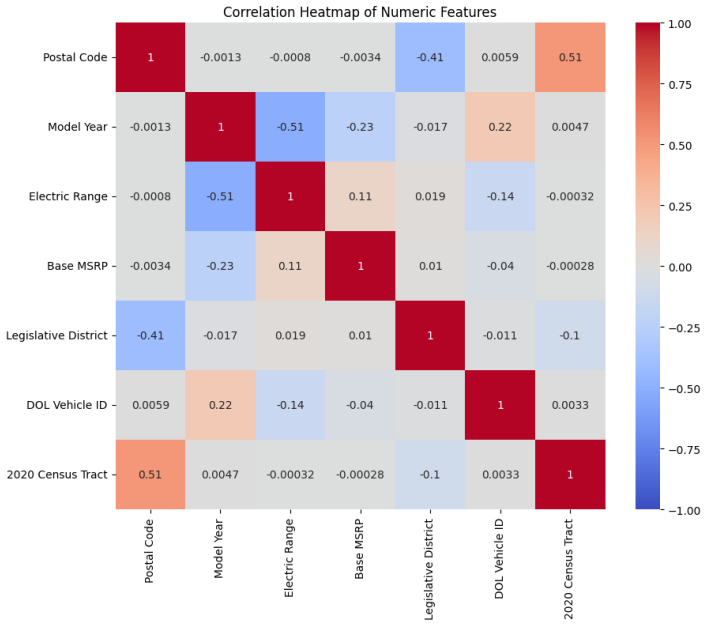
We analyzed the popularity of different EV models (categorical data) by finding the frequencies of each model in the dataset. This analysis is useful in finding the most frequent and popular models and examining users' attitudes and inclinations toward electric vehicles.



This figure shows the most popular models in the dataset, with Model Y having the highest frequency of 44,038. This chart also highlights how user preferences vary between models, revealing trends in EV popularity.

8. Analysis of Correlation Between Numerical Features:

To study the relationships between numerical features in the dataset, we created a heat map of correlation. Through this map, we will be able to find the strength of the correlation between each pair of features.



If the value is close to zero, this means that there is no correlation, such as Postal Code with Base MSRP. However, if the value is negative, as is the case between model year and electric range (-0.51), this means there is a neutral negative correlation indicating that newer models may have a slightly lower electric range.Also, if the value is positive as between Postal Code and 2020 Census Tract (0.51), this means that

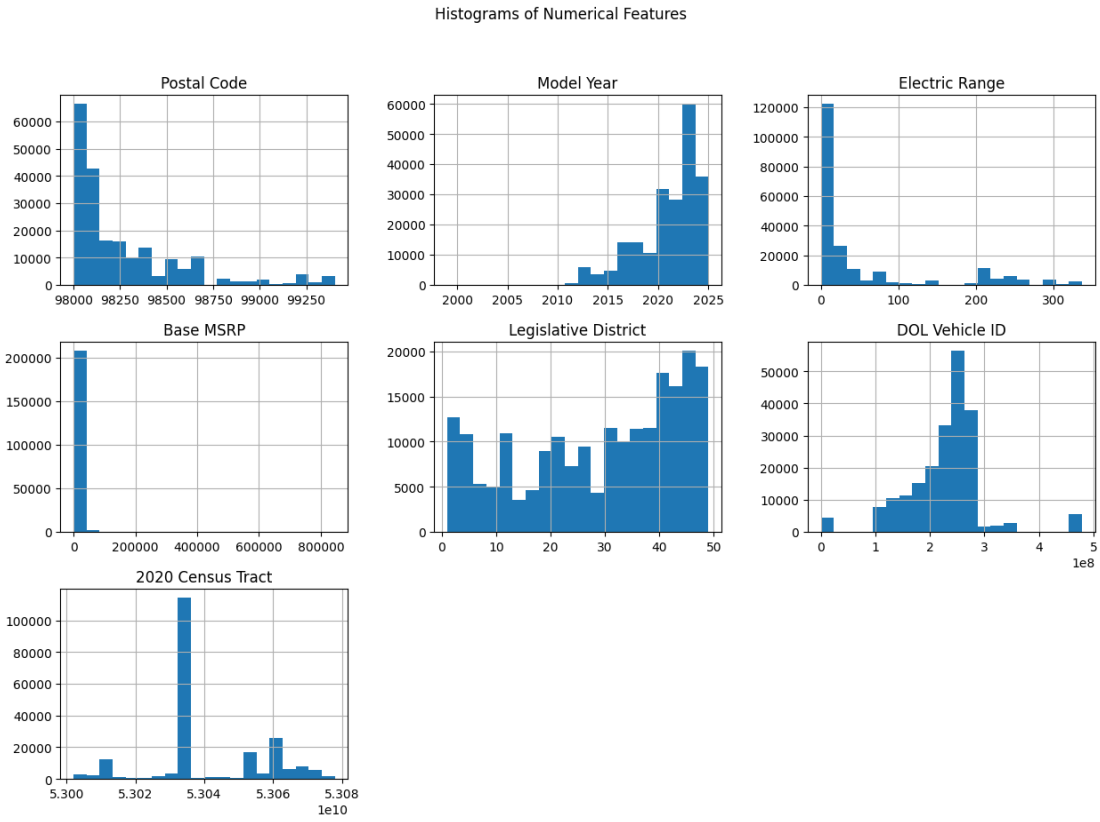
there is a moderate positive correlation that reflects a geographic distribution pattern associated with postal code areas.

Visualization

9. Data Exploration Visualizations

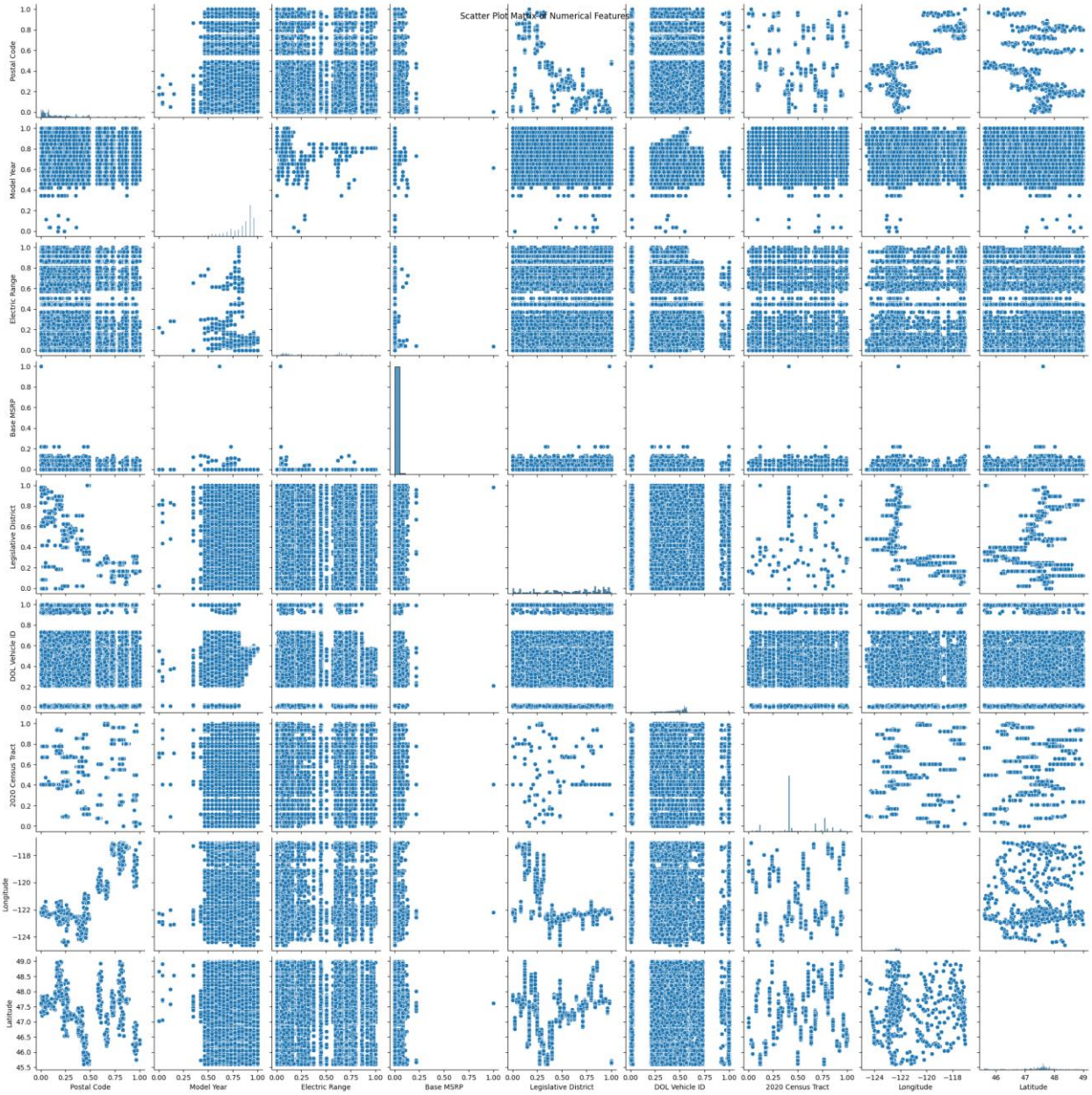
Below are visualizations exploring the dataset's distributions, relationships, and outliers.

- **Histograms for Numerical Features**



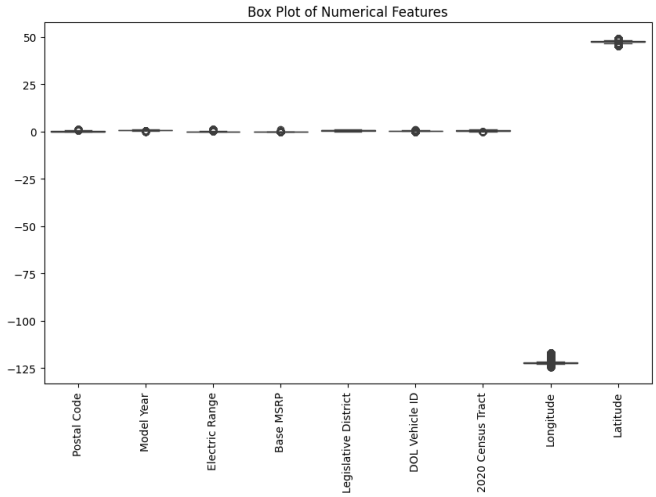
EV registrations are concentrated in particular regions, as indicated by the **Postal Code** histogram, indicating greater acceptance in those areas. Growing popularity over time is indicated by the **Model Year** histogram, which shows a consistent rise in EV registrations. Both **base MSRP** and **electric range** show a great deal of variation, with a few high-value **outliers** clearly visible and a large number of cars with low price and range. The dataset's wide range of attributes is further supported by the diversity of other factors, such as **Legislative District** and DOL **Vehicle ID**, which vary among entries.

- **Scatter Plot Matrix**



Relationships between numerical features are shown by the scatter plot matrix. The positive trend between Model Year and Electric Range indicates that newer models typically have longer ranges, which is consistent with EV technology developments. However, there is no discernible relationship between **Base MSRP** and **Model Year**, suggesting that EV prices vary from year to year. Latitude and longitude also show clustering, suggesting that EV registrations are concentrated in certain regions. Strong correlations are absent from other feature pairings, indicating a broad range of traits in the EV dataset.

- **Box Plot to Identify Outliers**

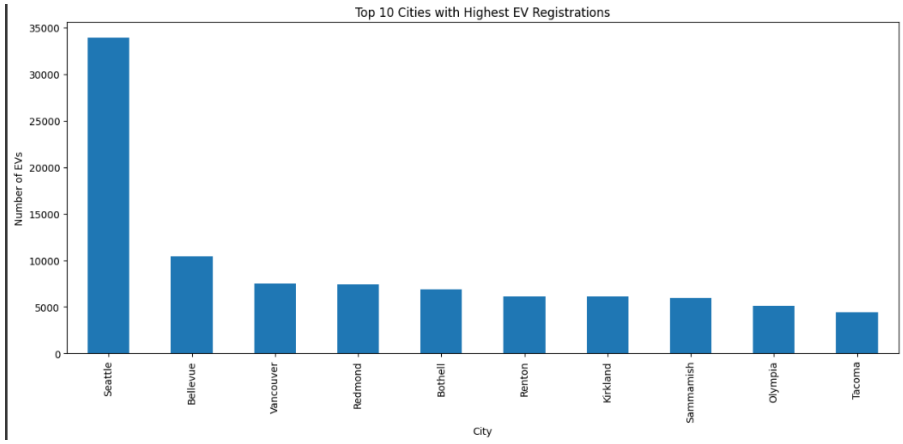


The box plot displays the distribution of values among the dataset's numerical features; the majority of the attributes have a small range, suggesting little variability. Significant **outliers** can be seen in both latitude and longitude, most likely as a result of inconsistent data entry or geographic concentration. Particularly in spatial distributions, these outliers point to locations that can benefit from additional study or data cleansing. The features that could affect EV distribution insights are highlighted in this summary.

These visualizations help us understand EV adoption patterns and feature variations in the dataset, providing insights into trends and outlier characteristics.

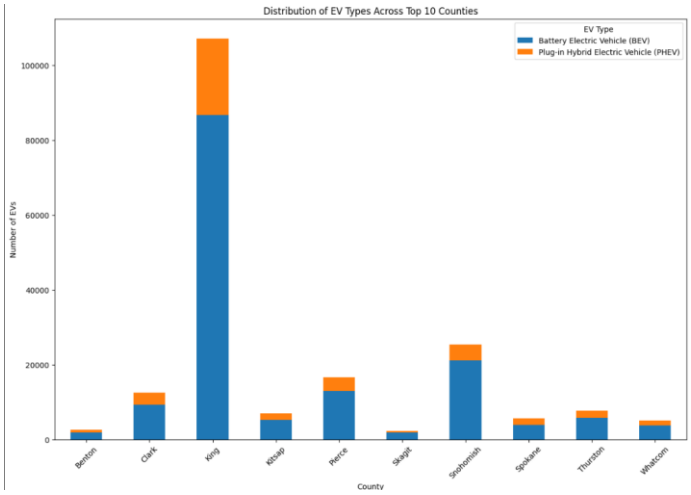
10. Comparative Visualization

- **Bar Chart for Top 10 Cities by Number of EV Registrations**



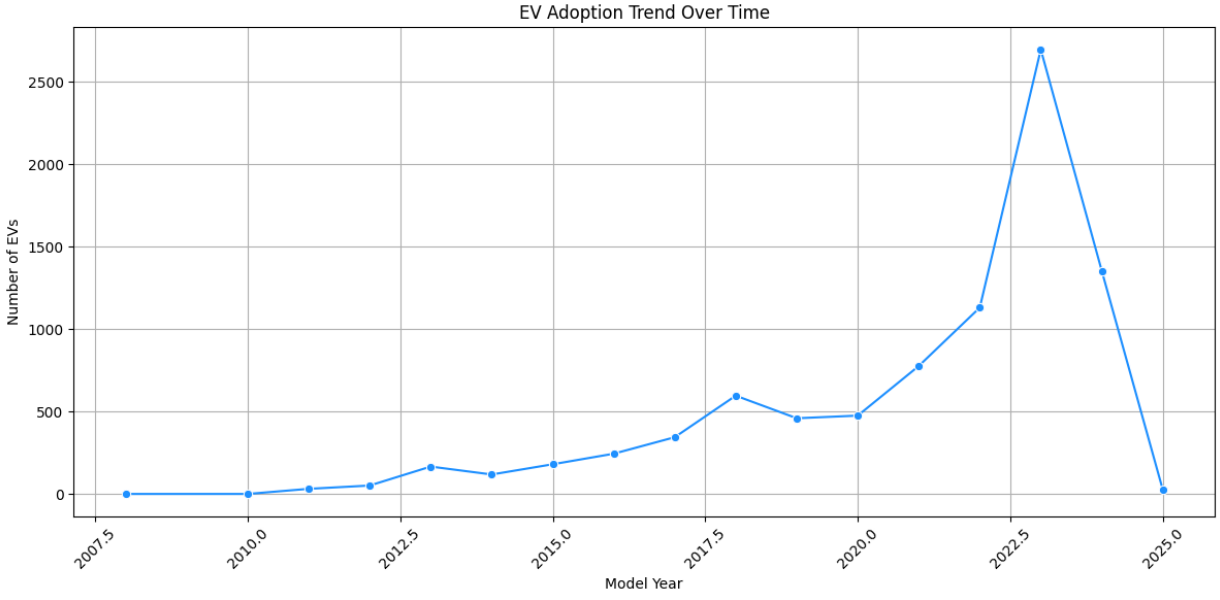
According to the visuals, EV adoption is greatest in urban regions, with Seattle far ahead of Bellevue, Vancouver, and Redmond. Concentrated EV use in Washington's major areas is reflected in these cities' greater populations and superior infrastructure.

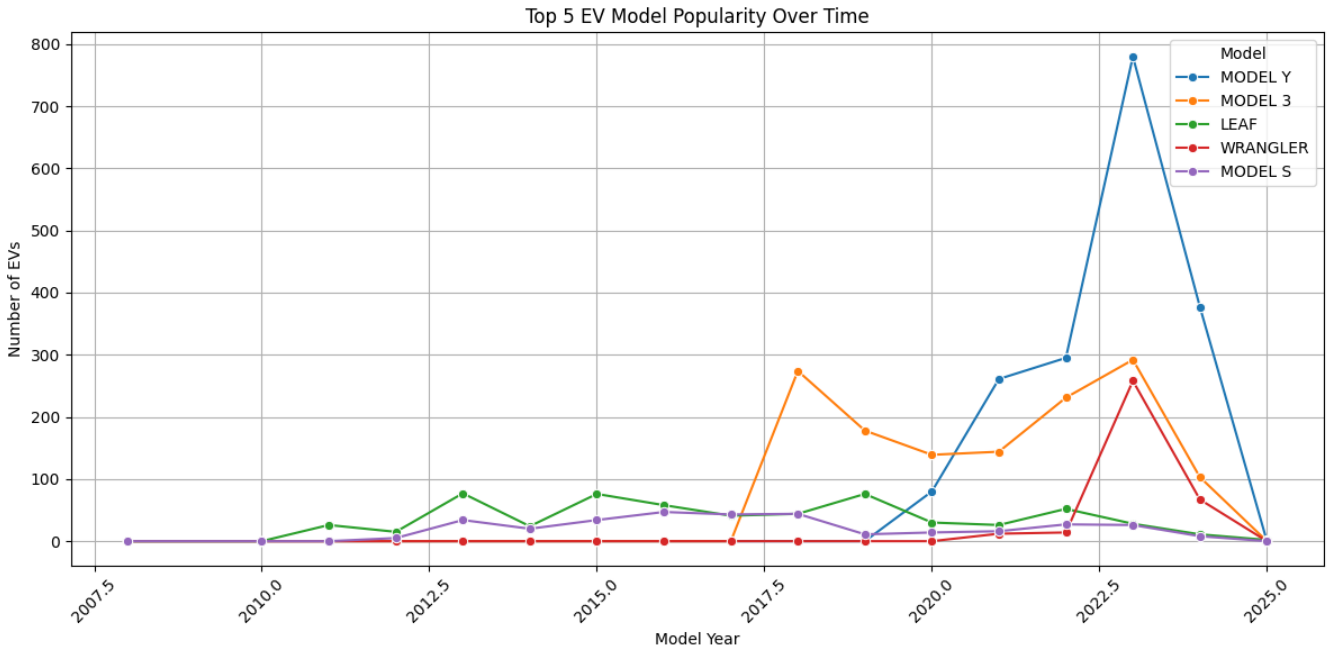
- **Stacked Barfor Top 10 Counties by Number of EV Registrations**



King County leads in EV registrations, primarily BEVs, according to the stacked bar graphic. Snohomish, Pierce, and Clark are next in line. This indicates that developed counties with supported infrastructure are more likely to adopt EVs.

11. Temporal Analysis





Based on their **model years**, this line graph shows the popularity patterns of the top 5 EV cars throughout time. Starting about 2021, Tesla's Model Y and Model 3 saw tremendous increase, demonstrating their robust market presence and rising adoption rates in recent years. While the Tesla Model S continues to have a steady, albeit slower, adoption trend over time, other vehicles, such as the Nissan Leaf and Jeep Wrangler, exhibit modest and irregular popularity. The graph illustrates the influence of new EV releases as well as the changing tastes of consumers. This research contextualizes changing EV choices in the dataset by capturing historical fluctuations in model popularity, with a noticeable spike in recent years.

Conclusion:

In conclusion, this assignment provided us with valuable and practical experience in working with data, and through data cleaning, feature engineering, and exploratory data analysis, we gained insights into the structure and trends of EV adoption in different regions.