"بسم الله الرحمن الرحيم"

Faculty of Engineering and Technology.

Department of Electrical and Computer Engineering.

MACHINE LEARNING AND DATA SCIENCE - ENCS5341.

Assignment #1.

Students name (Prepared by): Razan Abdelrahman –1200531.

Hidaya Mustafa - 1201910.

Instructor: Dr. Ismail Khater.

Section: 2.

Date: 31th | October.

1. Introduction.

This report analyzes electric vehicle (EV) registration data from Washington State, focusing on battery electric vehicles (BEVs) and plug-in hybrid electric vehicles (PHEVs). The aim is to apply data preprocessing and exploratory data analysis (EDA) techniques to uncover insights from the dataset titled "Electric Vehicle Population Data."

Key preprocessing steps included handling missing values, encoding categorical features, and normalizing numerical features. The analysis involved visualizations and statistical methods to identify trends in EV model popularity, spatial distribution, and relationships among vehicle characteristics.

The findings highlight trends in electric vehicle adoption and emphasize the importance of data preprocessing and EDA in machine learning workflows. This analysis provides valuable insights for stakeholders regarding EV infrastructure and policy development in Washington State.

2. Dataset overview.

The "Electric Vehicle Population Data" dataset, sourced from the Data.gov repository and provided by the Washington State Department of Licensing, contains comprehensive information about electric vehicles (EVs) registered in Washington State. It includes both Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs), reflecting EV adoption in the region.

- **Number of Records (Rows):** 210,165 entries, representing individual EV registrations.
- **Number of Features (Columns):** 17 features, detailing each vehicle's attributes and registration information.

Key Features

- **VIN (1-10):** The first 10 characters of the Vehicle Identification Number, unique to each vehicle.
- **County and City:** The location of registration, allowing for spatial analysis of EV distribution.
- **Model Year:** The production year of the vehicle.
- **Make and Model:** The manufacturer and specific model of the vehicle, providing insights into brand popularity.
- **Electric Vehicle Type:** Indicates whether the vehicle is a BEV or PHEV.
- **Electric Range:** The estimated range of the vehicle on electric power alone.
- **Base MSRP:** The Manufacturer's Suggested Retail Price, providing a general cost indication.
- **Legislative District:** The Washington legislative district where the vehicle is registered.
- **Vehicle Location:** The geographical coordinates of the registration, useful for spatial mapping.
- **Electric Utility:** The utility provider servicing the vehicle's registration area.

2. Data Preprocessing and Feature Engineering.

2.1 Document Missing Values.

To ensure data quality, an initial check was conducted to identify missing values across all features. The summary of missing values and their percentages is shown in the figures below.

Mising Values Summary:

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

dtype: int64

Filtered Summary of Missing Values (only columns with missing values)

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

The feature with the most missing values is *Legislative District*, with 445 missing entries, representing approximately 0.21% of the data. Other features, such as *Electric Range*, *Base MSRP*, and *Vehicle Location*, have very low percentages of missing data (less than 0.01% each). Overall, the low percentage of missing data across all features suggests that data quality is high and suitable for analysis following appropriate imputation strategies.

2.2 Missing Value Strategies.

To address missing data in the dataset, we applied a targeted approach based on the percentage of missing values for each feature.

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

For features with minimal missing values (e.g., *County*, *City*, *Postal Code*), rows with missing data were dropped to maintain data integrity without significant loss. For *Legislative District*, which had a higher percentage of missing values, median imputation was used to preserve central tendencies while reducing outlier impact. After applying these strategies, all missing values were resolved, ensuring a complete dataset as shown in Figure above.

2.3 Feature Encoding.

To make categorical features suitable for analysis, we applied one-hot encoding to convert each unique category into a binary feature. This transformation was applied to columns such as *Make*, *Model*, *County*, *City*, *Electric Vehicle Type*, and *Electric Utility*, among others.

ility_PORTLAND GENERAL ELECTRIC CO	Electric Utility_PUD NO 1 OF CHELAN COUNTY	Electric Utility_PUD NO 1 OF DOUGLAS COUNTY	Electric Utility_PUD NO 1 OF OKANOGAN COUNTY	Electric Utility_PUD NO 1 OF PEND OREILLE COUNTY	Electric Utility_PUD NO 1 OF WHATCOM COUNTY	Electric Utility_PUD NO 2 OF GRANT COUNTY	Electric Utility_PUGET SOUND ENERGY INC	Electric Utility_PUGET SOUND ENERGY INC||CITY OF TACOMA - (WA)	Electric Utility_PUGET SOUND ENERGY INC||PUD NO 1 OF WHATCOM COUNTY
False	False	False	False	False	False	True	False	False	False
False	False	False	False	False	False	False	True	False	False
False	False	False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	False	True	False
False	False	False	False	False	False	False	False	False	False

One-hot encoding expanded the dataset's dimensionality by adding a binary column for each unique category, as seen in Figure above. This approach ensures that categorical information is accurately represented without implying any ordinal relationships, which is essential for machine learning models to interpret the data correctly. Although it increased the number of columns, it preserved the integrity of categorical data, allowing for precise analysis.

2.4 Normalization.

In this part *Electric Range* and *Base MSRP* were selected for normalization due to their significant value ranges, which can disproportionately impact analysis and machine learning models if left unprocessed. The Electric Range feature, varying between 0 and 337 miles, is central to understanding the vehicle's electric performance. Similarly, Base MSRP has a broad scale, extending from $0 up to $845,000, and requires normalization to minimize any skew caused by high-cost outliers. Applying normalization to these features maintains balance across attributes and improves model performance by reducing bias from numerical magnitude.

	Electric Range	Base MSRP
count	2.101500e+05	2.101500e+05
mean	-2.407359e-17	9.196651e-18
std	1.000002e+00	1.000002e+00
min	-5.818120e-01	-1.172560e-01
25%	-5.818120e-01	-1.172560e-01
50%	-5.818120e-01	-1.172560e-01
75%	-9.890965e-02	-1.172560e-01
max	3.292904e+00	1.103029e+02

Z-score normalization was applied to the Electric Range and Base MSRP features, which have widely varying ranges. Standardization is ideal here as it centers on a mean of zero with a standard deviation of one, neutralizing the impact of large value ranges. This approach improves comparisons in analysis, reduces biases in statistical models, and enhances the clarity of visualizations by putting features on a similar scale, enabling more accurate insights.

3. Exploratory Data Analysis.

3.1 Descriptive Statistics.

Descriptive statistics were calculated for key numerical features to understand their central tendency and variability. The **mean** provides the average value, while the **median** offers a midpoint that is less influenced by outliers. The **standard deviation** shows the spread of values, indicating which features have more variability.

```
Mean values:
 Postal Code              9.817819e+04
Model Year                2.021049e+03
Electric Range           -2.407359e-17
Base MSRP                 9.196651e-18
Legislative District      2.893684e+01
DOL Vehicle ID            2.290765e+08
2020 Census Tract         5.297929e+10
```

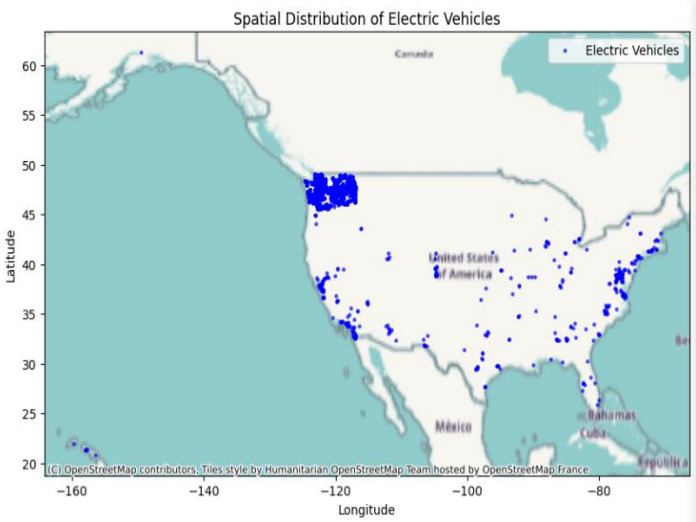
```
Median values:
 Postal Code              9.812500e+04
Model Year                2.022000e+03
Electric Range           -5.818120e-01
Base MSRP                -1.172560e-01
Legislative District      3.200000e+01
DOL Vehicle ID            2.405161e+08
2020 Census Tract         5.303303e+10
```

```
Standard deviation values:
 Postal Code              2.445491e+03
Model Year                2.988946e+00
Electric Range            1.000002e+00
Base MSRP                 1.000002e+00
Legislative District      1.489343e+01
DOL Vehicle ID            7.115445e+07
2020 Census Tract         1.551507e+09
```

The descriptive statistics highlight key trends: Model Year averages show most EVs are recent, while standardized Electric Range and Base MSRP confirm effective normalization. Legislative District and Postal Code variations reflect regional diversity, and high standard deviations in DOL Vehicle ID and 2020 Census Tract indicate wide-ranging unique identifiers. These insights set the stage for further analysis of EV adoption trends

3.2 Spatial Distribution.

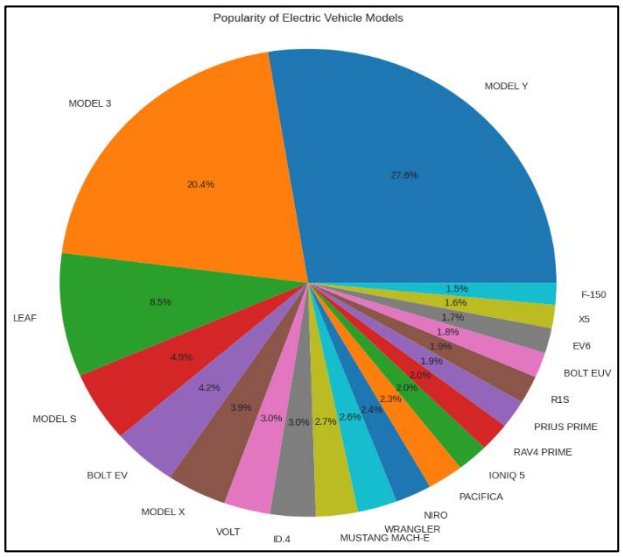
In this part we visualize the spatial distribution of EVs across locations using GeoDataFrame



The plot reveals clusters of EV registrations, particularly concentrated in the state of Washington. By visualizing these spatial data points, we can identify regions with **high** or **low** EV adoption rates, which can inform infrastructure planning and targeted initiatives. This map provides an overview of the geographic distribution of electric vehicles across the country, showing where EVs are most commonly registered.

3.3 Model Popularity.

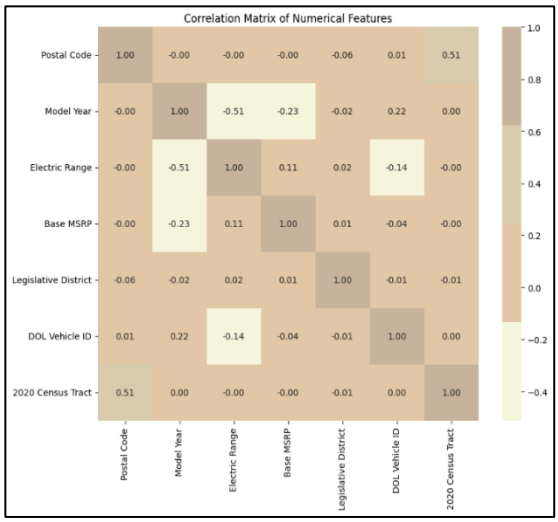
This section analyzes the popularity of different EV models.



The chart shows that Model Y and Model 3 are the most popular electric vehicles, accounting for 25.1% and 18.6% of the dataset, respectively. Other models, such as the Leaf and Model S, also hold significant shares but are noticeably less popular than the top two models. This trend suggests that Tesla models dominate the EV market, particularly with their Model Y and Model 3 vehicles. The data provides insights into consumer preferences, which could be valuable for automakers and policy makers in understanding market demand.

3.4 Investigate the relationship between every pair of numeric features.

In this section, we examine relationships between pairs of numerical features.

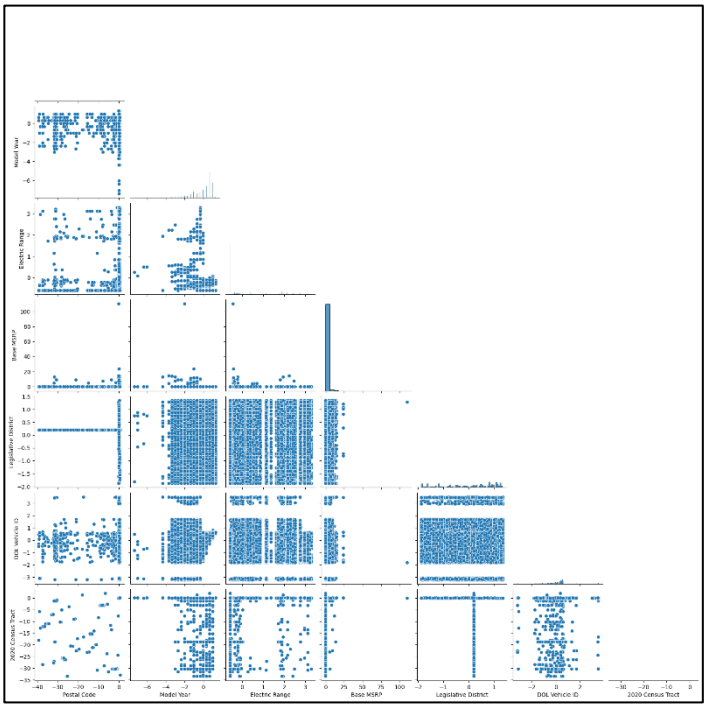


The **correlation matrix** reveals a significant negative correlation (-0.51) between **Model Year** and **Electric Range**, indicating that older vehicles generally have shorter ranges, likely due to advancements in battery technology. There's also a moderate positive correlation (0.51) between **Postal Code** and **2020 Census Tract**, possibly reflecting geographic clustering within regions. Other features, like **Base MSRP** and **Legislative District**, show low correlation with each other, suggesting these variables are largely independent.

4. Visualization.

4.1 Data Exploration Visualizations.

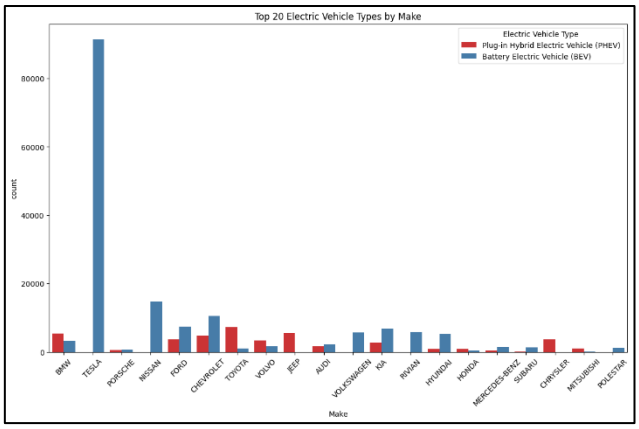
The scatter matrix analysis reveals several trends among the variables. A notable trend is that as the model year increases, the base MSRP also tends to rise, indicating that newer models generally have higher prices. There is a slight positive correlation between electric range and base MSRP, suggesting that vehicles with a greater electric range are typically more expensive. Additionally, newer models may feature longer electric ranges due to advancements in technology



In contrast, the distribution of DOL Vehicle ID and legislative district appears random, indicating no significant relationship between them. The 2020 Census Tract also shows a random distribution concerning other variables, suggesting no notable relationships.

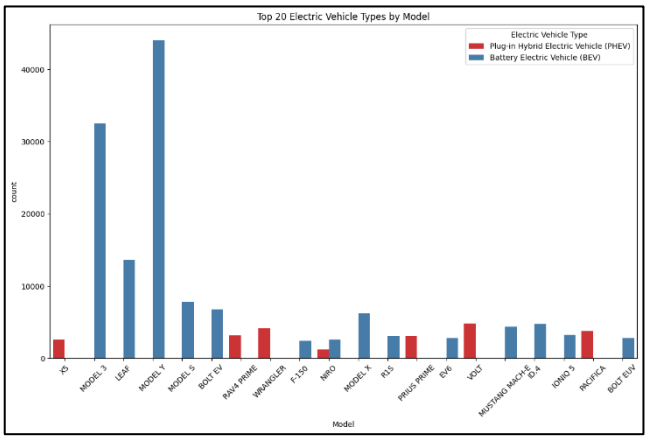
Price variations based on postal codes indicate market diversity, while certain legislative districts may have a higher concentration of newer model vehicles. Furthermore, electric range could vary by postal code, reflecting urban versus rural preferences. Overall, the analysis highlights potential connections between vehicle characteristics and their geographical or socioeconomic contexts.

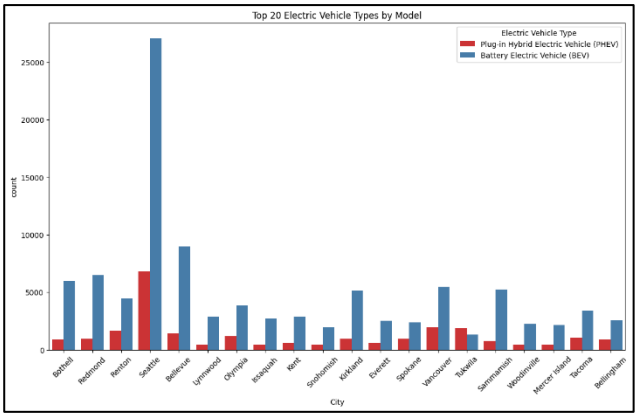
Electric vehicle Types Relations:



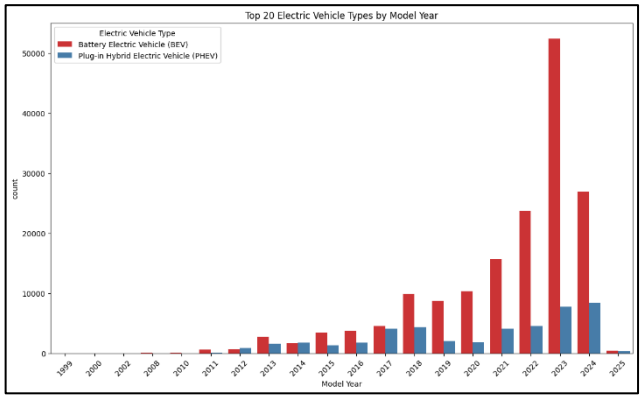
The bar plot illustrates the distribution of Electric vehicle (EV) types—Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs)—among the top 20 car makes. Tesla ranks highest, primarily with BEVs, underscoring its dominance in the EV market, while brands like Nissan and Chevrolet present a mix of BEVs and PHEVs.

A separate plot highlights the top 20 vehicle models, showcasing Tesla's Model Y and Model 3 as popular BEVs. Other models, such as the Nissan Leaf and Chevrolet Bolt EV, also demonstrate significant registrations, reflecting the demand for specific EV offerings across different manufacturers.





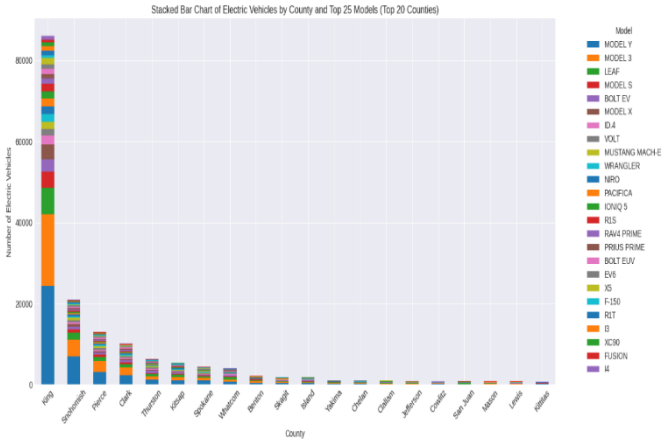
When examining the distribution of EV types in major cities, Seattle emerges as a leader in registered EVs, primarily BEVs from Tesla. Cities like Bellevue and Vancouver feature a combination of BEV and PHEV registrations, indicating a concentration of urban EV adoption.



Finally, the analysis of EV types by model year reveals a notable increase in both BEV and PHEV adoption, particularly since 2020, fueled by advancements in technology and growing environmental awareness.

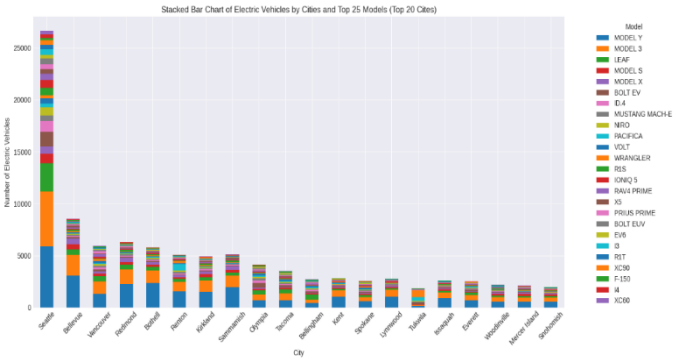
4.2 Comparative Visualization.

We compared EV distribution across the top 20 counties and cities using stacked bar charts.



Counties Analysis: The chart shows the distribution of EV models across the top 20 counties with the highest EV registrations. King County stands out with the highest EV count, particularly of popular models like Model Y, Model 3, and LEAF, indicating a high EV concentration. Snohomish and Pierce counties also have significant numbers, but with a more balanced distribution of models. These regional trends help highlight areas with high EV adoption, valuable for planning and market analysis.

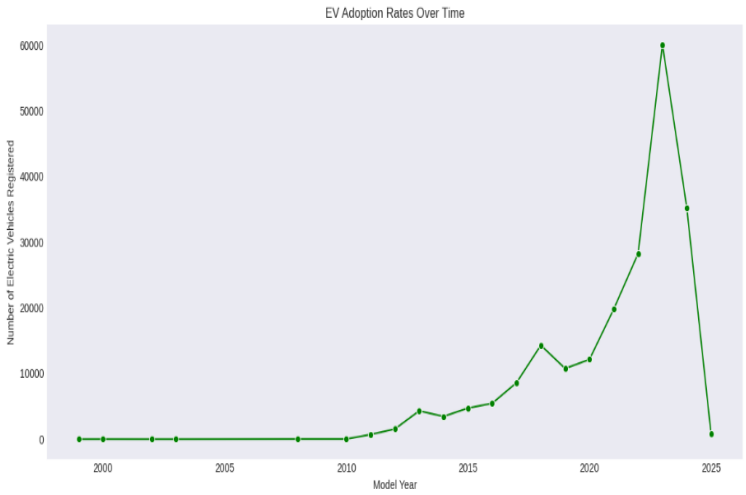
Cities Analysis: The city-level chart reveals similar trends, with Seattle leading in EV registrations and models such as Model Y and Model 3 dominating. Bellevue and Vancouver also exhibit high EV counts with a mix of models. While Seattle has the highest concentration, a diverse distribution in smaller cities suggests a broader spread of EV adoption across urban areas in Washington.



These charts provide insights into where specific EV models are most popular, supporting decisions in infrastructure and market strategies.

4.3 Analyze the temporal trends in EV adoption rates and model popularity.

We analyzing the dataset over multiple time points provides insights into EV adoption rates and model preferences.



EV Adoption over Time: The line plot of EV registrations by model year shows a steady growth in EV adoption up until 2018, followed by a sharp rise that peaks in 2022. This surge likely reflects advancements in EV technology, increased environmental awareness, and supportive policies for EV ownership. A recent dip may be due to incomplete data for the latest year or external factors affecting registrations.

Model Popularity Trends: The line plot of the top 10 EV models reveals a rapid increase in registrations for Tesla's Model Y and Model 3, underscoring their strong market dominance in recent years. Other models, such as the Leaf, Model S, and Bolt EV, also show steady popularity but remain lower in numbers compared to Tesla's leading models. These patterns highlight evolving consumer preferences, with newer models gaining traction, likely due to better features, improved range, and brand reputation.