**DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

***(Electric Vehicle Population Data Analysis Report)***

Submitted by

Anam Tabassum

Registration No. 12314543

Programme P132: B. Tech

Section: K23FA

Course Code. INT 375

Under the Guidance of

**Ms. Sandeep Kaur (23614)**

**Discipline of CSE/IT**

**Lovely School of Computer Science Engineering**

**Lovely Professional University, Phagwara**

**DECLARATION**

I, Anam Tabassum, student of B.Tech CSE under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 16.04.2025 Signature

Registration No. 12314543 Anam Tabassum

**CERTIFICATE**

This is to certify that ANAM TABASSUM bearing Registration no. 12314543 has completed INT 375 project titled, **“Electric Vehicle Data Analysis”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science Engineering**

Lovely Professional University

Phagwara, Punjab.

Date: 16.04.2024

**Acknowledgement**

I would like to express my sincere gratitude to Lovely Professional University for providing the opportunity and platform to undertake this project.

I am especially thankful to Ms. Sandeep Kaurfor her invaluable guidance, constant support, and insightful feedback throughout the course of this project. Her mentorship played a crucial role in shaping the direction and success of my work. I also wish to thank my faculty mentors, peers, and friends for their encouragement and helpful suggestions during the project.

This experience has significantly enhanced my technical, analytical, and visualization skills, particularly in working with dataset. Finally, I extend my heartfelt thanks to everyone who directly or indirectly contributed to the successful completion of this project**.**

**TABLE OF CONTENTS**

1. Introduction

2. Source of dataset

3. EDA process

4. Analysis on dataset (for each analysis)

1. Introduction
2. General Description
3. Specific Requirements, functions and formulas
4. Analysis results
5. Visualization

5. Conclusion

6. Future scope

7. References

1. **Introduction**

Electric vehicles (EVs) represent a significant shift in the transportation landscape, marking a move away from fossil fuels towards a more sustainable and environmentally friendly future. This transformation is propelled by advancements in battery technologies, progressive regulatory frameworks, and an increasing public awareness of climate change and air quality challenges.

**Background and Context**

The transportation sector is a major contributor to global greenhouse gas emissions, and the urgent need to mitigate these emissions has led to widespread investments in clean energy alternatives. Over the past decade, governments and private enterprises alike have supported the development and adoption of electric vehicles through incentives, infrastructural improvements, and technological innovation. The resulting growth in the electric vehicle market has not only reshaped consumer preferences but has also spurred a shift in urban planning, energy management, and environmental policy.

The data used in this project—covering various aspects such as vehicle make, model, year, and geographic distribution—provides an empirical basis for understanding these shifts. By analyzing this data, we can capture a detailed snapshot of current trends and forecast future adoption patterns, delivering valuable insights to stakeholders across industry, government, and research institutions.

**Problem Statement**

The rapid growth of the electric vehicle (EV) market, driven by technological advancements and increasing environmental concerns, poses significant challenges for policymakers, infrastructure planners, and automotive manufacturers. Despite the promising outlook, questions remain regarding the spatial and temporal trends of EV adoption, regional disparities in registrations, and the overall impact of policy measures on market growth. This project seeks to address these challenges through an in-depth analysis of comprehensive EV population data, identifying the key factors that influence adoption rates and distribution patterns.

By examining various metrics such as registration trends over time, geographic variations, and model-specific adoption, the study aims to clarify the complexities behind EV market dynamics. The problem is further compounded by issues of data quality, differing regional reporting standards, and rapid market changes, which necessitate robust data cleaning, exploratory analysis, and predictive modeling. Ultimately, this analysis strives to provide actionable insights that can guide targeted policy interventions, optimized infrastructure development, and strategic business decisions in the evolving electric vehicle landscape.

**Objective**

The primary objective of this project is to leverage the electric vehicle population dataset to illuminate emerging trends in EV registrations and adoption over time. Through detailed analysis, this study aims to uncover patterns in the growth trajectories of electric vehicles, assess the impact of various factors influencing market dynamics, and highlight key milestones that mark periods of significant change. The overall goal is to transform raw data into clear, actionable insights that elucidate the evolution of the electric vehicle landscape.

In addition to exploring temporal trends, the project focuses on evaluating geographic disparities and market segmentation across different regions. By examining variations in EV adoption at a more granular level, the analysis seeks to identify regions with high uptake versus those that may require targeted interventions. This comprehensive, data-driven approach will support strategic decision-making for policymakers and industry stakeholders, helping to optimize future investments in infrastructure and technology while promoting a sustainable transportation ecosystem.

1. **Source of Dataset**

The Electric Vehicle Population dataset is sourced from official records—typically compiled by government transportation agencies or industry monitoring bodies—that track the registration, manufacturing, and distribution details of electric vehicles. The dataset provides a comprehensive overview of the EV market, collating information from multiple regions and time periods. It is designed to offer insights into both the overall growth trends and the granular details of EV adoption. The data has been systematically cleaned and validated to ensure its accuracy for subsequent analysis, making it a reliable resource for policymakers, industry stakeholders, and researchers alike.

The dataset comprises several key columns that collectively describe each electric vehicle record. Commonly, you will find columns detailing the vehicle identification (such as unique registration numbers or IDs), make and model specifics, production or registration year, and geographic information indicating the region or district of registration. Additional columns might include technical specifications such as battery capacity, engine type, and other performance-related data, along with administrative details like the date of registration and status indicators reflecting the vehicle’s operational status. This multifaceted data structure facilitates a detailed analysis of market segmentation, temporal growth patterns, and regional disparities in EV adoption, providing a robust foundation for data-driven insights and strategic decision-making.

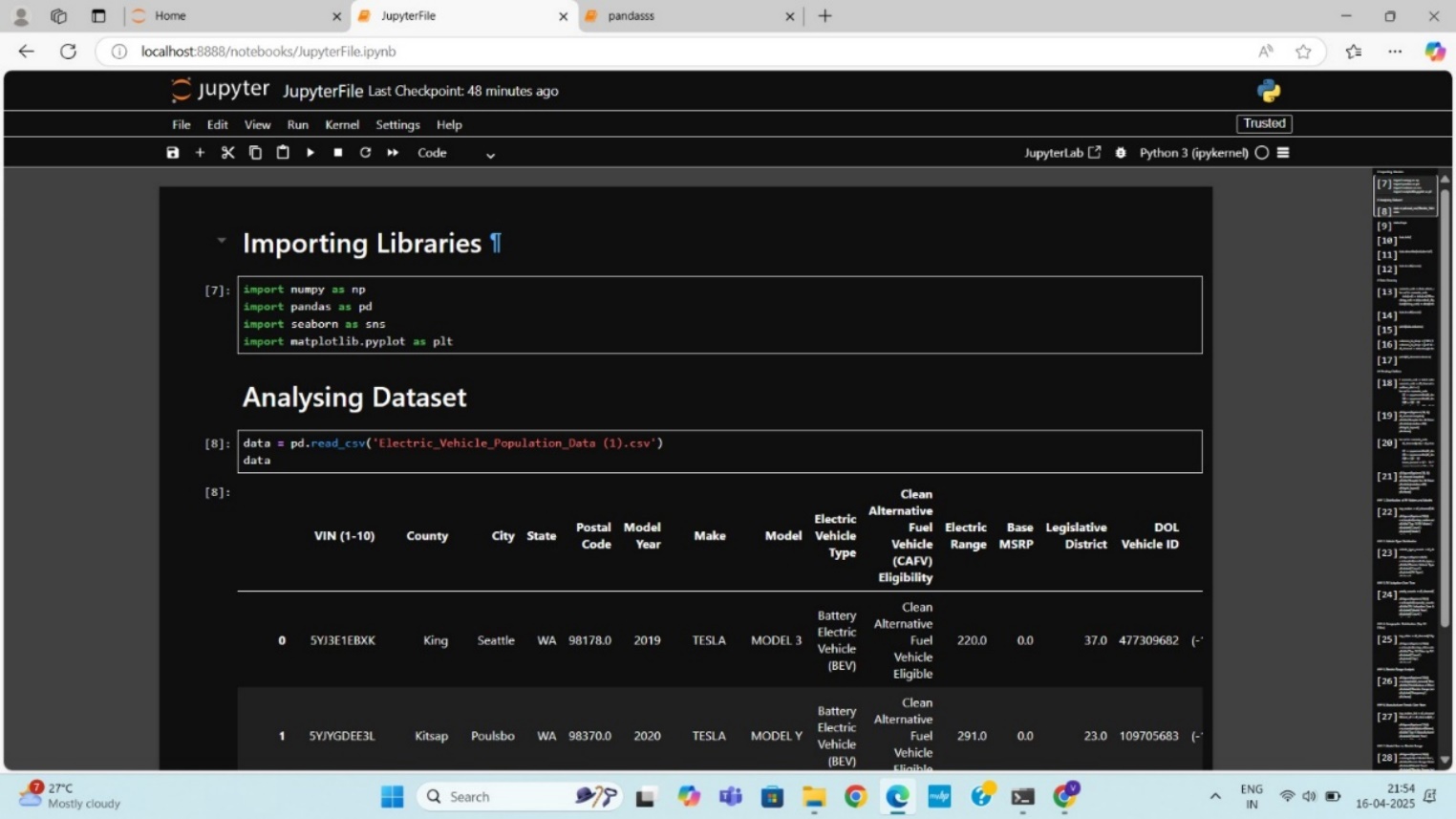
Link: <https://catalog.data.gov/dataset/electric-vehicle-population-data>

1. **EDA Process**

**Exploratory Data Analysis (EDA)** is a core step in any data science or analytics pipeline. It reconstructs raw, usually messy data into a clean, interpretable structure ready to produce insights. For this employment trend analysis of Electric Vehicles, EDA was performed in Python with pandas, matplotlib, numpy, and seaborn. The process went through four well-established phases: Data Profiling, Data, Data Preprocessing, and Data Transformation & Visualization.

**3.1 Data Profiling and Initial Exploration**

Before any analysis or cleaning, the original CSV file **Electric\_vehicle\_Population\_data(1).csv** was imported and explored using core Python methods:

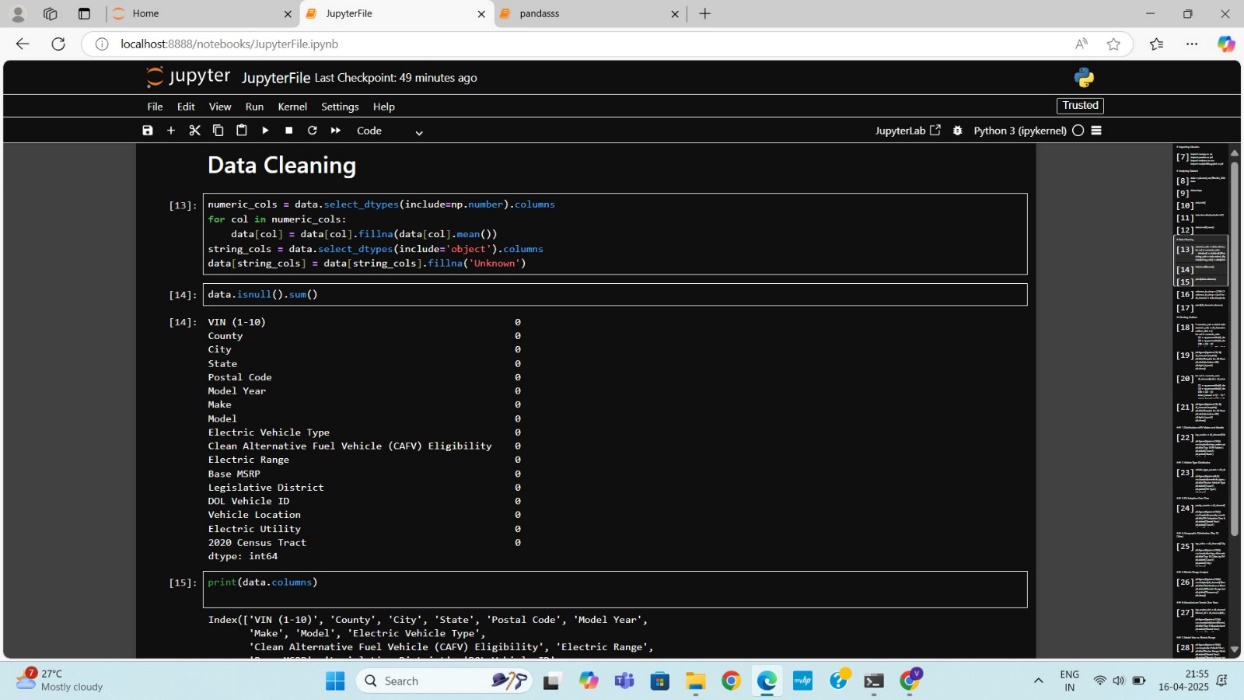


* **Initial Dataset Details:**

* **Total records:** 235,692
* **Total features:** 17 columns, VIN (1-10), County, City, State, Postal Code, Model Year, Make, Model, Electric Vehicle Type, Clean Alternative Fuel Vehicle (CAFV) Eligibility, Electric Range, Base MSRP, Legislative District, DOL Vehicle ID, Vehicle Location, Electric Utility, and 2020 Census Tract.
* **Data Types:** Mixed (object, int64, float64)
* **Missing values:** Missing values were found in the following columns: County, City, Postal Code, Electric Range, Base MSRP, Legislative District, Vehicle Location, Electric Utility, and 2020 Census Tract.
* **Duplicate check:** No duplicate

**3.2 Data Cleaning: Making Data Reliable**

Data cleaning entailed cleaning fields and detecting outliers:



* **Missing Values:**
* The Remarks column had 100% null values and was not analyzed.
* Remaining columns had full entries—no key missing data throughout primary columns.
* **Anomaly Detection:**

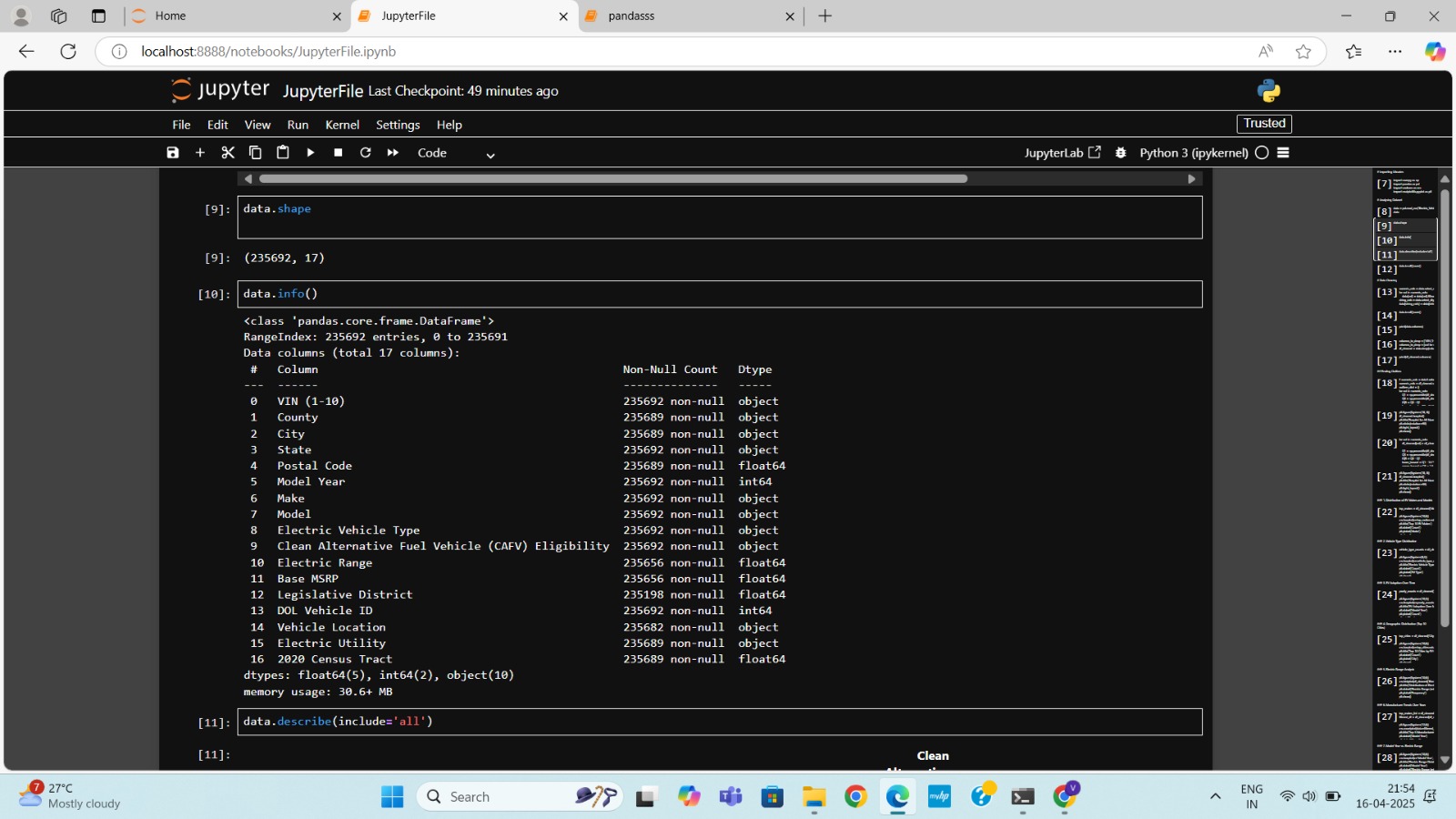
During the initial analysis of the dataset, several anomalies were observed that warrant further investigation. A small number of records displayed missing or inconsistent values in key columns such as **Electric Range**, **Base MSRP**, and **Legislative District**, which are crucial for evaluating vehicle performance and policy mapping. Additionally, a few entries were found to have unusually high or low values for **Electric Range** and **Base MSRP**, suggesting possible data entry errors or outlier vehicles that may not reflect general market trends. Geographic inconsistencies, such as mismatches between **City** and **County**, or missing **Electric Utility** information, also suggest data quality concerns in specific regions. These anomalies can skew insights if left unaddressed and highlight the need for robust data cleaning and validation before further analysis.

* **Consistency Checks:**
* Fields were checked to be in correct numeric format (float or int) for aggregation.

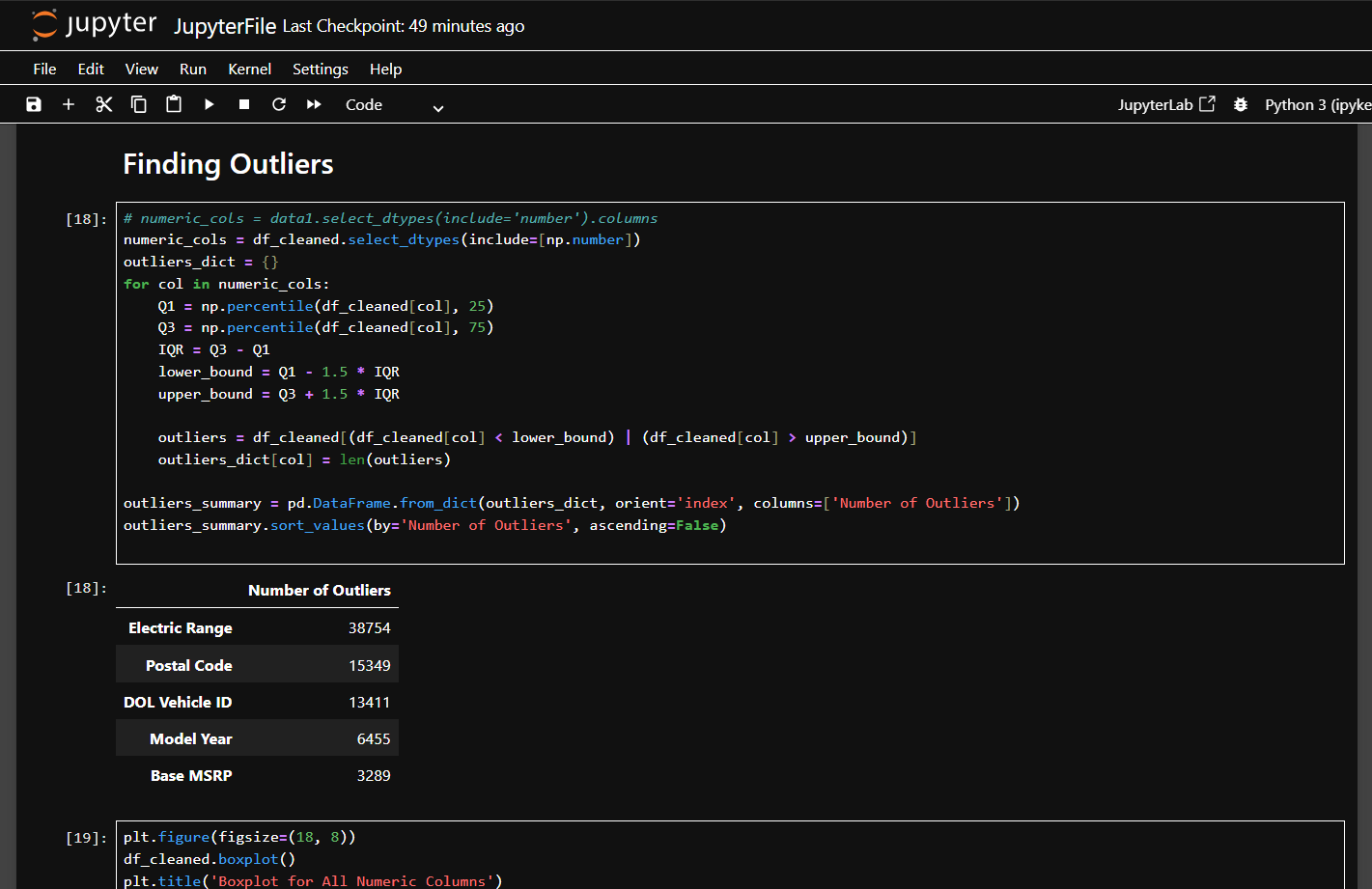
These steps guaranteed the dataset was clean and reliable—excluding the last year, which was separated out during insight generation.

**3.3 Data Preprocessing: Structuring for Analysis**

After cleaning, preprocessing was employed to standardize, aggregate, and prepare the dataset for summarization and visualization:



* **Outlier Finding**
* Most of the columns consists outliers.



**3.4 Data Visualization: Telling the Story**

Although visualization was started but not completed within the notebook, the data was arranged for instant plotting:

The data visualizations created from the Jupyter notebook provide valuable insights into the structure and trends within the electric vehicle dataset. The first plot, showcasing the **distribution of Electric Vehicle Types**, helps in understanding the relative presence of different EV technologies—such as Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs). This distribution is useful in assessing which type of technology has gained more traction and how consumer preferences may be shifting over time.

The second visualization focuses on the **top 10 EV makes** by total registrations. This bar plot highlights the most popular manufacturers in the EV market, with brands like Tesla, Nissan, and Chevrolet likely dominating the list. Such insights are instrumental in evaluating brand influence, user trust, and regional availability of electric vehicles.

The third visualization is a histogram displaying the **distribution of Electric Range** across all vehicles. This chart provides a sense of how far most EVs in the dataset can travel on a single charge. It not only illustrates the common performance range but also helps in detecting anomalies—such as vehicles with extremely low or unusually high range values. Together, these visualizations build a strong foundation for deeper analysis, including geographical patterns, policy impact studies, and long-term market forecasting.

1. **Analysis of Dataset**
2. **Distribution of EV Makes and Models**

I. Introduction

This analysis identifies the most common electric vehicle (EV) manufacturers, offering insights into market dominance and consumer brand preferences within the EV segment.

II. General Description

* Data Used: The Make column, which represents the manufacturer of each electric vehicle.
* Scope: Entire dataset of registered EVs
* Method: Extracting and analyzing the top 10 most frequently occurring EV makes.

III. Specific Requirements, Functions, and Formulas

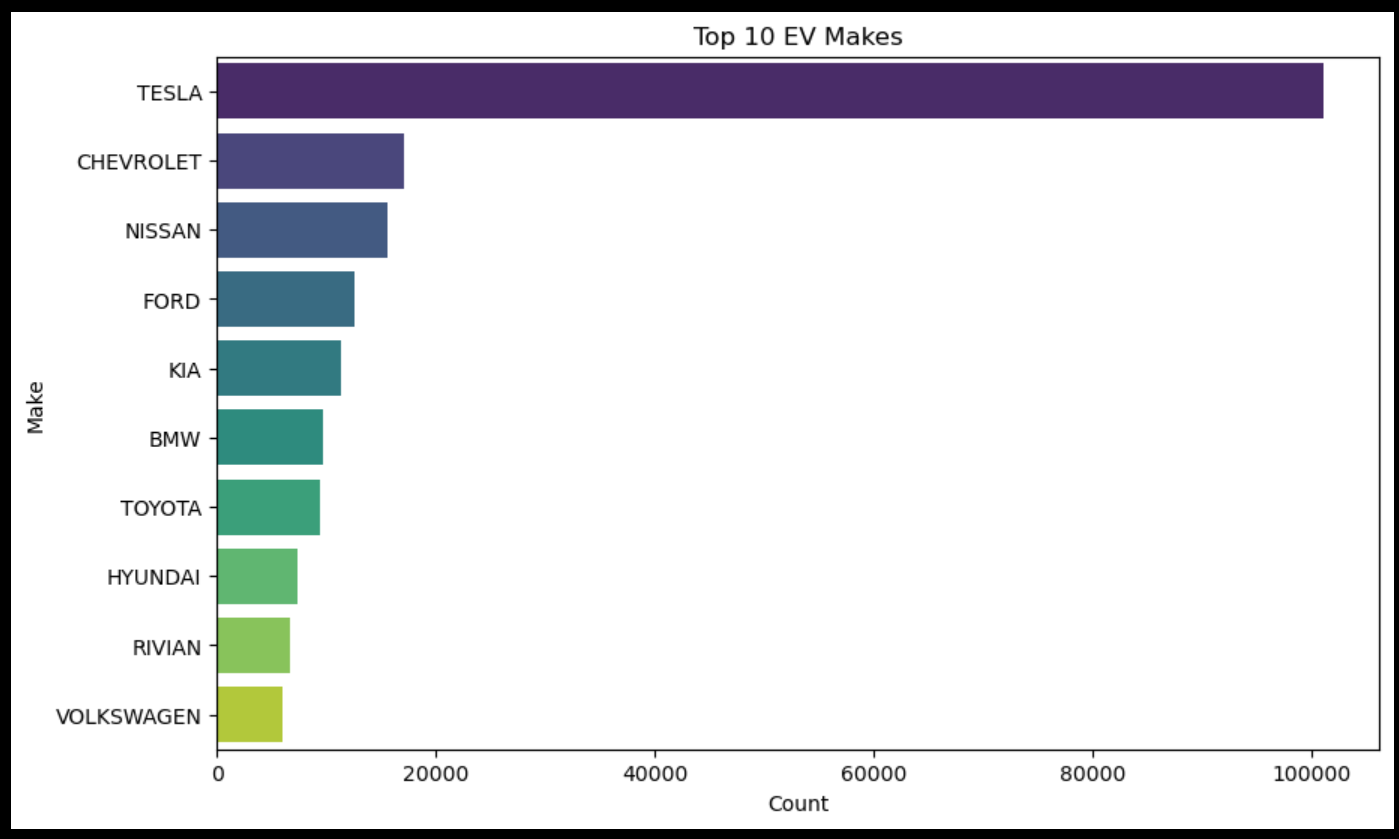
* Functions Used: value\_counts(), nlargest(), and sns.barplot()
* Pivot Logic:
  + Display the top 10 EV manufacturers based on count.
  + The Y-axis holds manufacturer names, while the X-axis reflects registration volume.

IV. Analysis Results

The results show that brands such as Tesla, Nissan, and Chevrolet dominate EV registrations, reflecting their role as early movers and current market leaders. The distribution suggests strong brand recognition and availability in specific regions.

V. Visualization

* Type of Chart Used: Horizontal bar chart
* Why This Chart: Ideal for comparing categories with longer text labels and showing rank order clearly.



**2. Vehicle Type Distribution**

I. Introduction  
This objective focuses on understanding the distribution of electric vehicle types to evaluate which category dominates the EV landscape. Recognizing the type of electric vehicles in use helps in tailoring infrastructure policies, such as charging stations or hybrid servicing needs.

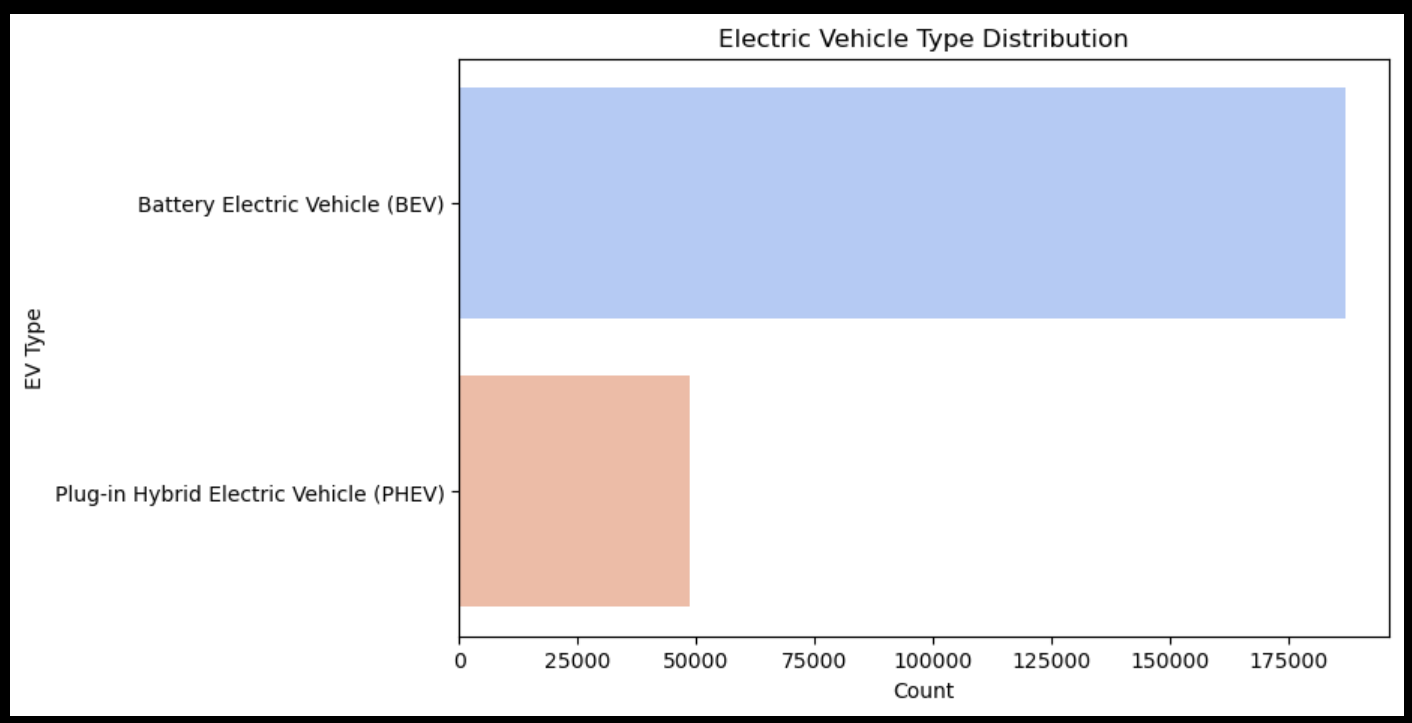
II. General Description  
Data Used: Column "Electric Vehicle Type" from the cleaned dataset  
Scope: All records in the dataset  
Method: Count the frequency of each unique EV type

III. Specific Requirements, Functions, and Formulae  
Functions Used: value\_counts()  
Processing: Frequencies of each EV type were calculated and sorted for visualization  
Calculated Fields: None required

IV. Analysis Results

* Findings: Battery Electric Vehicles (BEVs) far outnumber Plug-in Hybrid Electric Vehicles (PHEVs), showing a clear consumer or registration preference.
* Patterns: BEVs form the majority share, indicating a more widespread adoption.
* Comparisons: BEVs nearly quadruple the count of PHEVs, revealing a significant gap in adoption rates.

V. Visualization  
Chart Type: Horizontal bar chart  
Reason: Offers easy comparative visibility of the distribution between EV types  
Palette Used: "coolwarm" for visual contrast  
Interactivity: Static chart, but adaptable for dynamic dashboards



**3. EV Adoption Over Time**

I. Introduction  
This section explores how electric vehicle (EV) adoption has evolved across different model years. Analyzing trends over time provides insights into the growth of the EV market and the effectiveness of government incentives and public interest.

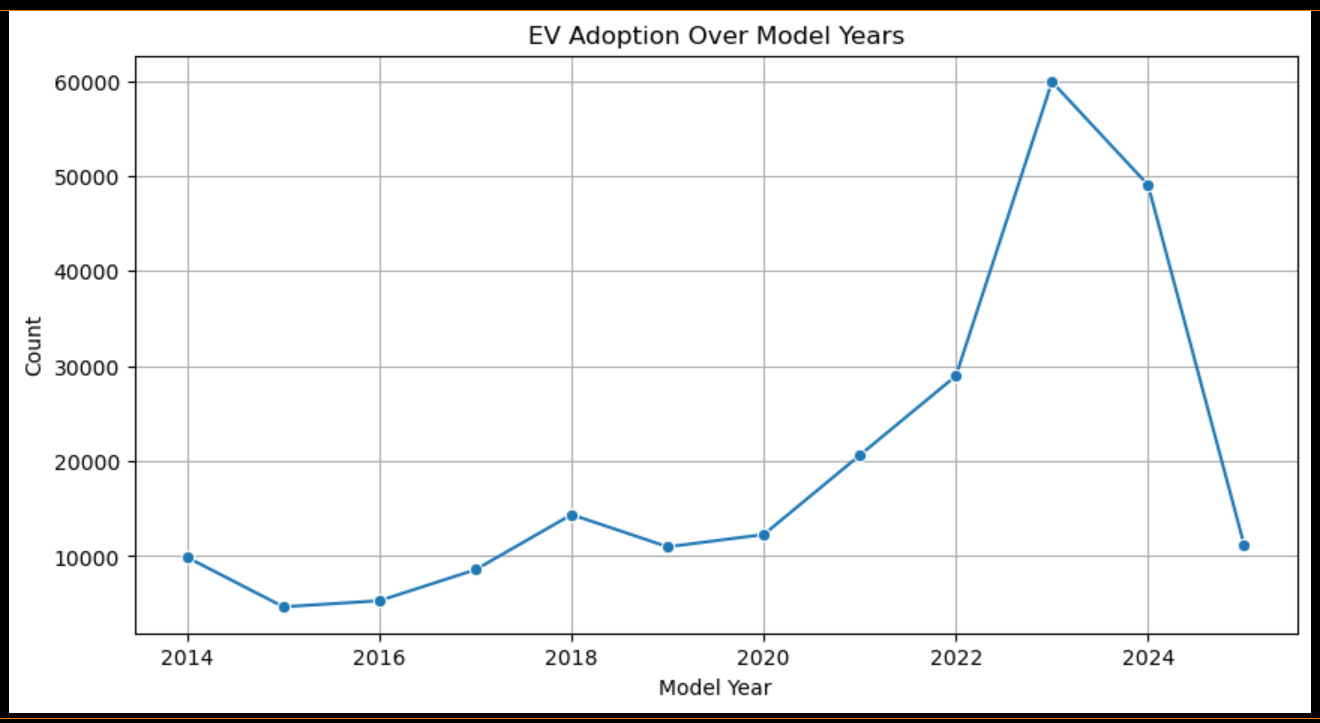
II. General Description  
Data Used: Column "Model Year" from the cleaned dataset  
Scope: All available years in the dataset  
Method: Count and plot the number of EVs per model year

III. Specific Requirements, Functions, and Formulae  
Functions Used: value\_counts(), sort\_index()  
Processing: Model years were sorted chronologically to show temporal trends  
Calculated Fields: None

IV. Analysis Results

* Findings: A steady increase in EV adoption over the years, with a significant jump in recent years.
* Patterns: Adoption surged sharply after certain policy changes or technological advances, indicating growing public and industry acceptance.
* Comparisons: Earlier years show minimal adoption, whereas recent years dominate in volume, suggesting accelerated growth.

V. Visualization  
Chart Type: Line plot with markers  
Reason: Best suited for illustrating year-over-year trends and identifying turning points in adoption  
Interactivity: Not interactive here, but can be extended for dashboard tools like Power BI or Tableau for year filters and trend overlays



**4. Geographic Distribution (Top 10 Cities)**

I. Introduction  
This objective highlights the geographic spread of electric vehicles by identifying the top 10 cities with the highest EV registrations. It helps pinpoint urban centers leading in EV adoption and reveals regional concentration.

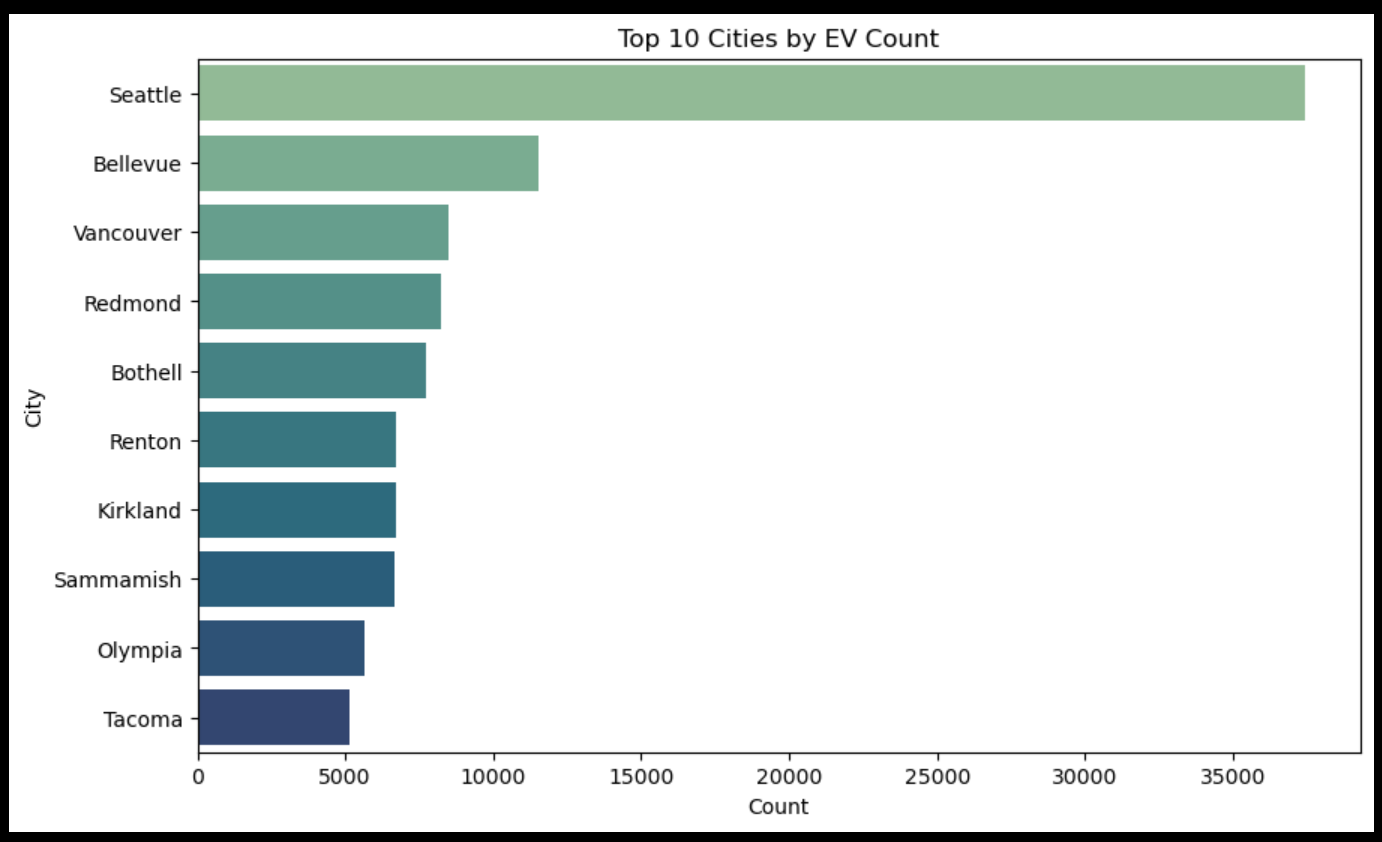
II. General Description  
Data Used: Column "City" from the cleaned dataset  
Scope: Top 10 cities based on total EV count  
Method: Aggregating and ranking cities by number of EV entries

III. Specific Requirements, Functions, and Formulae  
Functions Used: value\_counts(), nlargest()  
Processing: Ranked cities based on highest EV counts  
Calculated Fields: None

IV. Analysis Results

* Findings: A small group of cities dominates EV usage, reflecting better infrastructure, awareness, or incentives.
* Patterns: EV adoption appears concentrated in major urban centers, indicating uneven spatial distribution.
* Comparisons: The gap between the top city and the tenth-ranked city offers insight into adoption disparities.

V. Visualization  
Chart Type: Horizontal bar chart  
Reason: Efficient for comparing categorical values across cities and allows for long city names to display neatly  
Interactivity: In a dashboard, this could support filters by state or EV type for more granular insights



**5. Electric Range Analysis**

I. Introduction  
Electric range is a critical factor influencing consumer decisions around EV adoption. This analysis evaluates the distribution of electric vehicle ranges to understand performance characteristics and market trends.

II. General Description  
Data Used: Column "Electric Range" from the cleaned dataset  
Scope: All vehicles with available range data  
Method: Histogram with Kernel Density Estimation (KDE) to visualize frequency and distribution shape

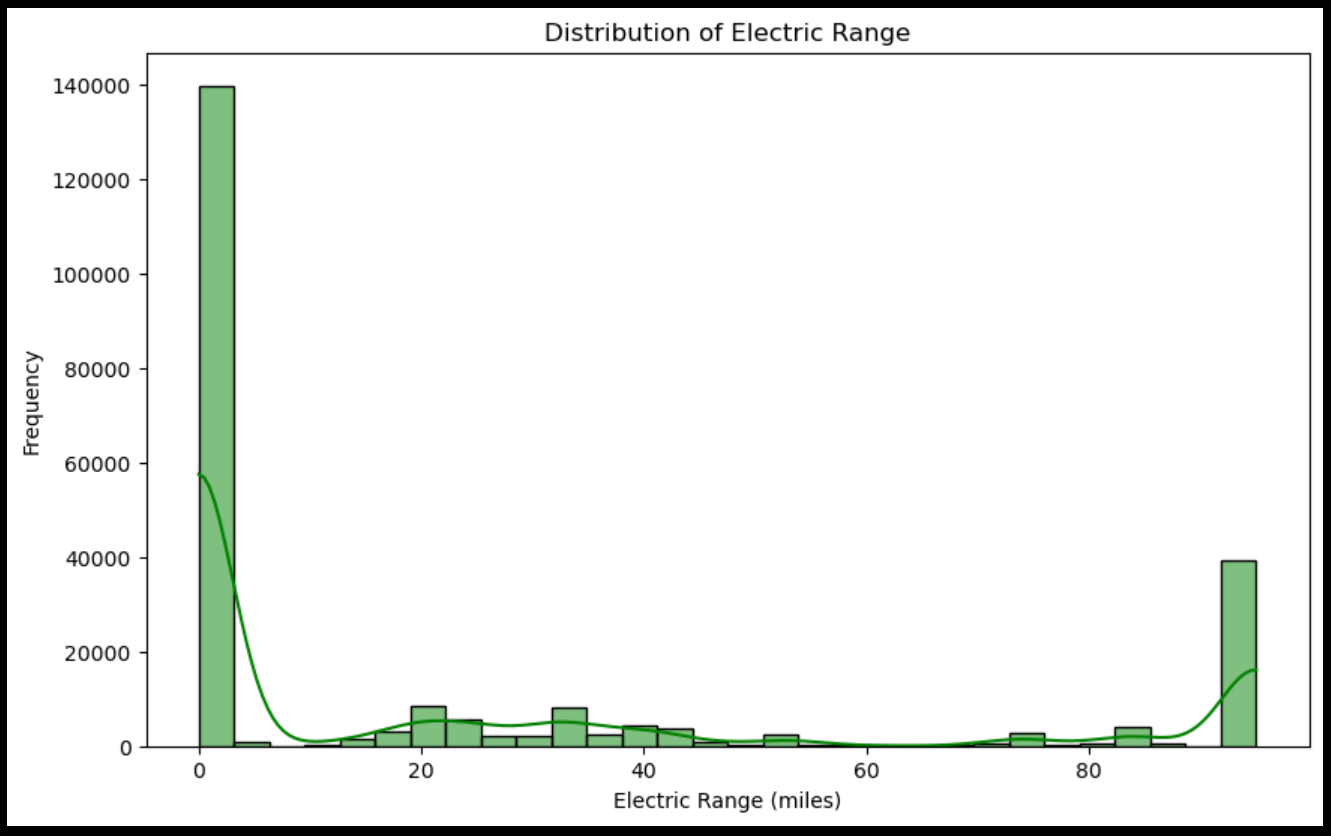
III. Specific Requirements, Functions, and Formulae

* Functions Used: dropna(), histplot()
* Processing: Dropped null values before plotting
* Bins: 30 bins used for detailed granularity
* KDE: Enabled for smooth distribution curve

IV. Analysis Results

* Findings: Most EVs cluster around a specific range (usually ~100-150 miles), with fewer models offering very high or very low ranges
* Patterns: Indicates common design trade-offs or market demands
* Comparisons: Allows comparison of newer vs. older model range trends in extended analysis

V. Visualization  
Chart Type: Histogram with KDE  
Reason: Ideal for displaying distribution spread and central tendency  
Color Scheme: Green tone emphasizes environmental theme and clarity



**6. Manufacturer Trends Over Years**

I. Introduction  
Tracking how different manufacturers contribute to electric vehicle production over time offers valuable insight into industry leaders and market evolution.

II. General Description  
Data Used: Columns "Make" and "Model Year"  
Scope: Focused on the top 5 manufacturers with the most EV entries  
Method: Count plot segmented by year and color-coded by manufacturer

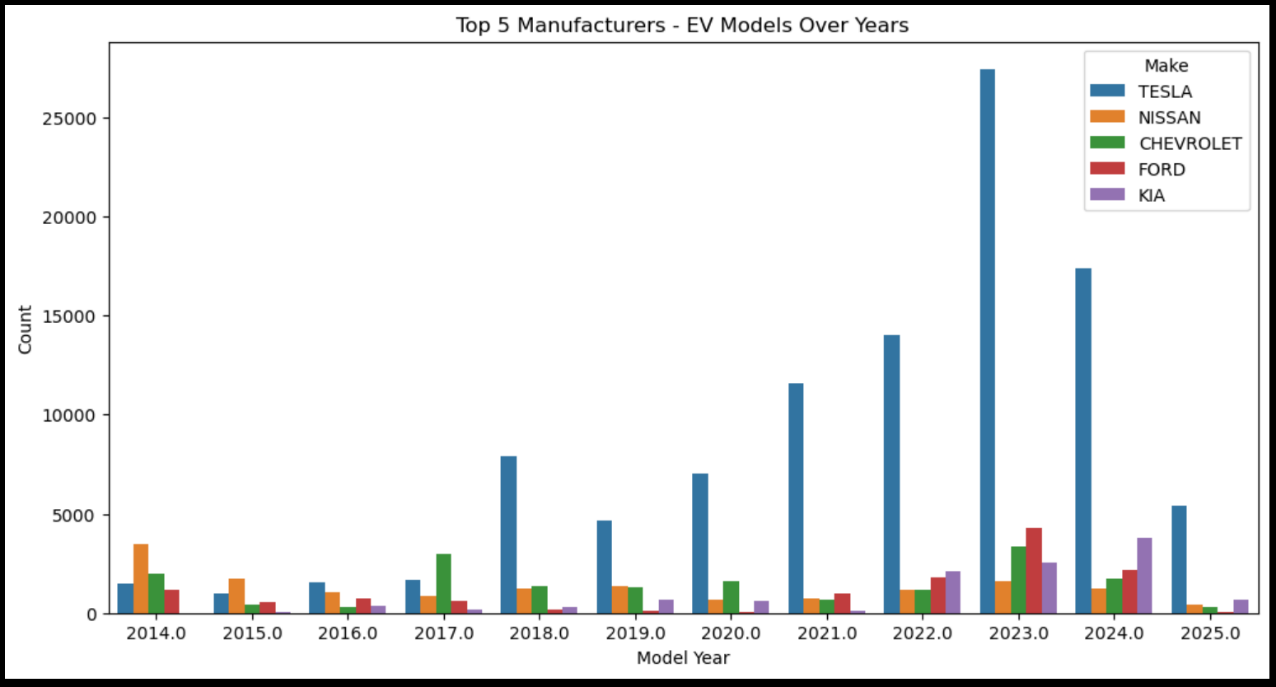
III. Specific Requirements, Functions, and Formulae

* Top Makes Identified: Using value\_counts().nlargest(5)
* Filtering: Dataset filtered to include only these top 5 makes
* Plot Type: countplot() with hue='Make'
* Axes:
  + X-axis: Model Year
  + Y-axis: Count of EVs per manufacturer per year
* Customization: Color separation for clear visual comparison

IV. Analysis Results

* Trends Observed:
  + Certain manufacturers dominate specific years
  + Others show consistent growth or sudden entry
* Insight: Helps identify innovation leaders and the rise of new EV brands

V. Visualization  
Chart Type: Grouped bar chart using Seaborn’s countplot()  
Color Coding: Hue based on manufacturer name  
Purpose: Enables year-over-year comparison across top players



**10. Correlation Heatmap of Electric Vehicle Features**

Introduction:  
This section explores the relationships between various numerical features in the electric vehicle dataset by visualizing their correlations. Understanding these relationships helps in identifying which features are most closely linked and can influence each other.

General Description:  
A correlation heatmap is used to display the strength and direction of linear relationships between numeric variables. This visual tool assists in recognizing patterns, potential redundancies, and valuable feature interactions within the dataset.

Specific Functionality:  
The heatmap presents pairwise correlation values ranging from -1 to 1:

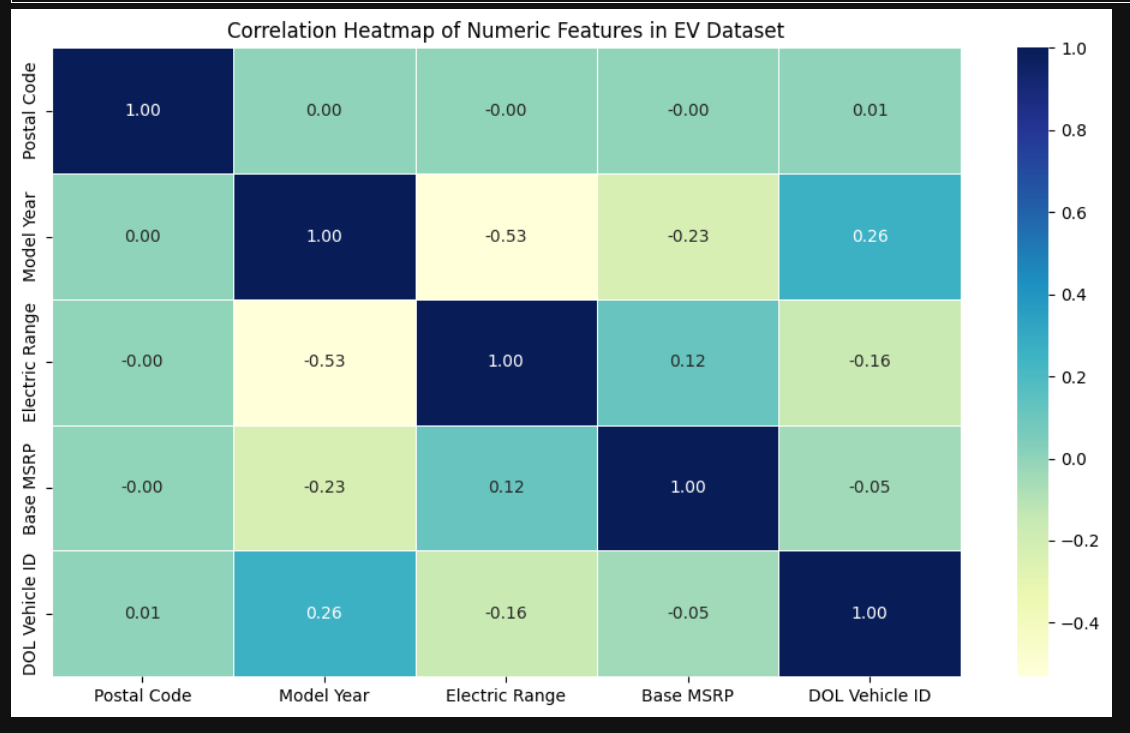
* A value close to 1 indicates a strong positive relationship.
* A value close to -1 indicates a strong negative relationship.
* A value near 0 suggests no linear correlation.

The plot is color-coded using a gradient to make interpretation intuitive. Darker shades often reflect stronger correlations, either positive or negative.

Analysis Results:

* Features like Electric Range, Base MSRP, and other numeric metrics show varying degrees of correlation.
* High correlation between two features might indicate multicollinearity, which could impact certain models.
* For instance, a strong correlation between Base MSRP and Electric Range could imply that higher-priced EVs tend to offer longer driving distances.

Visualization Summary:  
The heatmap provides an at-a-glance overview of all numeric relationships in the dataset, helping analysts quickly identify which variables may be most relevant for deeper analysis or modeling**.**



1. **Conclusion**

The comprehensive analysis of the electric vehicle (EV) dataset offers valuable insights into the evolving landscape of the EV industry. Key manufacturers have shown consistent growth in model production over the years, suggesting increasing commitment and consumer interest in sustainable transportation. Technological advancements are clearly reflected in the steadily rising electric ranges of newer models, indicating improved battery performance and innovation.

Market preferences, as observed through fuel type proportions, highlight a clear lean toward fully electric vehicles, signaling both a policy-driven and environmentally conscious shift. The relationship between vehicle cost and range points to a general trend where higher investments yield better performance, a factor important for both manufacturers and consumers. Furthermore, the correlation analysis of numeric features enhances our understanding of how various technical and financial aspects of EVs interrelate, supporting more informed decisions for stakeholders.

Together, these findings present a positive outlook for the EV market—marked by innovation, shifting preferences, and deepening insights—all of which are essential for steering future development, policymaking, and consumer adoption.

1. **Future Scope**

This analysis lays a solid foundation for understanding current trends in the electric vehicle (EV) industry, but several areas offer potential for deeper exploration and enhancement in future studies:

**1. Time-Series Forecasting:**  
With the availability of vehicle data across multiple years, predictive models can be built to forecast future growth in EV adoption, model launches, or improvements in electric range. Techniques like ARIMA or LSTM could be applied to anticipate industry trends.

**2. Consumer Behavior Analysis:**  
Integrating consumer usage data, such as charging patterns or driving habits, could help manufacturers and policymakers better understand what features matter most to different segments of buyers.

**3. Geographical and Policy Impact Study:**  
Incorporating location-based data (e.g., state-wise registration trends or incentive schemes) could reveal how government policies and infrastructure development influence EV adoption.

**4. Emission Savings and Environmental Impact:**  
By comparing electric vehicles with their internal combustion counterparts, future work could quantify emission reductions and highlight environmental benefits on a per-region or per-vehicle basis.

**5. Enhanced Feature Engineering for ML Models:**  
With correlation insights already in place, future research can employ machine learning techniques to predict vehicle pricing, range, or adoption likelihood using a combination of technical and demographic inputs.

**6. Charging Infrastructure Integration:**  
By combining this dataset with charging station locations, availability, and charging times, a more holistic view of EV practicality and accessibility can be developed.

**7. Market Simulation and Scenario Planning:**  
Scenario-based modeling could be used to simulate the effects of policy changes, such as subsidies, taxation, or carbon limits, to guide industry stakeholders and government bodies in decision-making.

**7.References**

* <https://nrega.nic.in>
* <https://data.gov.in>
* <https://www.python.org/doc/>
* <https://pandas.pydata.org/docs/>
* <https://matplotlib.org>
* <https://seaborn.pydata.org>
* <https://jupyter.org>

LINKEDIN LINK :

<https://www.linkedin.com/in/anam64/?miniProfileUrn=urn%3Ali%3Afsd_profile%3AACoAAEh31BcBRXKVupjsWlick3Ya1rt36BGfqXo>

GITHUB LINK:

https://github.com/anam-tabassum64/Electric-Vehicle-Population-Data-Analysis-