Project Report: Exploratory Data Analysis (EDA) on Electric Vehicle (EV) Data

1. Introduction

Electric vehicles (EVs) have gained significant attention in the past decade due to their environmental benefits and energy efficiency. Governments around the world are promoting the adoption of EVs through incentives and policy frameworks such as Clean Alternative Fuel Vehicle (CAFV) eligibility. This report aims to perform an Exploratory Data Analysis (EDA) on a dataset of electric vehicles, which includes detailed information such as vehicle model year, electric range, base MSRP, and geographical attributes like city, state, and postal code.

The analysis is carried out to understand the distribution of electric vehicles across different regions, the impact of EV features on market price and range, and how legislative and utility factors contribute to EV adoption.

2. Objectives

- Analyze the distribution of electric vehicles by region (state, city, county).
- Explore the relationship between electric range, price (MSRP), and vehicle type.
- Understand the legislative and geographical factors that influence EV adoption.
- Examine how CAFV eligibility impacts EV adoption and electric range.

3. Data Description

The dataset contains the following columns:

- **VIN (1-10)**: First 10 characters of the Vehicle Identification Number, unique for each vehicle.
- **County**: The county where the vehicle is registered.
- **City**: The city where the vehicle is registered.
- **State**: The state where the vehicle is registered.
- **Postal Code**: Postal code of the vehicle’s registered location.
- **Model Year**: Year of the vehicle's model.
- **Make**: Manufacturer or brand of the EV (e.g., Tesla, Nissan, etc.).
- **Model**: Specific model name of the EV.
- **Electric Vehicle Type**: Type of electric vehicle (e.g., Battery Electric Vehicle (BEV), Plug-in Hybrid Electric Vehicle (PHEV)).
- **CAFV Eligibility**: Whether the vehicle qualifies as a Clean Alternative Fuel Vehicle.
- **Electric Range**: The electric range of the vehicle in miles on a single charge.
- **Base MSRP**: Manufacturer's Suggested Retail Price for the vehicle.
- **Legislative District**: The legislative district where the vehicle is located.
- **DOL Vehicle ID**: Department of Licensing's unique vehicle ID.
- **Vehicle Location**: Vehicle location determined by city and state.

- **Electric Utility**: The electric utility company serving the area where the vehicle is registered.
- **2020 Census Tract**: The census tract for the vehicle's registered location based on the 2020 census.

4. Steps Involved in EDA

4.1 Data Cleaning

- **Missing Values**: Checked for missing data in key columns such as Electric Range, Base MSRP, County, and CAFV Eligibility. Missing values were handled using appropriate imputation methods, such as filling missing location data with the mode or interpolating price values.
- **Outlier Detection**: Identified and treated outliers in columns such as Base MSRP and Electric Range using boxplots and Z-score analysis.
- **Data Type Conversion**: Ensured that categorical variables like State, City, and Make were converted to appropriate data types for analysis.

4.2 Univariate Analysis

- **Vehicle Distribution by State**: Analyzed the count of EVs registered in each state using bar charts, identifying which states have the highest number of EV registrations.
- **Electric Range Distribution**: Visualized the distribution of the electric range across different EV models using histograms and boxplots. Most EVs had a range between 100-300 miles, with some luxury models exceeding 300 miles.
- **Base MSRP**: Assessed the price distribution of electric vehicles using histograms and identified the median MSRP for EVs. This provides insight into the price range that consumers are willing to pay.
- **CAFV Eligibility**: Analyzed the proportion of vehicles that are eligible for Clean Alternative Fuel Vehicle (CAFV) status, using pie charts or bar graphs.

4.3 Bivariate Analysis

- **Electric Range vs. Base MSRP**: Investigated the relationship between the electric range and price using a scatter plot. Generally, vehicles with higher ranges had a higher MSRP, but there were some exceptions for premium models.
- **CAFV Eligibility vs. Electric Range**: Compared CAFV-eligible vehicles with non-eligible vehicles in terms of their electric range. CAFV-eligible vehicles tended to have better ranges, suggesting that eligibility might incentivize better performance.
- **State vs. Electric Vehicle Type**: Analyzed how the distribution of Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) varies by state, using stacked bar charts.

4.4 Multivariate Analysis

- **Geographical Distribution**: Used a map visualization to plot the geographical distribution of EV registrations across different counties and states. This helped identify clusters where EV adoption is highest.

- **Vehicle Model, Make, and Year Trends**: Explored the popularity of different vehicle makes (e.g., Tesla, Nissan) and models over the years. This was done through grouped bar charts and line charts showing the distribution of EV sales by year.
- **CAFV Eligibility Impact on Pricing and Range**: Investigated the combined effect of CAFV eligibility, vehicle type (BEV or PHEV), and MSRP on the electric range of vehicles. Multivariate scatter plots helped identify how eligibility impacts the relationship between range and price.
- **Legislative District and EV Adoption**: Explored the distribution of EVs across legislative districts using heatmaps to analyze if policy-driven factors are encouraging EV adoption in certain areas.

5. Findings and Insights

- **State-Wise Adoption**: California, Washington, and New York lead the country in terms of EV registrations. These states also tend to have more supportive policies and better charging infrastructure.
- **Electric Range and Price Correlation**: There is a positive correlation between electric range and MSRP. Premium vehicles with longer ranges are priced significantly higher, with some models exceeding $100,000.
- **CAFV Eligibility**: Vehicles that are CAFV-eligible tend to have a higher electric range and a slightly higher base MSRP, indicating that these vehicles are better optimized for clean energy standards.
- **Model Year Trends**: Newer models (post-2018) show a significant improvement in electric range, with technological advancements allowing manufacturers to offer better performance at competitive prices.
- **Legislative Influence**: Certain legislative districts with EV incentives have higher adoption rates, suggesting that policy and subsidies are critical in driving the growth of the EV market.
- **Brand Dominance**: Tesla dominates the EV market, particularly in the BEV segment, while other brands like Nissan and Chevrolet have a larger presence in the PHEV market.

6. Conclusion

The EDA on the electric vehicle dataset provides a comprehensive understanding of the current state of EV adoption, including geographical trends, the influence of CAFV eligibility, and the relationship between vehicle features and price. The findings suggest that while premium EV models dominate the market, there is a growing trend of affordable EVs with moderate ranges, making them accessible to a wider audience. Additionally, states with supportive legislation and better infrastructure show higher adoption rates.

7. Future Work

- Develop predictive models to forecast the adoption of EVs based on factors such as legislative support, geographical location, and vehicle features.
- Perform clustering analysis to identify consumer segments based on price sensitivity and performance preferences.

- Investigate the impact of charging infrastructure density on EV adoption by incorporating external data on EV charging station availability.

8. Tools Used

- **Python Libraries**: Pandas, NumPy, Matplotlib, Seaborn, Plotly.
- **Jupyter Notebook**: For interactive data exploration and visualization.

9. References

- EV Market Reports and Registrations.
- Research papers on electric vehicle market adoption and CAFV eligibility.