Report on Electric Vehicle (EV) Population

Data Analysis

Name – Amrendra Kumar

Date – 7/01/2025

Linkedin: https://www.linkedin.com/in/amrendra-kumar-9954b9225/

Github: https://github.com/Amrendra-kumar7

Detailed Explanation: Introduction

1. The Role of Electric Vehicles (EVs)

Electric Vehicles (EVs) are pivotal in reducing greenhouse gas emissions, combatting climate change, and lessening the world's dependence on fossil fuels. Unlike traditional Internal Combustion Engine (ICE) vehicles, EVs rely on electricity, which can be generated from renewable sources such as solar, wind, and hydropower. This makes EVs an essential component of sustainable transportation systems.

2. Why Analyze EV Population Data?

Analyzing the population of EVs provides valuable insights into their adoption patterns, growth trajectories, and market penetration. Such analysis helps in understanding:

- Trends: How quickly EVs are being adopted globally or regionally.
- Barriers: Challenges that hinder EV adoption, such as lack of infrastructure or high costs.
- Success Factors: Elements that contribute to higher adoption rates, such as government incentives or advancements in battery technology.

3. Purpose of the Report

The report aims to go beyond just presenting raw data by:

- Identifying Patterns: Detecting key behaviors and shifts in the EV market.
- Highlighting Trends: Showing how the adoption of EVs has changed over time.
- Deriving Insights: Extracting actionable conclusions for policymakers, manufacturers, and other stakeholders.

4. Importance of Visual Representations

Visual representations like charts, graphs, and heatmaps are integral to the report. They:

- Simplify complex data, making it easier to understand and interpret.
- Highlight critical aspects such as growth rates, regional disparities, or leading manufacturers.
- Engage stakeholders by providing clear, visually appealing insights.

5. Statistical Interpretations

Statistical tools allow for:

- Measuring the growth rate of EV adoption.
- Comparing EV adoption across regions or time periods.

- Analyzing the distribution of different EV types, models, or manufacturers.

6. Supporting Stakeholder Decision-Making

The ultimate goal of this analysis is to empower stakeholders such as:

- Governments: To refine policies and incentives.
- Manufacturers: To optimize production and marketing strategies.
- Investors: To identify growth opportunities in the EV market.
- Consumers: To understand the environmental and economic benefits of EVs.

Data Overview

About the Data

This dataset provides comprehensive information on Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) registered with the Washington State Department of Licensing (DOL) as of April 2023. It serves as a valuable resource for analyzing the adoption and distribution of electric vehicles across the state.

Dataset link: https://catalog.data.gov/dataset/electric-vehicle-population-data

Objective

The primary objective of this analysis is to:

1. Identify the growth trends in EV populations.
2. Analyze geographical adoption patterns.
3. Study the distribution of EV types and manufacturers.
4. Assess the environmental impact of increased EV adoption.

Implementation Packages/Tools Used:

1**. Pandas:** Data manipulation and analysis library.

2. **Numpy:** Numerical computing library.

3**. Matplotlib:** Data visualization library.

4. **Seaborn:** Statistical data visualization library.

5. **Scipy:** To provide a comprehensive set of numerical algorithms and tools for scientific computing in Python.

6. **Shapely:** Shapely is a BSD-licensed Python package for manipulation and analysis of planar geometric objects.

Data-Preprocessing

Data Cleaning

The data collected is compact and is partly used for visualization purposes and partly for clustering. Python libraries such as NumPy, Pandas, Scikit-Learn, and SciPy are used for the workflow, and the results obtained are ensured to be reproducible.

```
#import some important libraries
# import some basic libraries
import pandas as pd
import numpy as np

# visualization libraries
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

# for scientific and statistical computing
import scipy as sp
```

```
# Display the data
df.head()
```

Python

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
3MW5P9J05N	Arapahoe	Englewood	CO	80111.0	2022	BMW	330E	Plug-in Hybrid Electric Vehicle (PHEV)	Not eligible due to low battery range	22	0	NaN	200589147	POINT (-104.89239 39.61914)	NaN
5YJXCBE27J	Island	Greenbank	WA	98253.0	2018	TESLA	MODEL X	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	238	0	10.0	326880081	POINT (-122.575569 48.08489)	NaN
1N4AZ0CP3F	Kings	Lemoore	CA	93245.0	2015	NISSAN	LEAF	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	84	0	NaN	182237457	POINT (-119.78637 36.30101)	NaN

Observation 1

Upon an initial review of the dataset, the following observations and considerations have been made regarding specific columns:

- **VIN (1-10):** Serves as a unique identifier for the dataset, and it will be used as the primary index.
- **Postal Code:** Currently stored as a float, this column should be converted to an integer format to ensure consistency and usability across the dataset.
- **Base MSRP:** Represents the Manufacturer's Suggested Retail Price (MSRP). While this is a key column, further investigation is needed to assess its data quality and completeness.

- **Model Year, Make, Model:** These attributes are derived from the VIN (1-10) identifier, and therefore, they may be redundant if the VIN is utilized effectively in the analysis.
- **Electric Utility:** This column describes the Electric Retail Services. The values are categorized as:
    - |: Single service provider from the same vendor.
    - ||: Multiple service providers from different vendors.
    - Blank: No data available.

Prior to renaming, we will inspect this column for NULL or missing values. The data can be grouped into the following categories:

- Single Type Utility
- Multi-Type Utility
- Not Available

- **Column Names with Spaces:** Several columns have spaces in their names, which may hinder analysis. These columns will be renamed for clarity and ease of use. The affected columns include:
    - Postal Code
    - Model Year
    - Electric Vehicle Type
    - Clean Alternative Fuel Vehicle (CAFV) Eligibility
    - Electric Range
    - Base MSRP
    - Legislative District
    - DOL Vehicle ID
    - Vehicle Location
    - Electric Utility
    - 2020 Census Tract
- **Potential Columns for Removal:** Based on the analysis requirements, the following columns may be considered for removal if found to be unnecessary:
    - Base MSRP
    - Legislative District

Descriptive Statistics of Dataset:

```
df.describe().style.background_gradient(cmap='Blues')
```

Python

	Postal Code	Model Year	Electric Range	Base MSRP	Legislative District	DOL Vehicle ID	2020 Census Tract
count	109480.000000	109481.000000	109481.000000	109481.000000	109205.000000	109481.000000	109480.000000
mean	98157.012943	2018.899197	90.211425	1849.194609	29.824120	198447551.859601	52968493402.103012
std	2640.605503	2.872853	102.575715	10946.085012	14.679959	95617186.541589	1675103687.859588
min	1730.000000	1997.000000	0.000000	0.000000	1.000000	4777.000000	1101001400.000000
25%	98052.000000	2017.000000	0.000000	0.000000	18.000000	146731322.000000	53033008500.000000
50%	98121.000000	2019.000000	35.000000	0.000000	34.000000	187411808.000000	53033029304.000000
75%	98370.000000	2021.000000	208.000000	0.000000	43.000000	216917571.000000	53053072506.000000
max	99701.000000	2023.000000	337.000000	845000.000000	49.000000	479254772.000000	56033000100.000000

Exploratory Data Analysis (EDA) and Visualization

Exploratory Data Analysis (EDA) is a critical step in the data analysis process, focusing on analyzing and summarizing the main characteristics of the dataset, often using visual methods. It helps to uncover underlying patterns, detect anomalies, test hypotheses, and check assumptions, all of which are vital before any modeling or predictive analysis. EDA provides insights into the structure and quality of the data, guiding the necessary steps for data wrangling, including cleaning and transformation.

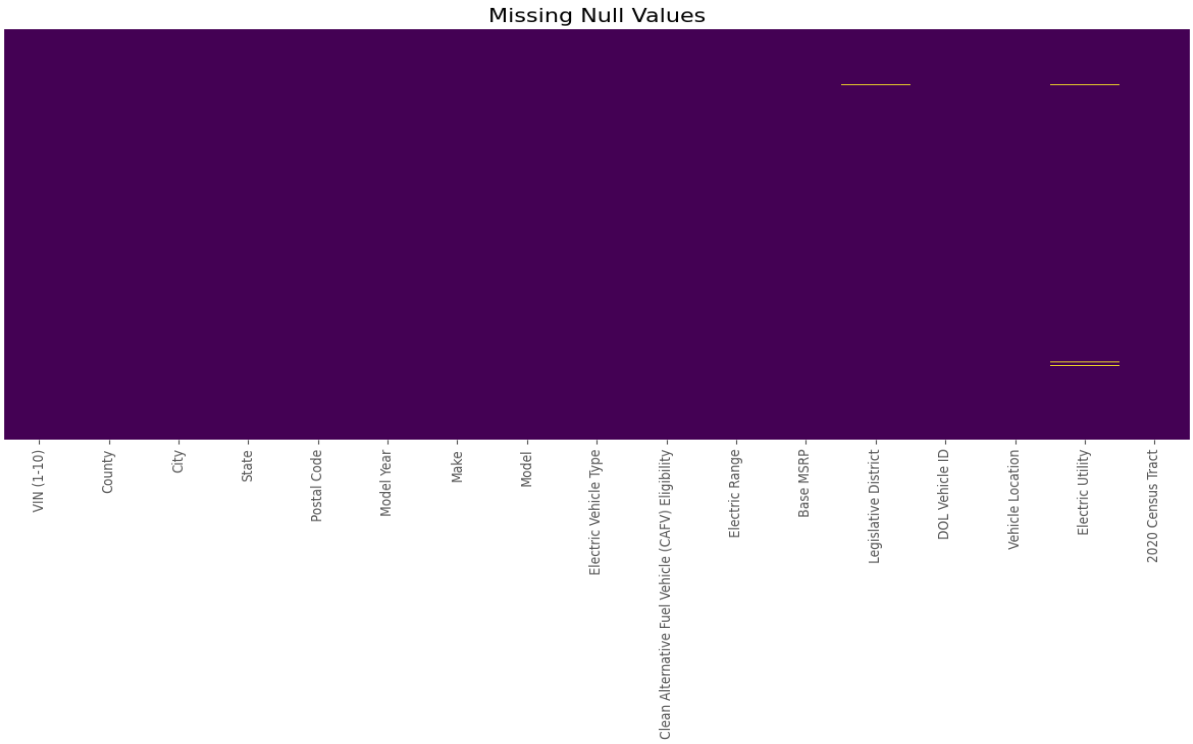
In this project, EDA will be used to explore the dataset, identify key trends, and gain a deeper understanding of the relationships between different variables. We will perform several key tasks during the EDA process:

1. **Summary Statistics:** This involves generating basic statistics like mean, median, mode, minimum, maximum, and standard deviation for numerical variables to get an overview of their distributions and identify outliers.
2. **Missing Data Analysis:** Checking for missing, null, or inconsistent values in the dataset is crucial for proper data handling. This will help us decide whether to impute, remove, or flag missing data.
3. **Data Distribution:** We will visualize the distribution of key variables (e.g., base MSRP, electric range, model year) using histograms or boxplots to understand the spread and identify any skewness or outliers.
4. **Correlation Analysis:** Understanding the relationships between different variables is essential for identifying potential patterns or trends. We will create correlation matrices and scatter plots to explore how variables like MSRP, electric range, and vehicle type are interrelated.

5. **Categorical Data Analysis:** For categorical variables like electric utility and vehicle make, we will use bar charts or pie charts to visualize the frequency distribution of each category.
6. **Visualizations:**
    - **Histograms and Box Plots:** To inspect the distribution and identify outliers in numerical features.
    - **Scatter Plots:** To visualize the relationship between two continuous variables.
    - **Heatmaps:** To display correlation matrices and the strength of relationships between variables.
    - **Bar Charts/Pie Charts:** To analyze the frequency of categorical variables such as electric utility and vehicle type.

Visualization plays an important role in EDA as it allows us to quickly understand the underlying patterns, distributions, and relationships within the dataset. Effective visualizations will be integral to presenting our findings clearly and concisely in the report, enabling easier interpretation of complex data patterns.

Missing Null Values

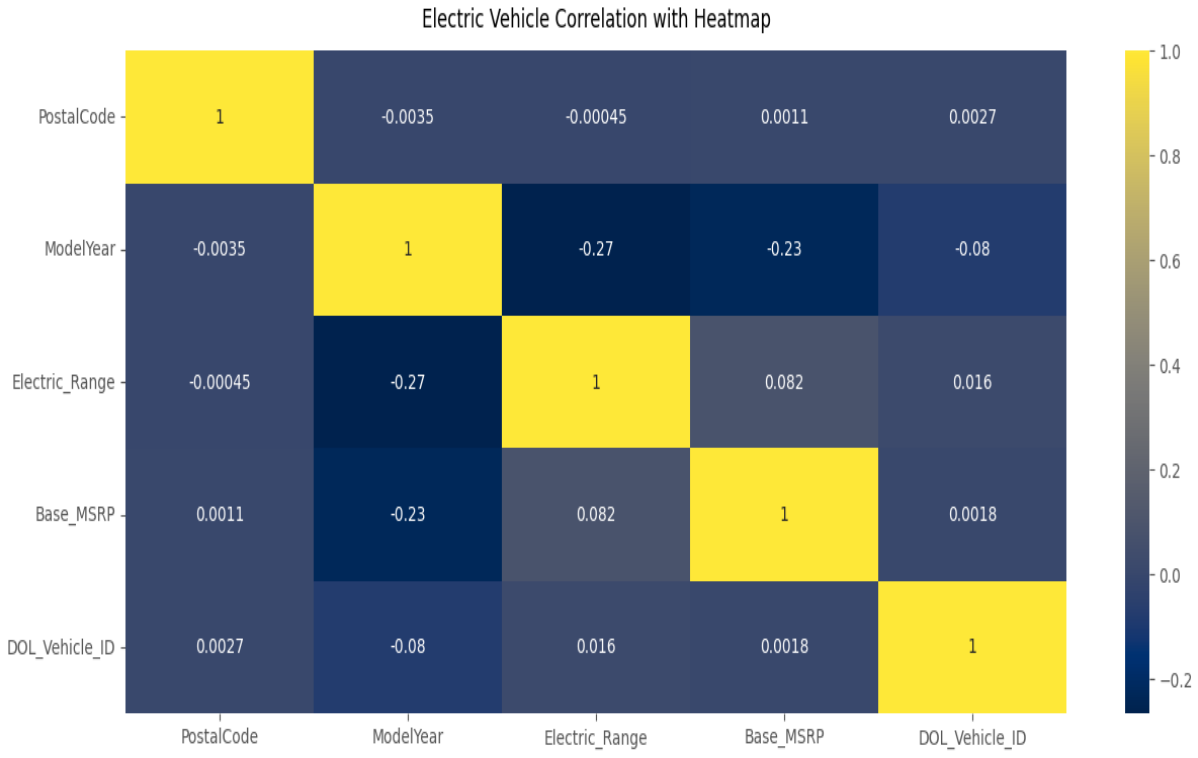


Handling Missing Values: To ensure data consistency and completeness, the following strategies will be employed for handling missing (null) values in the dataset:

- **Electric Utility:** Missing values will be replaced with the text "Utility Not Available" to indicate the absence of data related to electric utility services.
- **Legislative District:** Missing values will be populated with the text "Unknown" to represent unknown or missing district information.
- **Vehicle Location:** Missing values will be replaced with the text "Unknown" to signify the absence of vehicle location data.
- **Model:** Missing values will be filled with the text "Unknown" to account for vehicles with no model data available.
- **2020 Census Tract:** Missing values will be filled with the text "Unknown" to handle cases where census tract information is missing.
- **City:** Missing city values will be populated with the text "Unknown" to maintain consistency in the city field.
- **Postal Code:** Missing postal code values will be filled with the mean value of the existing postal codes to provide a reasonable estimate for the missing data.
- **County:** Missing values will be replaced with the text "Unknown" to handle cases where county information is not available.

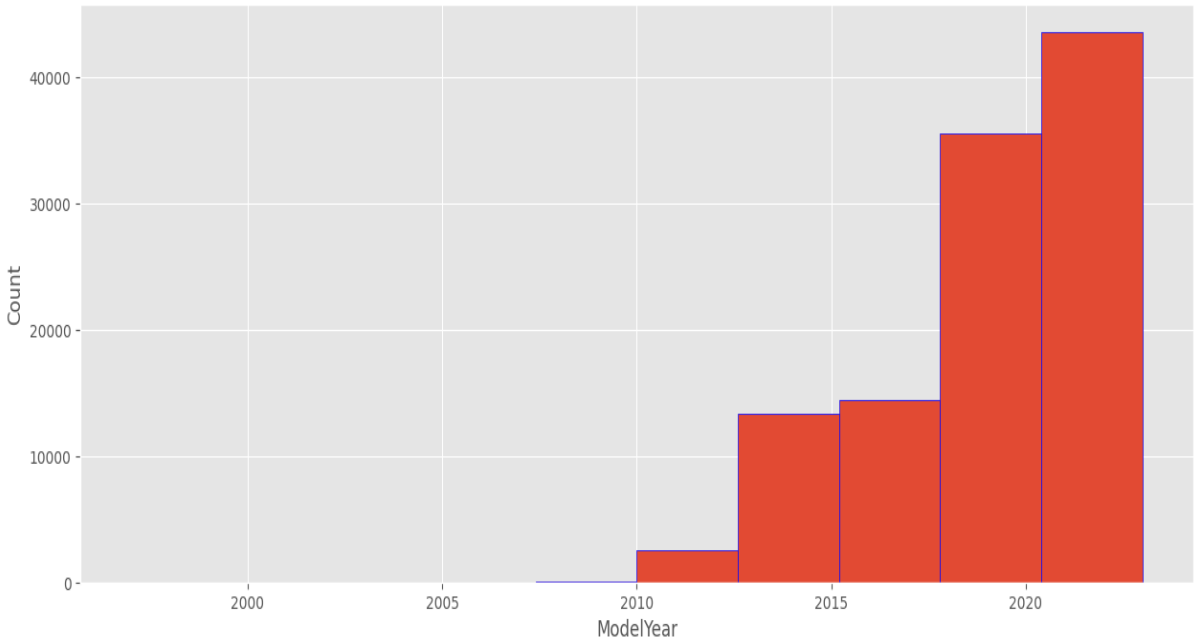
These strategies will ensure that the dataset is complete and ready for further analysis, while maintaining the integrity of the information.

Electric Vehicle Correlation with Heatmap

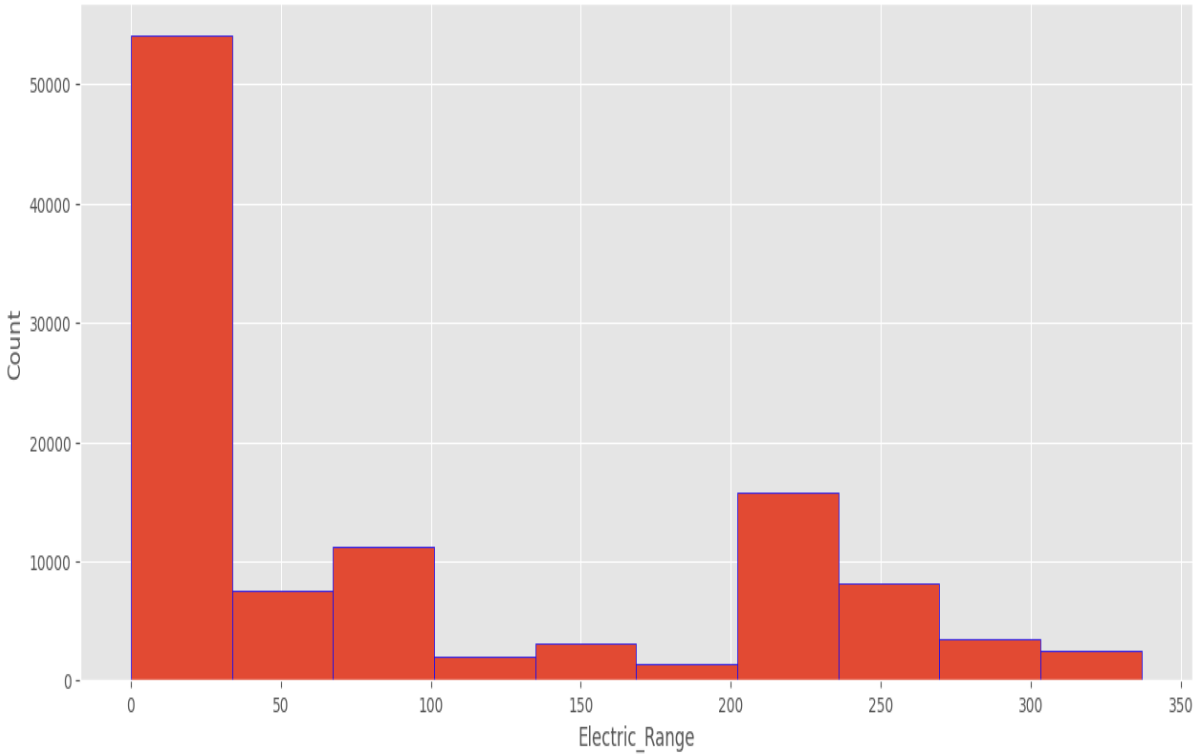


Distribution of numerical variables:

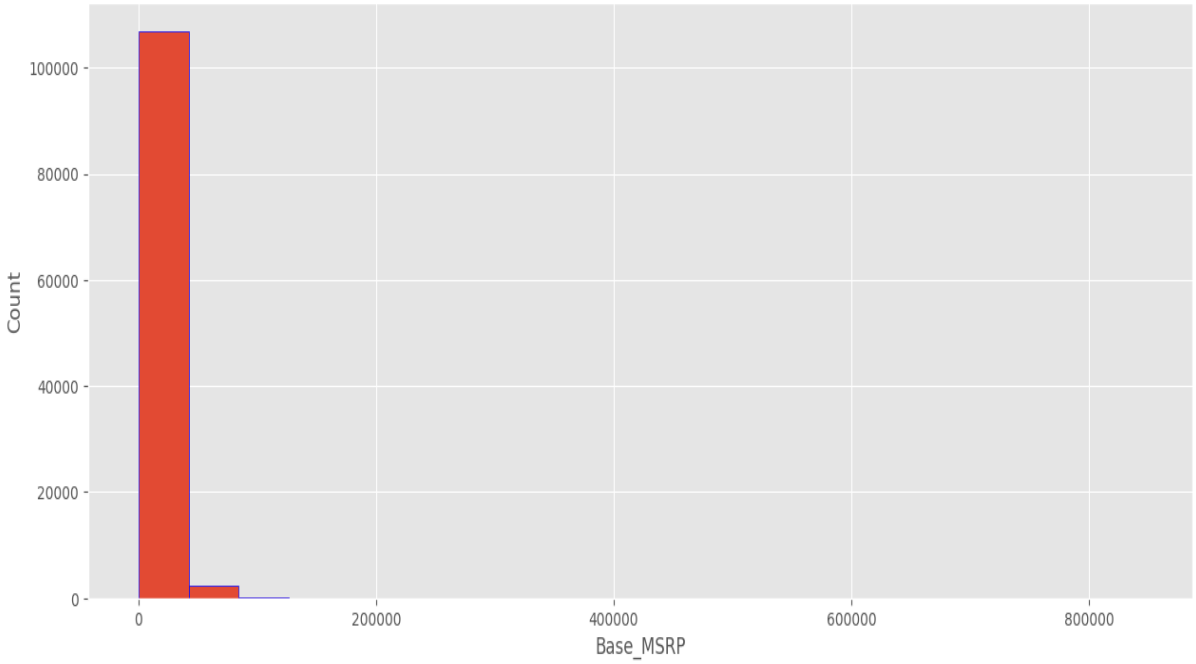
ModelYear, Electric_Range, Base_MSRP, DOL_Vehicle_ID



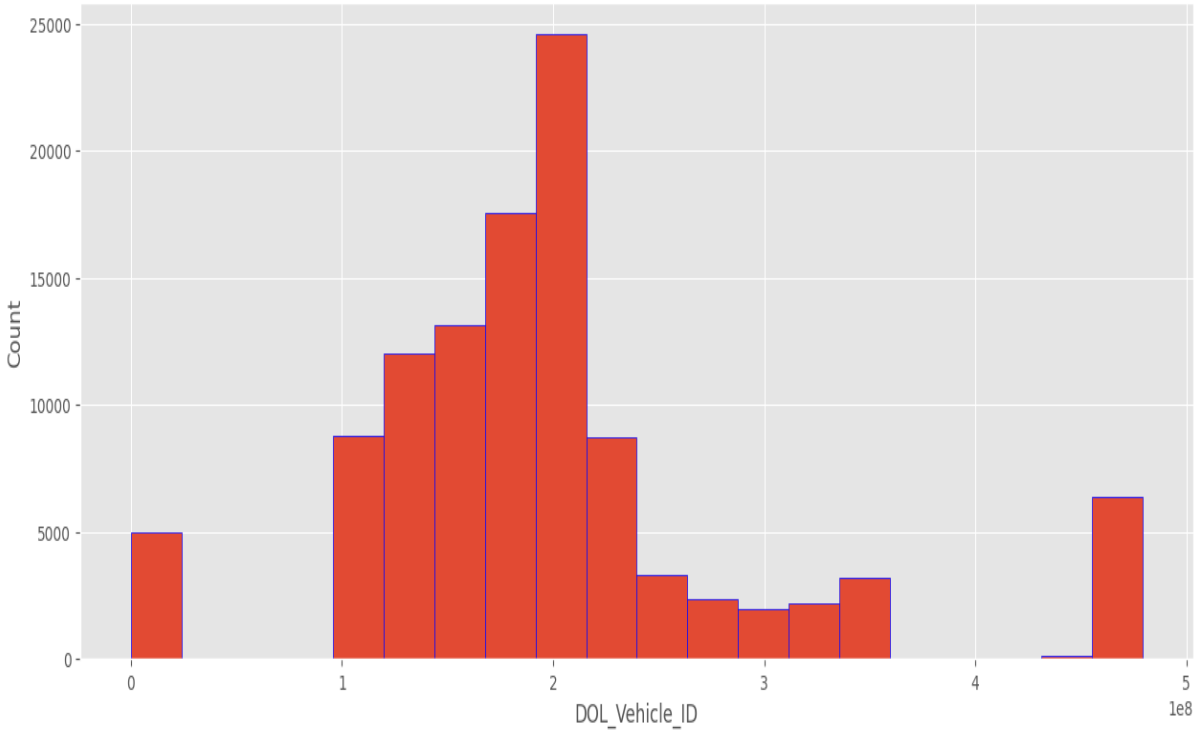
Histogram to check the normality and distribution of Model Year attribute:



Histogram to check the normality and distribution of Electric Range of the Cars in one charge:



Histogram to check the normality and distribution of Base MRP:



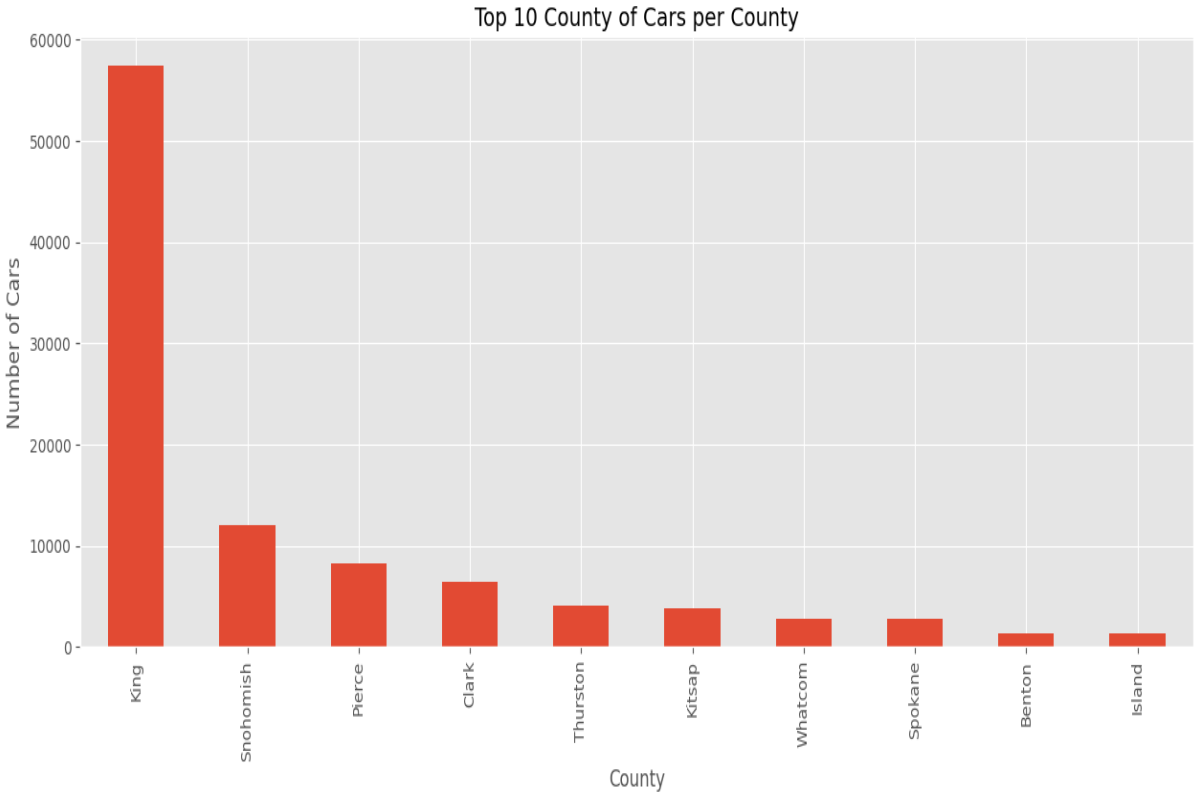
Observation 2

Upon analyzing the histogram plots, the following observations have been made:

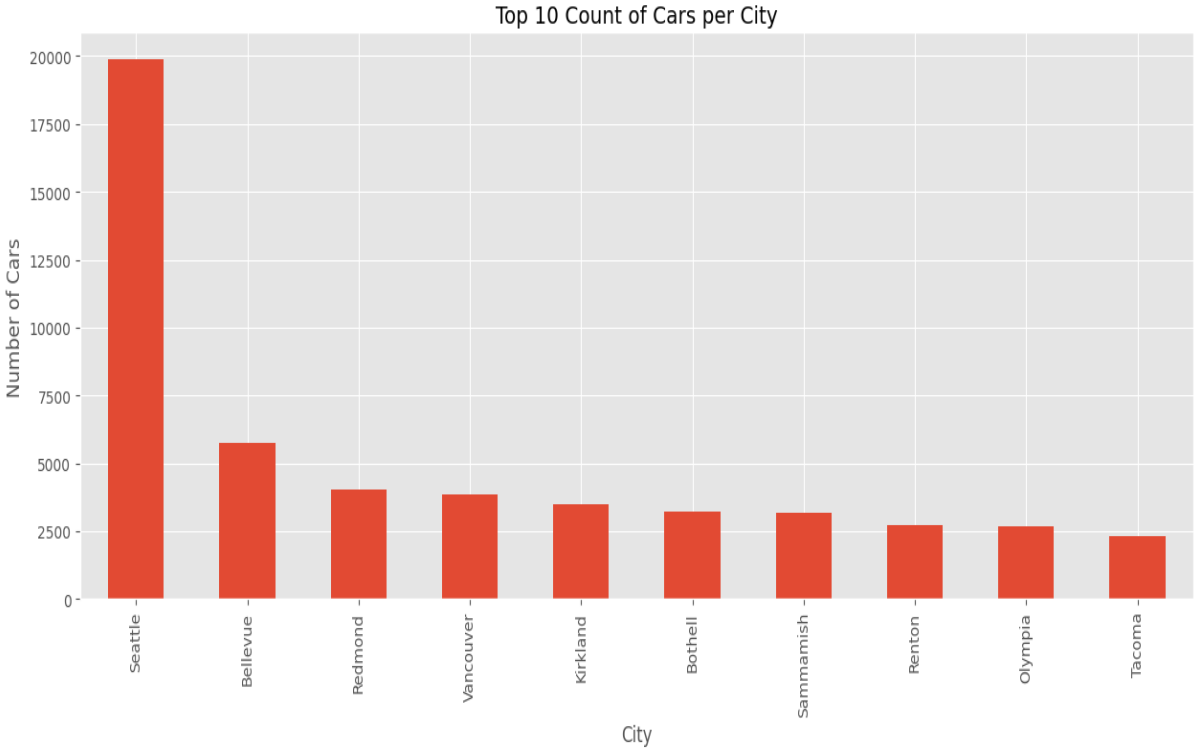
- **Base MSRP:** Despite being a key column, it is evident that the Base MSRP data is largely missing, as the information is not provided by the vendors. Therefore, this column will be dropped from the dataset due to its lack of completeness and relevance for the analysis.
- **Electric Range:** It was observed that approximately 50,000 vehicle records have invalid data in the Electric Range column, with the values incorrectly filled as zero. This represents missing or erroneous information, and further action will be taken to address this issue in the data cleaning process.

Some Important Visualization

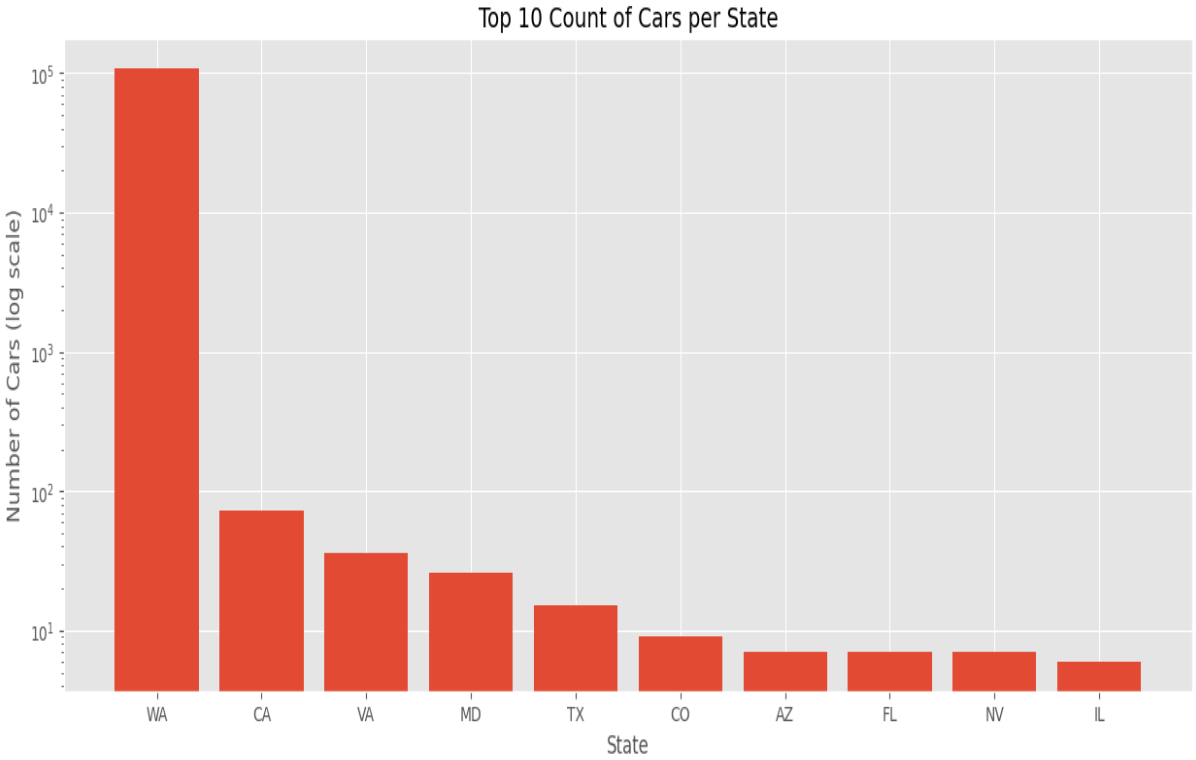
Top 10 County of Cars per County:



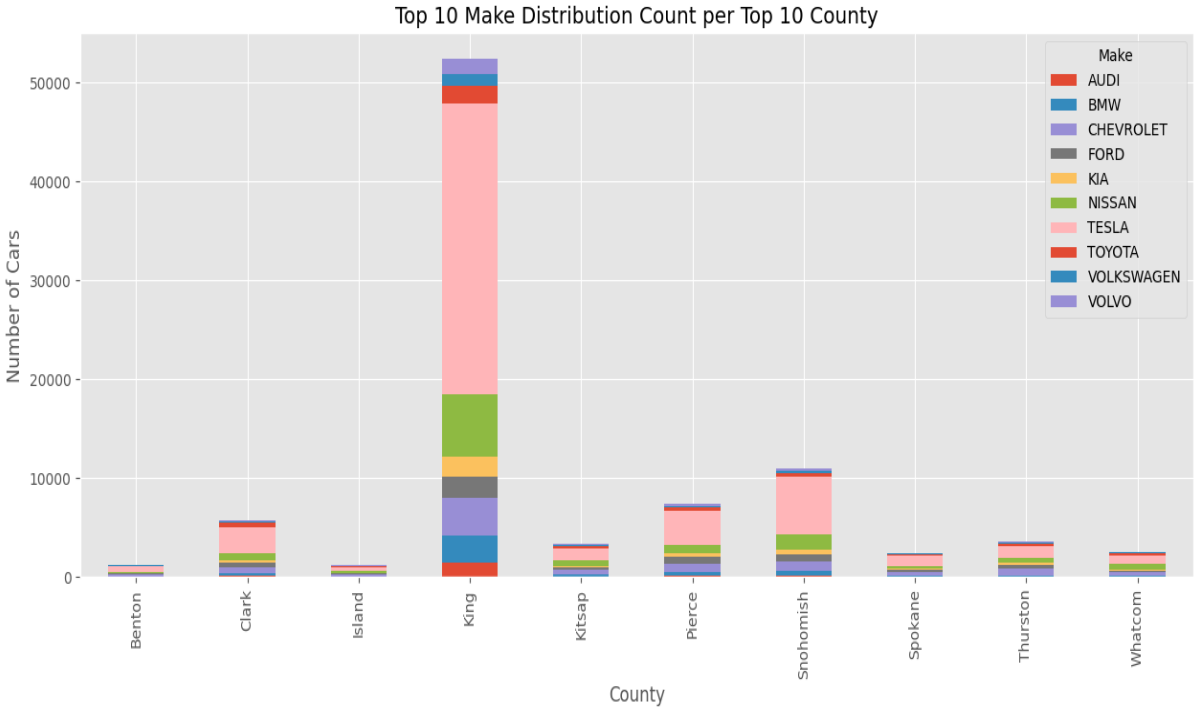
Top 10 Count of Cars per City:



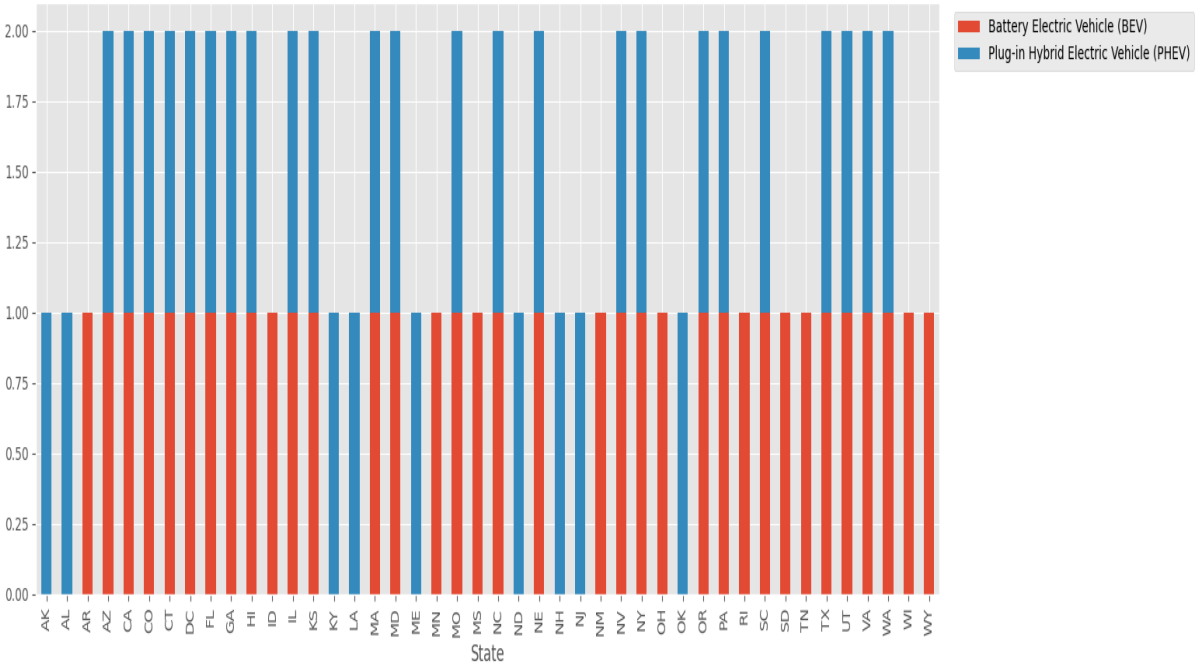
Top 10 Count of Cars per State:



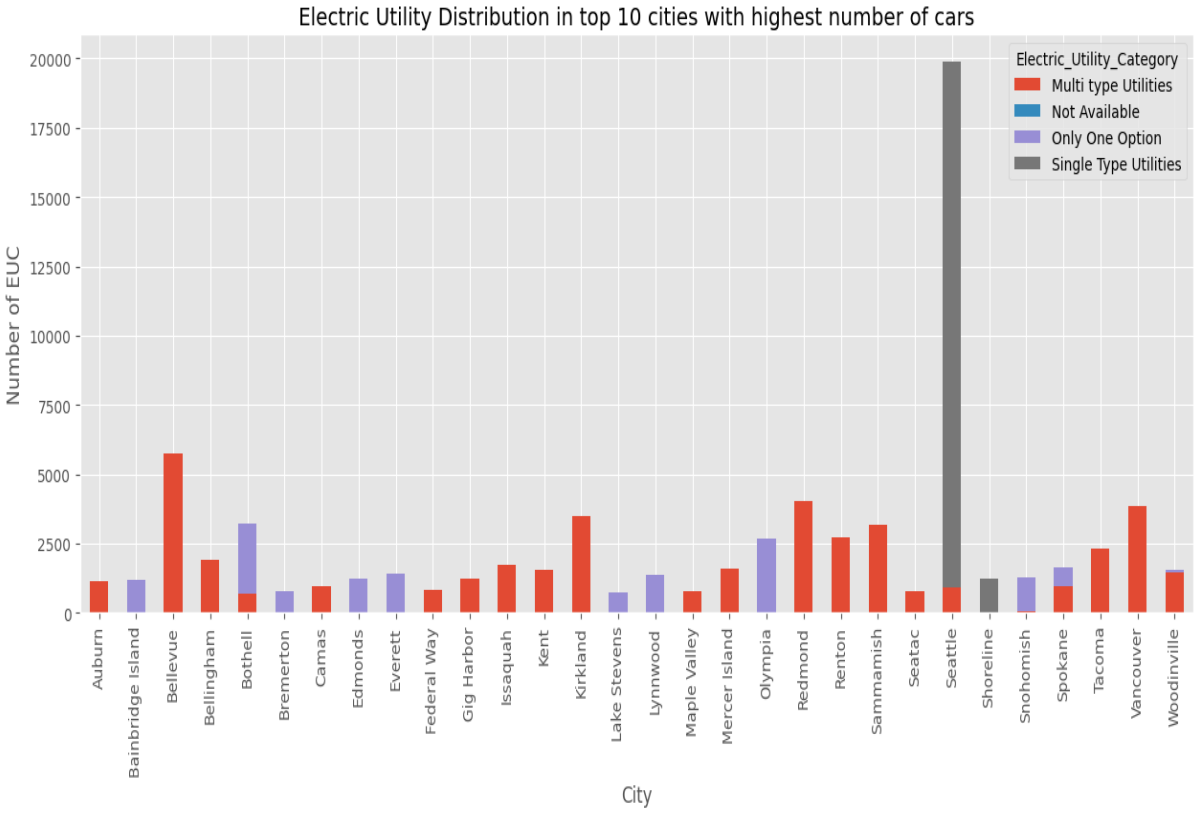
Top 10 Make Distribution Count per Top 10 County:



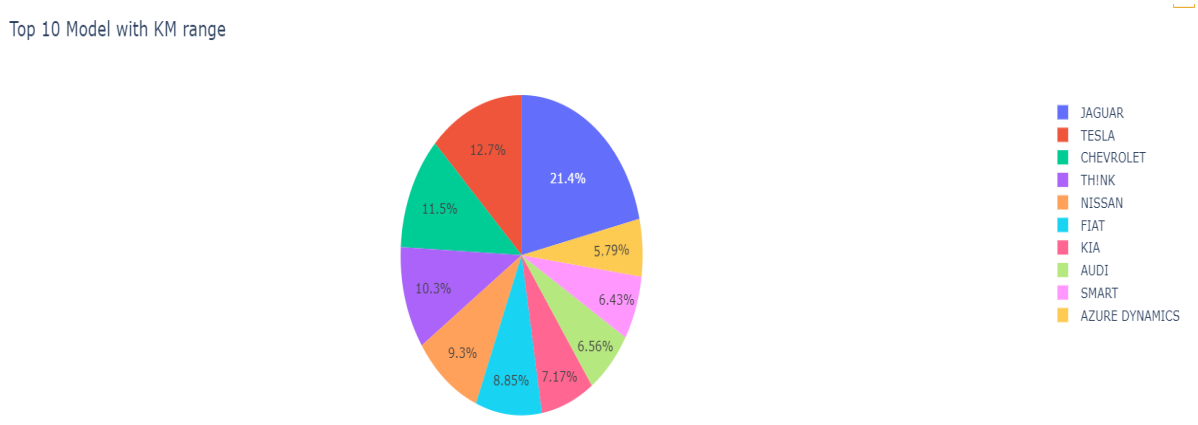
EVT Distribution count per state:



Electric Utility Distribution in top 10 cities with highest number of cars:



Now visualize the Distances Travel by vehicle make per electric charge:



Summary & Conclusion

The Electric Vehicle Population dataset offers detailed information on Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) registered with the Washington State Department of Licensing (DOL). Through exploratory data analysis, key insights were gained, including the top 10 vehicle counts by county, city, state, and postal code. King County emerged as the leading county in terms of vehicle registrations, followed by Snohomish and Pierce counties. Seattle topped the list of cities, with Bellevue and Redmond following closely behind. Among states, Washington had the highest number of registered electric vehicles, with California and Virginia ranking next.

The analysis also highlighted the top 10 car manufacturers by county, city, and state, with Tesla standing out as the most popular brand overall. This provides valuable insights into market dynamics, suggesting that car manufacturers like Audi and BMW could explore untapped markets in other states. Furthermore, the top 10 postal codes were identified, offering additional opportunities for targeted marketing and potential upselling strategies.

References

1. Global EV Outlook 2024
2. EV Adoption Statistics
3. Google for images