BIRZEIT UNIVERSITY

Faculty of Engineering & Technology

Electrical & Computer Engineering Department

MACHINE LEARNING AND DATA SCIENCE

ENCS5341

Assignment #1 Report

Prepared by:

Student Name: Razi Atyani **Student ID:** 1200028

Student Name: Abdalkarim Eiss **Student ID:** 1200015

Instructor: Dr. Yazan Abu Farha && Dr. Ismail Khater

Section: 1 & 2

Date: 10/29/2024

1 Abstract

This report analyzes the "Electric Vehicle Population Data" from Washington State, focusing on data preprocessing, exploratory data analysis (EDA), and insight generation. The dataset includes information on battery electric vehicles (BEVs) and plug-in hybrid electric vehicles (PHEVs), covering attributes like make, model, year, type, electric range, and registration locations. Key steps include handling missing values, encoding categorical variables, and normalizing numerical features to prepare the data for analysis. Descriptive statistics, spatial visualizations, and temporal trends highlight patterns in EV adoption and model popularity across cities and counties. The findings offer valuable insights into the growth of EVs, supporting data-driven decisions for sustainable transportation initiatives.

Table of Contents

List of Figures

2 Libraries

Figure 1 shows the libraries that are used in the assignment code.

Libraries:

```
[ ] import pandas as pd
    import seaborn as sns
    import numpy as np
    import matplotlib.pyplot as plt
    from sklearn.preprocessing import OneHotEncoder
    from sklearn.preprocessing import MinMaxScaler
    from sklearn.preprocessing import LabelEncoder
    from scipy.stats import zscore
    import geopandas as gpd
    from shapely import wkt
    import folium
```

Figure 1: All Libraries

- **pandas**: Used for data manipulation and analysis, such as reading datasets and handling dataframes.
- **seaborn**: A statistical data visualization library that helps create attractive and informative plots.
- **numpy**: Provides support for numerical operations, including arrays and mathematical functions.
- **matplotlib.pyplot**: A plotting library for creating static, interactive, and animated visualizations.
- **OneHotEncoder (sklearn)**: Converts categorical variables into a binary (one-hot) format for model compatibility.
- **MinMaxScaler (sklearn)**: Normalizes numerical data to a fixed range (e.g., 0 to 1) for better model performance.
- **LabelEncoder (sklearn)**: Encodes categorical labels into numeric values for model processing.

- **zscore (scipy)**: Computes the z-score to standardize data by measuring how many standard deviations a value is from the mean.
- **geopandas**: Extends pandas to handle geospatial data, useful for analyzing and visualizing location-based data.
- **shapely**: Facilitates the manipulation of geometric objects for spatial operations.
- **folium**: Helps create interactive maps, especially useful for visualizing geospatial data.

3 Data Cleaning

3.1 Dataset Reading

Figure 2 shows the code of dataset reading and its results.

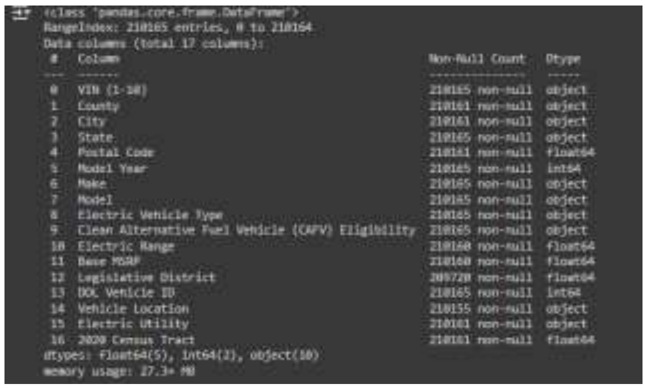
Figure 2: Dataset Reading Results

As noted, here we explore the data and the columns inside it and know each column what is doing here, the csv dataset is read and its results are displayed properly. It has 17 columns some of them are categorical and the others are numerical. Figure 3 shows all columns and the dataset shape that have 210165 rows and 17 columns before cleaning.

The dataset contains information about electric vehicles registered in Washington State, with 17 columns covering various attributes such as:

- VIN (Vehicle Identification Number) – A unique identifier for each vehicle.
- County, City, and State – Registration locations.
- Postal Code – Location postal codes.
- Model Year, Make, and Model – Details about the vehicle's brand and year.
- Electric Vehicle Type – Classification of the vehicle (e.g., Battery Electric Vehicle (BEV) or Plug-in Hybrid (PHEV)).
- CAFV Eligibility – Indicates whether the vehicle qualifies for alternative fuel benefits.

- Electric Range – Specifies the range on electric power.
- Base MSRP – The base manufacturer's suggested retail price.
- Legislative District and DOL Vehicle ID – Identifiers used by the Department of Licensing and for legislative purposes.
- Vehicle Location (Latitude, Longitude) – Coordinates showing where the vehicle is registered.
- Electric Utility – The energy provider for the registered location.
- 2020 Census Tract – A geographical region defined for census purposes.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210165 entries, 0 to 210164
Data columns (total 17 columns):
 #   Column                                              Non-Null Count   Dtype
---  ------                                              --------------   -----
 0   VIN (1-10)                                          210165 non-null  object
 1   County                                              210161 non-null  object
 2   City                                                210161 non-null  object
 3   State                                               210165 non-null  object
 4   Postal Code                                         210161 non-null  float64
 5   Model Year                                          210165 non-null  int64
 6   Make                                                210165 non-null  object
 7   Model                                               210165 non-null  object
 8   Electric Vehicle Type                               210165 non-null  object
 9   Clean Alternative Fuel Vehicle (CAFV) Eligibility   210165 non-null  object
 10  Electric Range                                      210160 non-null  float64
 11  Base MSRP                                           210160 non-null  float64
 12  Legislative District                                209720 non-null  float64
 13  DOL Vehicle ID                                      210165 non-null  int64
 14  Vehicle Location                                    210155 non-null  object
 15  Electric Utility                                    210161 non-null  object
 16  2020 Census Tract                                   210161 non-null  float64
dtypes: float64(5), int64(2), object(10)
memory usage: 27.3+ MB
```

Figure 3: The Information and the Shape of Dataset

3.2 Document Missing Values

Figure 4 shows all columns (features) with its number of null values.

Figure 4: The Sum of Missing Values in Each Column

As noted, here we calculate the sum of null values in the data set and now we are going to deal with them, there are small null values like (4, 5, 10). Also, the Legislative Distinct feature with largest number of null values.

So, the smallest number of null values can be ignored and dropped.

Also, the percent of null values in each feature and the null values in each record are found. About 456 rows that are about 21% of all dataset are found. You can find it in the notebook.

Also, the duplicated rows existence is checked and there is no any duplicated row, all values are zeros. Moreover, the empty records existence is checked and there is no any empty record.

Figure 5 shows all columns with its number of missing values after the smallest number of null values dropped.

```
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                               441
DOL Vehicle ID                                       0
Vehicle Location                                     0
Electric Utility                                     0
2020 Census Tract                                    0
dtype: int64
```

Figure 5: The Sum of Missing Values in Each Column After the Smallest Number of Null Values Dropped

3.3 Missing Values Handling

Figure 6 shows the code and results of missing values handling methods.

This analysis compares two strategies for handling missing values in the "Legislative District" column: dropping rows with missing data and using mean or median imputation. Dropping rows reduced the dataset size but preserved the variability in the column, maintaining consistent overall data distribution. In contrast, both mean and median imputation retained all rows, but the imputed column lost variability since missing values were replaced with a single value, either the mean or median. Descriptive statistics for other columns, such as "Electric Range" and "Base MSRP," remained stable across all methods. The decision between dropping rows or using imputation depends on whether preserving the full dataset or the variability in the "Legislative District" column is more important for the analysis.

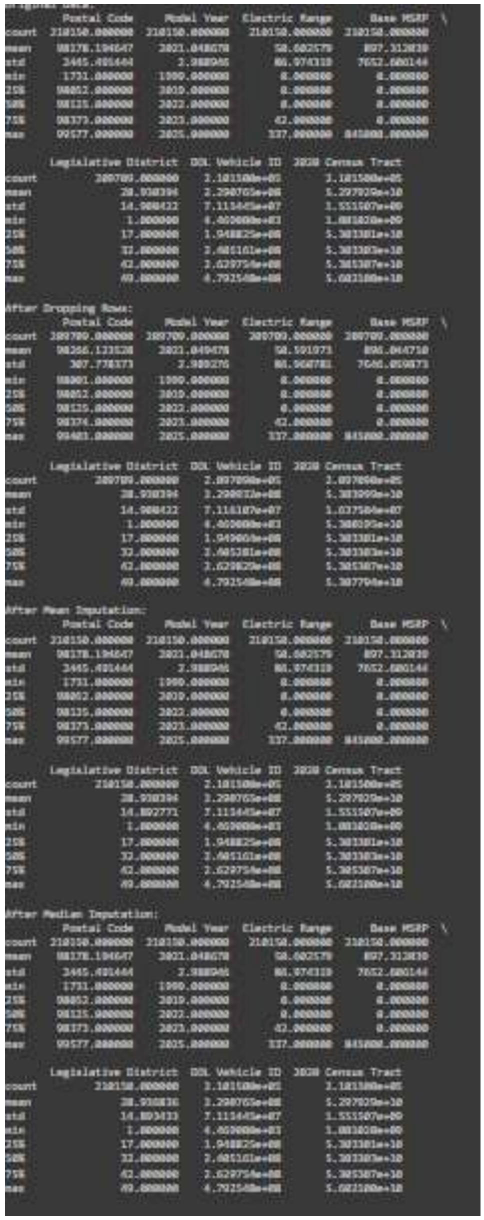


Figure 6: The Results of Missing Values Handling Methods

3.4 Data After Cleaning

Figure 7 shows the description of the dataset after cleaning.

	Postal Code	Model Year	Electric Range	Base MSRP	Legislative District	DOL Vehicle ID	2020 Census Tract
count	209709.000000	209709.000000	209709.000000	209709.000000	209709.000000	2.097090e+05	2.097090e+05
mean	98266.123528	2021.049478	50.591973	896.044710	28.930394	2.290932e+08	5.303999e+10
std	307.778373	2.989276	86.960781	7646.059873	14.908422	7.116107e+07	1.637504e+07
min	98001.000000	1999.000000	0.000000	0.000000	1.000000	4.469000e+03	5.300195e+10
25%	98052.000000	2019.000000	0.000000	0.000000	17.000000	1.949064e+08	5.303301e+10
50%	98125.000000	2022.000000	0.000000	0.000000	32.000000	2.405261e+08	5.303303e+10
75%	98374.000000	2023.000000	42.000000	0.000000	42.000000	2.629829e+08	5.305307e+10
max	99403.000000	2025.000000	337.000000	845000.000000	49.000000	4.792548e+08	5.307794e+10

Figure 7: The Description of the Dataset After Cleaning

After cleaning we have seen the outliers but we have not deal with them. You can see that we made the boxplots for each feature in the notebook.

4 Feature Engineering

4.1 Categorical To Numerical Conversion

Figure 8 shows the conversion of categorical columns to numerical column. As noted, the one hot encoding is used.as it changes each category to a feature and each occurrence to put 1 at it and zero for others It increases the number of columns based on the categorical columns. So, it makes the data big.

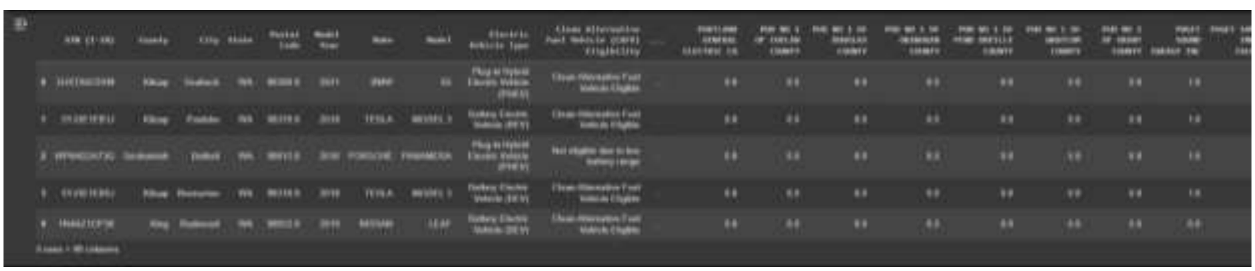


Figure 8: The OneHotEncoding Results

4.2 The Normalization

Figure 8 shows the code and results of using the min-max method. Min-Max Scaling transforms data to a fixed range, typically 0 to 1, by subtracting the minimum value and dividing by the range. Z-score Standardization rescales data by subtracting the mean and dividing by the standard deviation, ensuring the data has a mean of 0 and a standard deviation of 1, which is useful for normalizing features with different units or scales.

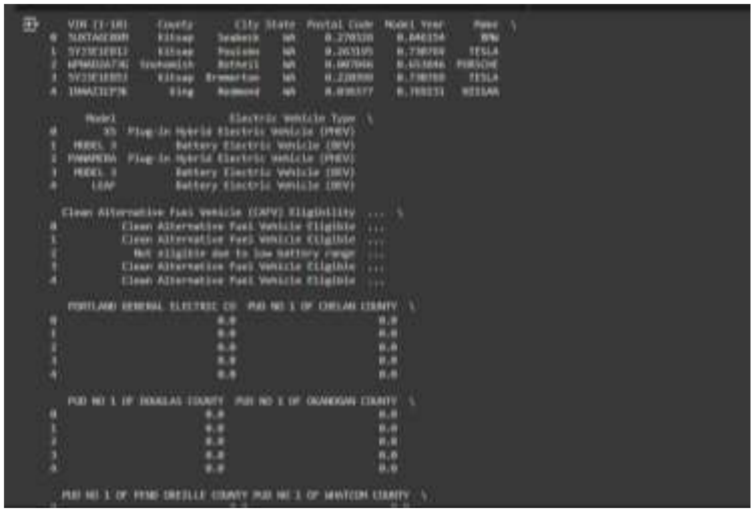


Figure 9: Min-Max Method Code and Results

Also, the Z-score method is used. You can find it in the notebook.

5 EDA

5.1 Descriptive Statistics

Figure 10 shows the descriptive statistics of the dataset. We here calculate the mean and median and standard deviation for each feature.

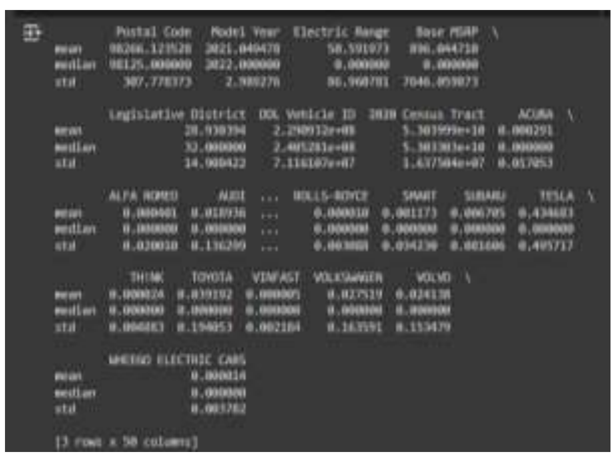


Figure 10: Dataset Descriptive Statistics

5.2 Spatial Distribution

Figure 11 shows the map for a spatial distribution of EVs across locations.

Figure 11: Spatial Distribution

As noted, we create an interactive map using Folium to display the locations of electric vehicles as blue circle markers based on their geographical coordinates.

5.3 Model Popularity

Figure 12 shows the model popularity based on the cleaned dataset. As shown, the model Y (Tesla) is the most trend vehicle used according to the dataset.

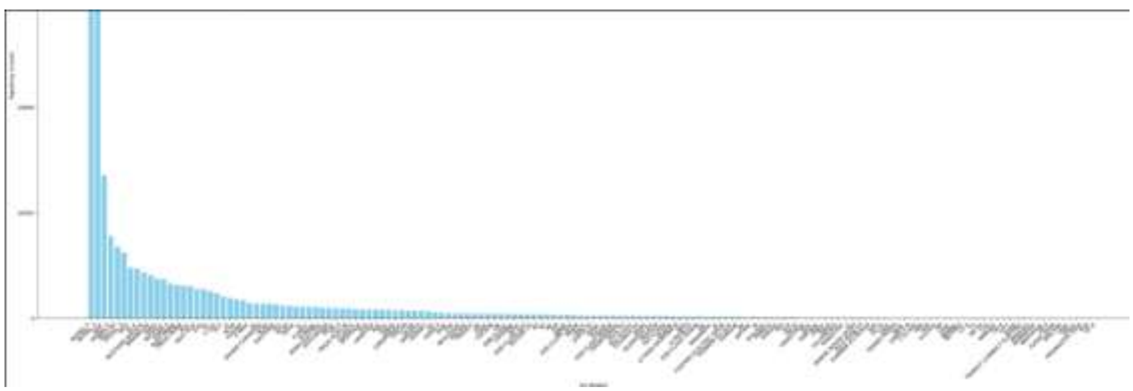


Figure 12: Model Popularity

5.4 Correlation Between the Numeric Features

Figure 12 shows the relationship between every pair of numeric features.

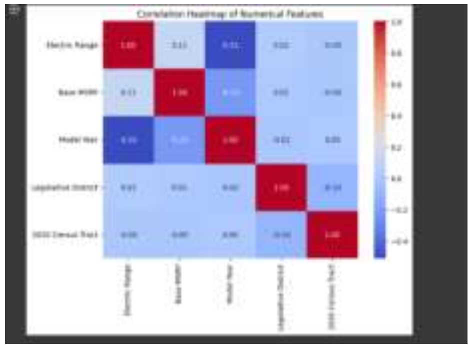


Figure 13: The Correlation Between the Numeric Features

As noted, if the value of correlation is positive and high like Base MSRP with Electric Range, there is a linear correlation between them. But if the value of correlation is negative like Model Year with Electric Range, there is a negative linear correlation between them. Also, if the value of correlation is zero, there is no correlation.

6 Visualization

6.1 Data Exploration Visualizations

Figure 14 shows the distribution of electric range, distribution of Base MSRP and distribution of model year.

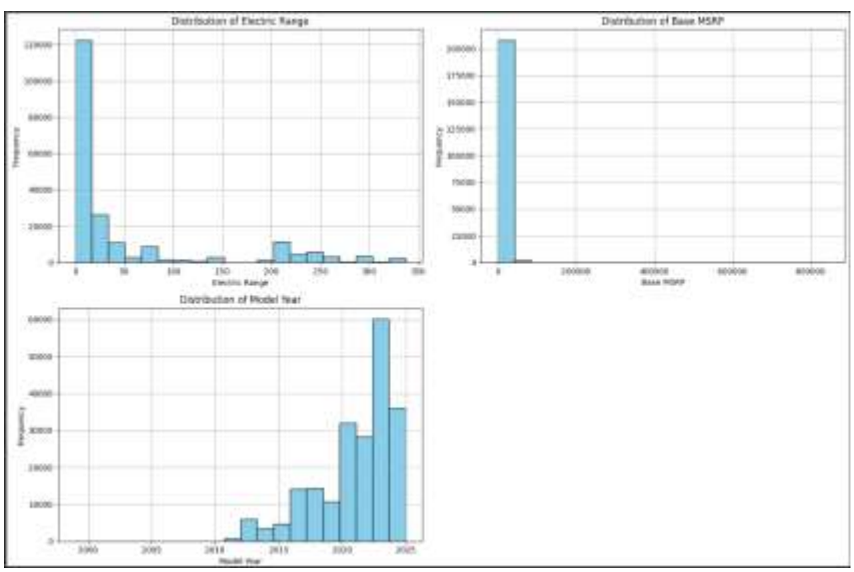


Figure 14: Histograms

The three histograms as following:

- **Distribution of Electric Range**: This histogram shows the frequency of electric vehicle models across different electric range categories. The x-axis represents the electric range in miles, and the y-axis indicates the number of models in each range. The distribution appears to be right-skewed, with a peak around 250 miles.

- **Distribution of Base MSRP**: This histogram displays the frequency of electric vehicle models across different base MSRP (Manufacturer's Suggested Retail Price) categories. The x-axis shows the base MSRP in dollars, and the y-axis represents the number of models in each price range. The distribution is also right-skewed, with the majority of models priced below $80,000.

- **Distribution of Model Year**: This histogram shows the frequency of electric vehicle models across different model years. The x-axis displays the model year, and the y-axis represents the number of models in each year. The distribution appears to be bimodal, with peaks around 2005 and 2020. This suggests a potential increase in electric vehicle production and popularity in recent years.

Figure 15 shows two scatter plots. The left plot compares Base MSRP with Electric Range, where most EVs with higher ranges have lower MSRPs. The right plot compares Base MSRP with Model Year, showing most recent Evs (post-2010) have lower MSRPs, with one outlier around $800,000. Both plots highlight Evs clustering at lower prices.

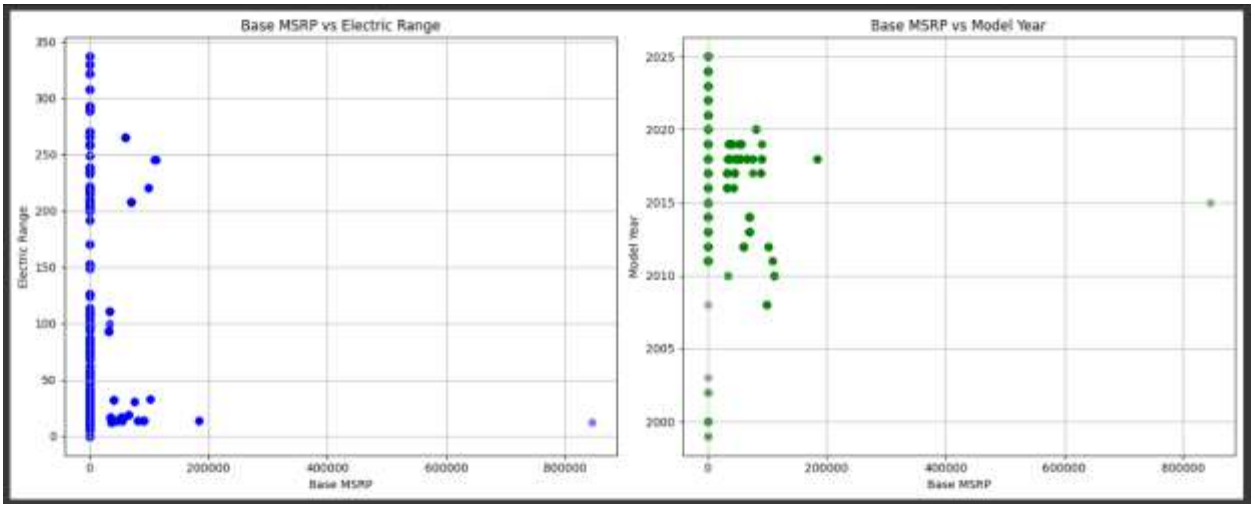


Figure 15: The Scatter Plots

Figure 16 shows a boxplot that contains the Base MSRP Distribution Across Different Makes.

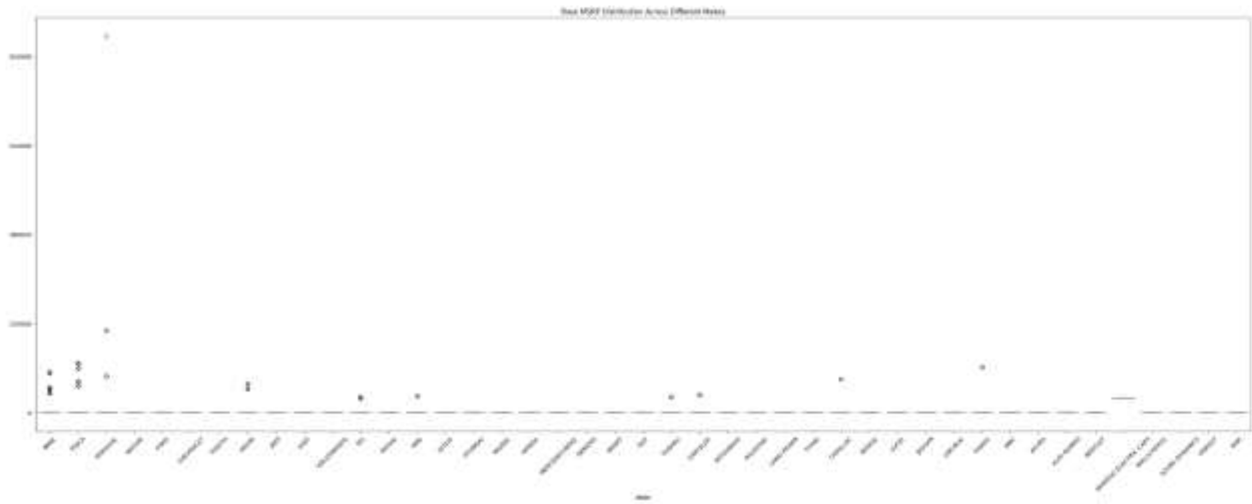


Figure 16: The Boxplots

6.2 Comparative Visualization

Figure 17 shows the vehicle count for the top 10 cities, displayed in descending order. The y-axis represents the cities, while the x-axis shows the corresponding number of vehicles.

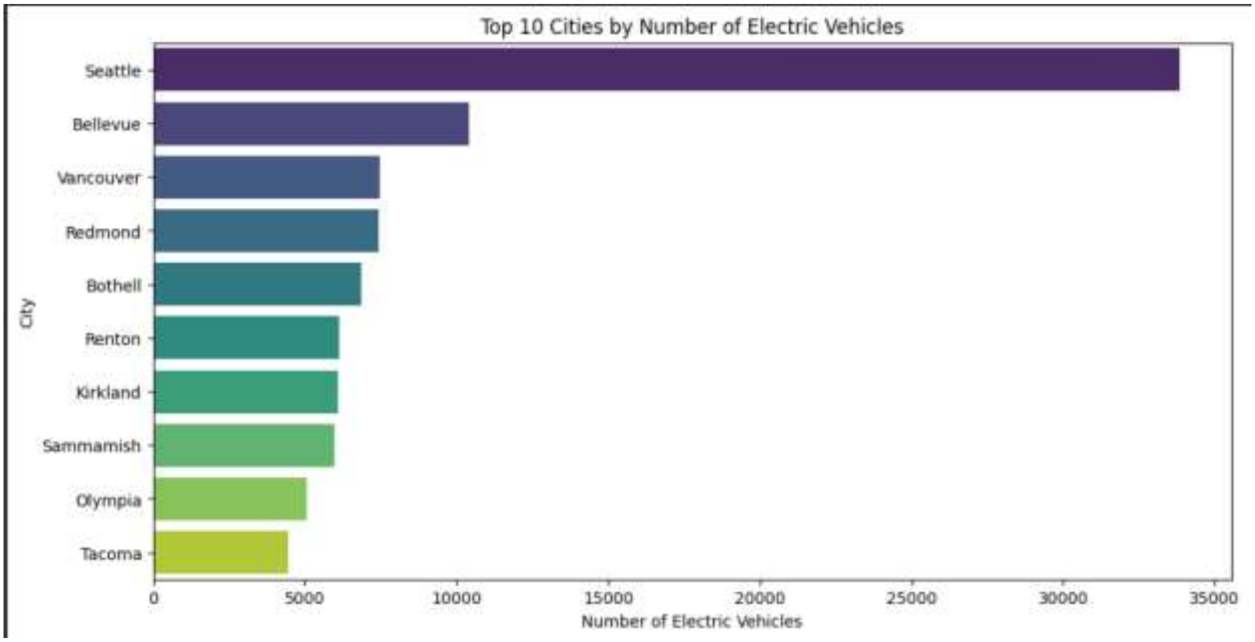


Figure 17: The Comparative of the Distribution of EVs Across Different Cities

Figure 17 shows the vehicle count for the top 10 counties, displayed in descending order. The y-axis represents the counties, while the x-axis shows the corresponding number of vehicles.

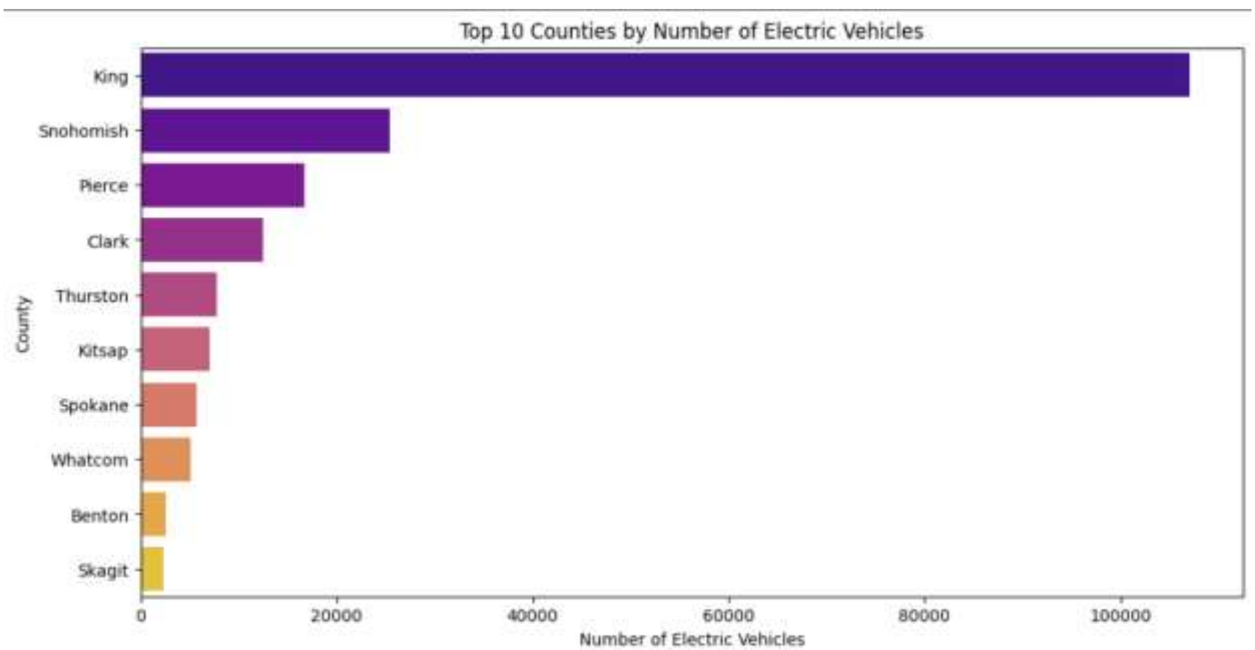


Figure 18: The Comparative of the Distribution of EVs Across Different Counties

6.3 Temporal Analysis

Figure 19 depicts the trend of electric vehicle (EV) registrations per year, with the x-axis representing model years from 2000 to 2025 and the y-axis showing the number of registrations. The graph shows a relatively flat trend in registrations from 2000 to about 2010. Afterward, there is a noticeable increase, particularly from 2015 onwards, with registrations peaking sharply around 2023 at approximately 60,000. After this peak, registrations drop drastically by 2025. The graph illustrates rapid growth followed by a significant decline in EV registrations.

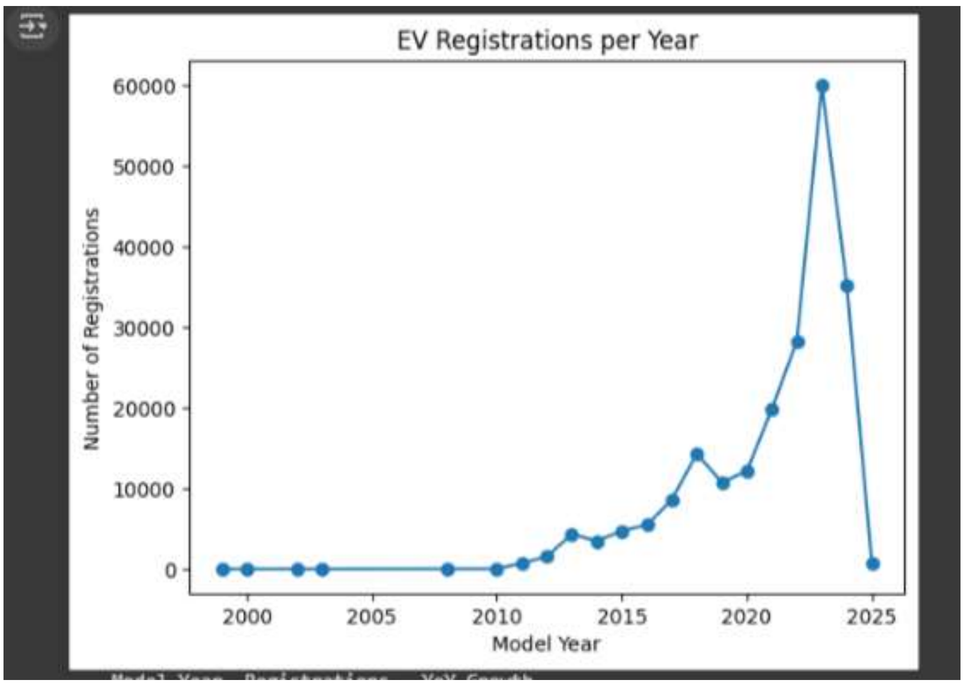


Figure 19: Temporal Analysis - Model Year

Figure 20 shows EV adoption by type from 1999 to 2025, with BEVs (in blue) and PHEVs (in orange). EV registrations remained low until 2017, after which they rose sharply, especially for BEVs. The peak occurs in 2023, followed by a sharp decline in registrations for both vehicle types. BEVs consistently outpace PHEVs throughout the period.

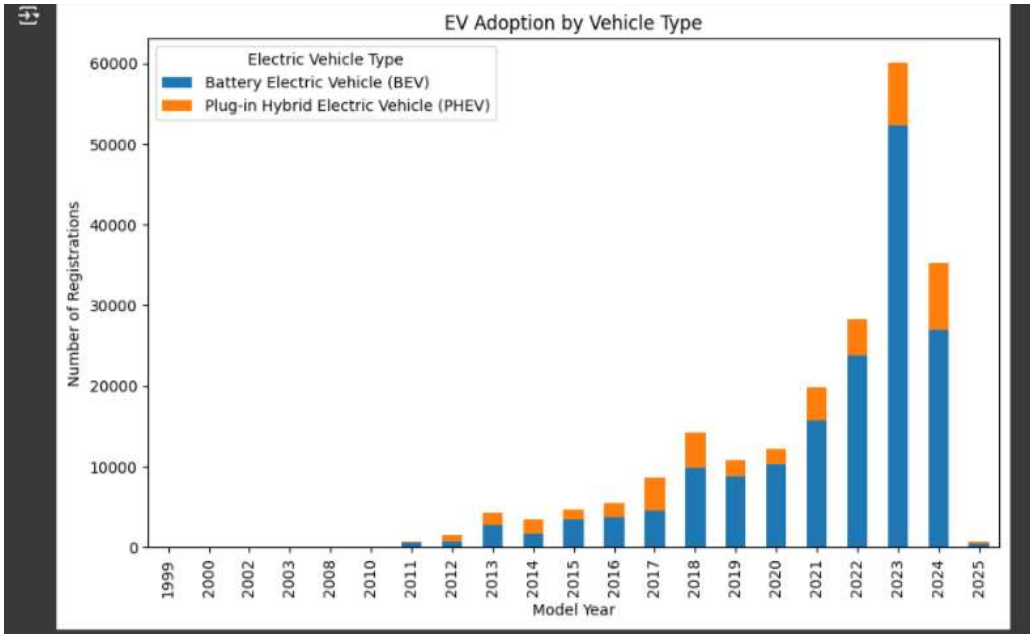


Figure 20: Temporal Analysis - Model Year/Model