BIRZEIT UNIVERSITY

Faculty of Engineering and Technology

Electrical and Computer Engineering Department

Machine Learning and Data Science - ENCS5341

Assignment #1

Prepared by: Ibaa Taleeb **ID** : 1203073

Instructor : Dr.Ismail Khater

Section: 2

Date: 30/10/2024

Abstract

Registration information about plug-in hybrid electric cars (PHEVs) and battery electric vehicles (BEVs) registered with the Washington State Department of Licensing is included in this dataset, which was made available by the State of Washington. Each vehicle's unique VIN, county and city of registration, make and model, electric vehicle type, and anticipated electric range are all included in the dataset, which has 17 columns of precise information. This regularly updated dataset, which spans model years from 2013 to the present, offers important insights about the distribution and uptake of electric vehicles in Washington State. The data provides a thorough resource for researching EV infrastructure and growth over time in the area and is well-suited for examining trends in EV popularity, geographic distribution, and performance measures.

Table of Contents

Table of Figures

Data Cleaning and Feature Engineering:

In this part, First we need to read the dataset and print the first row using .head(1) to show how the dataset is, after that print the dataset information using .info() where we can see the columns with information about them (Count of non null values, Datatype of each columns) .

1.Document Missing Values:

using .isnull().sum() we can check for missing values and to document their frequency and distribution across features, by using len() method we can find the number of items in the data set then calculate the percentage of missing values.

```
The number of items in the dataset: 205439

                      Missing Values  Percentage
County                             3    0.001460
City                               3    0.001460
Postal Code                        3    0.001460
Model                              1    0.000487
Electric Range                     8    0.003894
Base MSRP                          8    0.003894
Legislative District             442    0.215149
Vehicle Location                   8    0.003894
Electric Utility                   3    0.001460
2020 Census Tract                  3    0.001460
```

Figure 1:number of missing values with there percentage

2.Missing Value Strategies:

we can apply multiple strategies (mean/median imputation, dropping rows) to fill the missing values, columns with low missing values then drop rows and using .mean() , .median() to calculate the mean and median to fill the missing values, check the null values so we will get 0 null value.

From statistics calculation and Correlation Matrix we can note that the imputation process was effective in filling missing values with minimal impact on central tendencies, variability, or correlations, ensuring that analysis integrity is maintained across features.

```
Initial Summary Statistics (Before Imputation):

         Postal Code    Model Year  Electric Range      Base MSRP  \
count  205436.000000  205439.000000  205431.000000  205431.000000
mean    98177.971870    2020.960363      52.164342     922.670532
std      2419.037479       2.989059      88.075859    7761.753602
min      1731.000000    1997.000000       0.000000       0.000000
25%     98052.000000    2019.000000       0.000000       0.000000
50%     98125.000000    2022.000000       0.000000       0.000000
75%     98372.000000    2023.000000      48.000000       0.000000
max     99577.000000    2025.000000     337.000000  845000.000000

       Legislative District  DOL Vehicle ID  2020 Census Tract
count         204997.000000    2.054390e+05       2.054360e+05
mean              28.970848    2.277156e+08       5.297704e+10
std               14.910052    7.205737e+07       1.588435e+09
min                1.000000    4.469000e+03       1.001020e+09
25%               17.000000    1.935324e+08       5.303301e+10
50%               33.000000    2.382368e+08       5.303303e+10
75%               42.000000    2.618718e+08       5.305307e+10
max               49.000000    4.792548e+08       5.602100e+10
Initial Correlation Matrix:
```

```
Post-Imputation Summary Statistics:

         Postal Code    Model Year  Electric Range      Base MSRP
count  205430.000000  205430.000000  205430.000000  205430.000000
mean    98177.958409    2020.960405      52.164632     922.294009
std      2419.071523       2.989057      88.075118    7760.599223
min      1731.000000    1997.000000       0.000000       0.000000
25%     98052.000000    2019.000000       0.000000       0.000000
50%     98125.000000    2022.000000       0.000000       0.000000
75%     98372.000000    2023.000000      48.000000       0.000000
max     99577.000000    2025.000000     337.000000  845000.000000

       Legislative District  DOL Vehicle ID  2020 Census Tract
count         205430.000000    2.054300e+05       2.054300e+05
mean              28.979604    2.277149e+08       5.297704e+10
std               14.895345    7.205607e+07       1.588459e+09
min                1.000000    4.469000e+03       1.001020e+09
25%               17.000000    1.935324e+08       5.303301e+10
50%               33.000000    2.382366e+08       5.303303e+10
75%               42.000000    2.618718e+08       5.305307e+10
max               49.000000    4.792548e+08       5.602100e+10
```

Figure 2: statistics calculation before and post imputation

```
Initial Correlation Matrix:

                      Postal Code  Model Year  Electric Range  Base MSRP  \
Postal Code              1.000000   -0.001019       -0.001556  -0.002685
Model Year              -0.001019    1.000000       -0.507739  -0.231280
Electric Range          -0.001556   -0.507739        1.000000   0.113545
Base MSRP               -0.002685   -0.231280        0.113545   1.000000
Legislative District    -0.410291   -0.015640        0.019888   0.010440
DOL Vehicle ID           0.006023    0.200597       -0.131015  -0.037803
2020 Census Tract        0.496433    0.005724       -0.001186   0.000878

                      Legislative District  DOL Vehicle ID  2020 Census Tract
Postal Code                      -0.410291        0.006023           0.496433
Model Year                       -0.015640        0.200597           0.005724
Electric Range                    0.019888       -0.131015          -0.001186
Base MSRP                         0.010440       -0.037803           0.000878
Legislative District              1.000000       -0.009259          -0.101356
DOL Vehicle ID                   -0.009259        1.000000           0.003559
2020 Census Tract                -0.101356        0.003559           1.000000
```

```
Post-Imputation Correlation Matrix:

                      Postal Code  Model Year  Electric Range  Base MSRP  \
Postal Code              1.000000   -0.001019       -0.001554  -0.002688
Model Year              -0.001019    1.000000       -0.507727  -0.231269
Electric Range          -0.001554   -0.507727        1.000000   0.113568
Base MSRP               -0.002688   -0.231269        0.113568   1.000000
Legislative District    -0.061594   -0.015723        0.019907   0.010466
DOL Vehicle ID           0.006024    0.200607       -0.131007  -0.037782
2020 Census Tract        0.496433    0.005724       -0.001185   0.000877

                      Legislative District  DOL Vehicle ID  2020 Census Tract
Postal Code                      -0.061594        0.006024           0.496433
Model Year                       -0.015723        0.200607           0.005724
Electric Range                    0.019907       -0.131007          -0.001185
Base MSRP                         0.010466       -0.037782           0.000877
Legislative District              1.000000       -0.009331          -0.011730
DOL Vehicle ID                   -0.009331        1.000000           0.003559
2020 Census Tract                -0.011730        0.003559           1.000000
```

Figure 3:Correlation matrix before and post imputation

3.Feature Encoding : using One-Hot Encoding to encode categorical features into a numerical format suitable for analysis, create binary columns for each category, making it easier to process and analyze these categorical features in the model. This is an example of encoded columns County is a feature before encoding, after _ represent a data from the feature ,the 1 appears when the county is Ada, else 0 .

```
County_Ada  \  Electric Utility_PUGET SOUND ENERGY INC  \
       0.0                                        1.0
       0.0                                        1.0
       0.0                                        0.0
       0.0                                        0.0
       0.0                                        1.0
```

Figure 4:one hot encoding sample

4. Normalization : normalization adjusts the scale of all numerical features to fit a specific range, By applying Min-Max scaling the data becomes consistent and suitable for models that may be sensitive to feature scaling, improving the analysis accuracy and performance. As we can see in the below figure that the values between 0 and 1 .

$$v' = \frac{v - min_F}{max_F - min_F}(new\_max_F - new\_min_F) + new\_min_F$$

```
Legislative District  DOL Vehicle ID  2020 Census Tract
            0.708333        0.502200           0.945730
            0.708333        0.989419           0.945730
            0.875000        0.236026           0.945693
            0.937500        0.225736           0.945692
            0.395833        0.368168           0.946311
```

Figure 5:Normalize the numerical features

Exploratory Data Analysis:

5. Descriptive Statistics: using methods to calculate the mean, median, and standard deviation, we can notice that the descriptive statistics for selected numerical features show varying central tendencies and dispersions. The "Model Year" has a high mean (0.856) and median (0.893), indicating recent data distribution, while "Electric Range" and "Base MSRP" show low means and medians, reflecting a skewed distribution with many low or zero values. "Legislative District" and "DOL Vehicle ID" have moderate means and medians, and "2020 Census Tract" has a high mean and low standard deviation, indicating limited variance across tracts. The table below shows the skewness for each feature

```
Descriptive Statistics for specified numerical features:
                          Mean    Median  Standard Deviation
Postal Code           0.985702  0.985160            0.024723
Model Year            0.855727  0.892857            0.106752
Electric Range        0.154790  0.000000            0.261353
Base MSRP             0.001092  0.000000            0.009186
Legislative District  0.582726  0.666667            0.310626
DOL Vehicle ID        0.475140  0.497094            0.150354
2020 Census Tract     0.944675  0.945693            0.028870
```

Figure 6:Descriptive Statistics for Numerical values

Postal Code	Mean ≈ Median	the distribution is likely close to symmetric or has minimal skew.
Model Year	Mean < Median	slightly left-skewed (negatively skewed) distribution
Electric Range	Mean > Median	right-skewed (positively skewed) distribution
Base MSRP	Mean > Median	right-skewed (positively skewed) distribution
Legislative District	Mean < Median	left-skewed (negatively skewed) distribution
DOL Vehicle ID	Mean ≈ Median	indicating a fairly symmetric distribution or minimal skew.
2020 Census Tract	Mean ≈ Median	symmetric distribution or minimal skew.

6. Spatial Distribution: The map displays the spatial distribution of electric vehicles (EVs) across various regions in Washington State. Each colored circle represents clusters of EVs, with the number inside indicating the count of vehicles in that area.

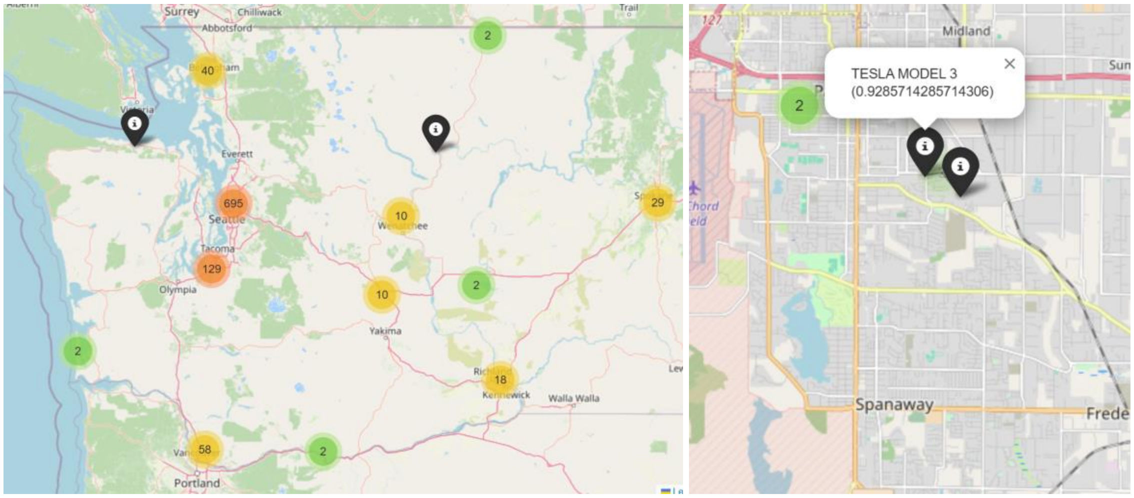

Figure 7: spatial distribution of electric vehicles (EVs)

7. Model Popularity: To Analyze the popularity of different EV models, we can use different methods, for example, the figures bellow shows that the **Tesla Model Y** is the most popular electric vehicle, making up approximately **21.14%** of all EVs, followed by the **Tesla Model 3** at **15.63%**. Other prominent models include the **Nissan Leaf** and **Tesla Model S**, though their shares drop significantly compared to the top two. Together, the top 10 models account for a substantial portion of the market, indicating a strong preference for Tesla and other high-performance or affordable EV options. This concentration in a few models highlights consumer leanings toward popular, established brands.

```
Top 10 Most Popular EV Models:
Model
MODEL Y          43437
MODEL 3          32113
LEAF             13488
MODEL S           7881
BOLT EV           6727
MODEL X           6249
VOLT              4829
ID.4              4564
MUSTANG MACH-E    4154
WRANGLER          4047
Name: count, dtype: int64

Total unique EV models: 152
```

```
Most popular model accounts for 21.14% of the total EVs.

Percentage of total EVs for each of the top 10 models:
Model
MODEL Y           21.143605
MODEL 3           15.631480
LEAF               6.565484
MODEL S            3.836194
BOLT EV            3.274467
MODEL X            3.041794
VOLT               2.350588
ID.4               2.221595
MUSTANG MACH-E     2.022021
WRANGLER           1.969937
Name: count, dtype: float64
```

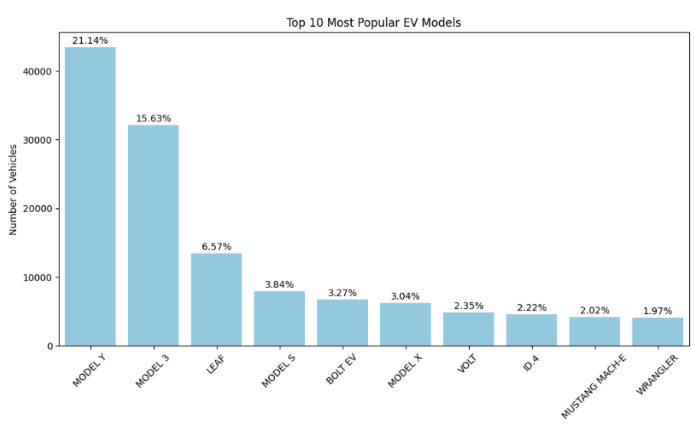

Figure 8:Model Popularity

Also we can see the Top EV Models by Manufacturer

```
    SUBARU             SOLTERRA    1174
     TESLA              MODEL Y   43436
     TH!NK                 CITY       5
    TOYOTA           RAV4 PRIME    3054
VOLKSWAGEN                 ID.4    4564
     VOLVO                 XC90    1695
```

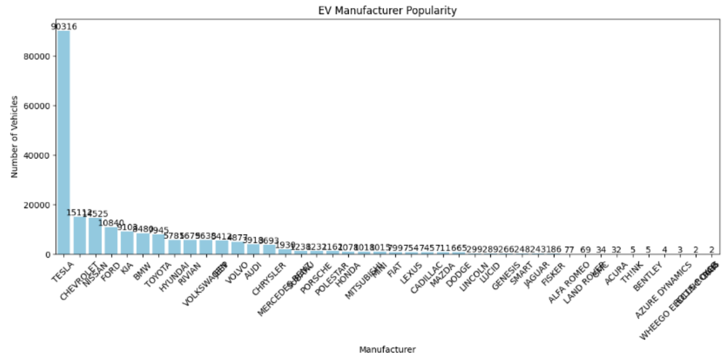

Figure 9: Top EV Models by Manufacturer

8. Investigate the relationship between every pair of numeric features. Are there any correlations?

Yes, there is a correlation between numeric features, the correlation matrix below shows the correlation, we can see that there is a moderate negative correlation between the model year and the electric range. This suggests that newer models tend to have higher electric ranges, which is consistent with advancements in technology and battery efficiency over time also there is a moderate negative correlation, suggesting that vehicles registered in certain postal codes may be associated with specific legislative districts. This could indicate regional differences in vehicle popularity or regulations, and moderate positive correlation indicates that postal codes are somewhat related to census tracts.

```
Correlation Matrix:
                      Postal Code  Model Year  Electric Range  Base MSRP  \
Postal Code              1.000000   -0.001019       -0.001556  -0.002685
Model Year              -0.001019    1.000000       -0.507739  -0.231280
Electric Range          -0.001556   -0.507739        1.000000   0.113545
Base MSRP               -0.002685   -0.231280        0.113545   1.000000
Legislative District    -0.410291   -0.015640        0.019888   0.010440
DOL Vehicle ID           0.006023    0.200597       -0.131015  -0.037803
2020 Census Tract        0.496433    0.005724       -0.001186   0.000878

                      Legislative District  DOL Vehicle ID  2020 Census Tract
Postal Code                      -0.410291        0.006023           0.496433
Model Year                       -0.015640        0.200597           0.005724
Electric Range                    0.019888       -0.131015          -0.001186
Base MSRP                         0.010440       -0.037803           0.000878
Legislative District              1.000000       -0.009259          -0.101356
DOL Vehicle ID                   -0.009259        1.000000           0.003559
2020 Census Tract                -0.101356        0.003559           1.000000
```

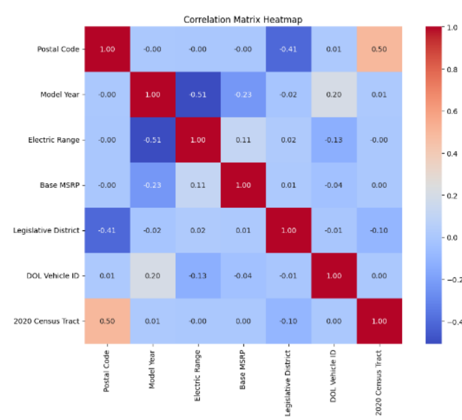

Figure 10:Correlation Matrix

Visualization: we can explore the relationships between features using various visualizations (e.g., histograms, scatter plots, boxplots)

9. Data Exploration Visualizations

The figure below shows the Count Plot of Make and Electric Vehicle Type, we can see for example, NISSAN doesn't have plug-in hybrid electric vehicles(PHEV), TEASLA manufacturing a battery electric vehicle (BEV) more than (PHEV),and so on .

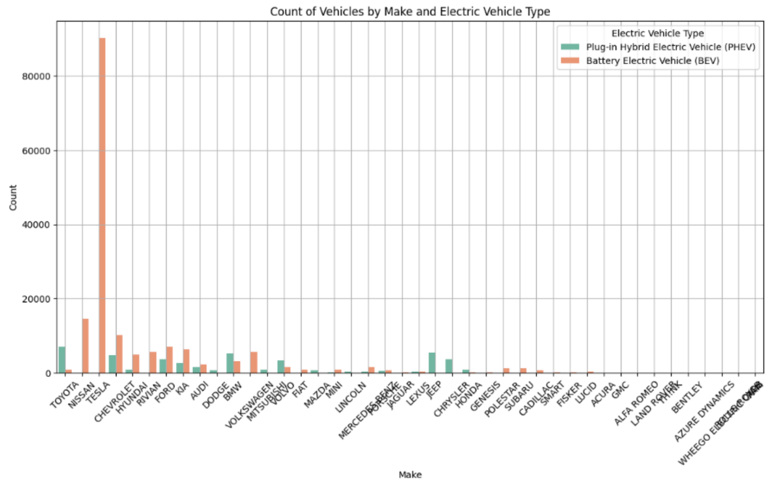

Figure 11: Count Plot of Make and Electric Vehicle Type

Bar plot Show the average electric range of different makes.

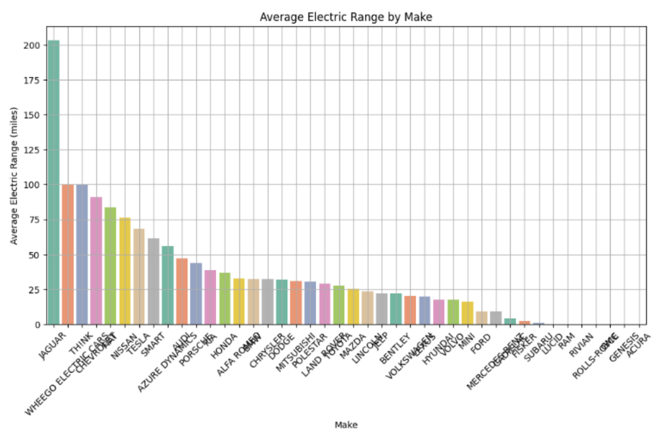

Figure 12:Avg electric range by make

Scatter Plot of Model Year vs. Electric Range: Analyze how electric range has changed over the years.

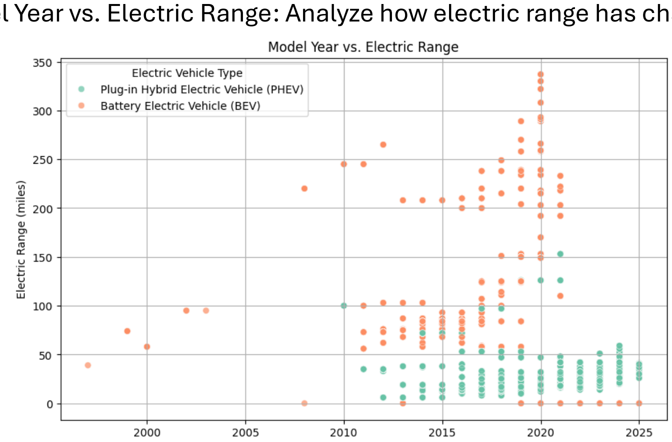

Figure 13:Model year vs electric range

bar and pie plots of Count of Vehicles by Electric Vehicle Type: Show the distribution of the number of vehicles across different EV types.

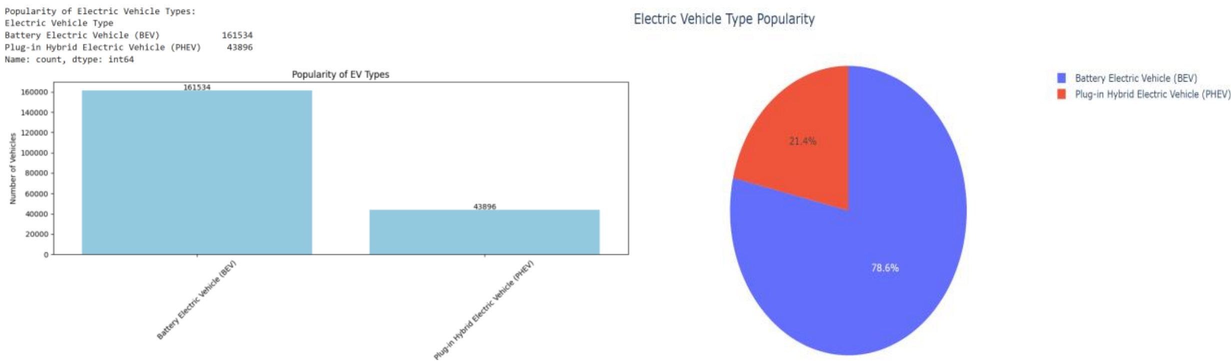

Figure 14: distribution of the number of vehicles across different EV types.

In figure 15 , Box Plot of Electric Range by Legislative District: Examine the distribution of electric ranges in different legislative districts, figure 16 shows the Histogram of Electric Range: Visualize the distribution of electric ranges across all vehicles and figure 17 histograms for all numerical features:

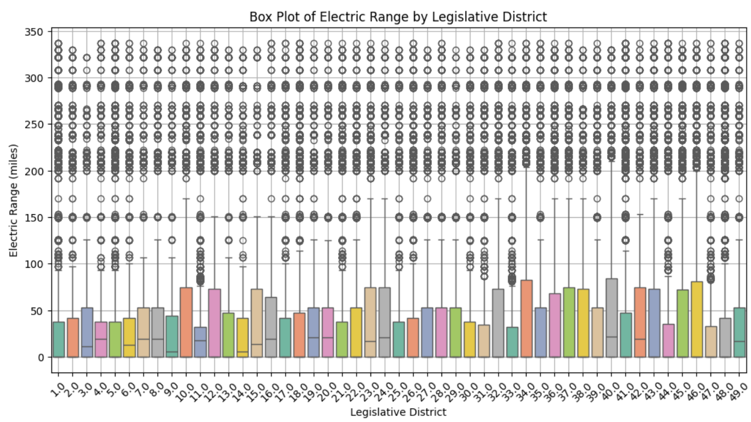

Figure 15: the distribution of electric ranges in different legislative districts.

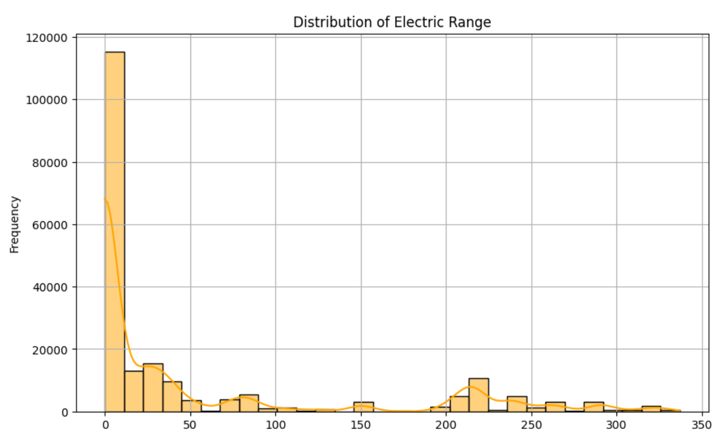

Figure 16: the distribution of electric ranges across all vehicles

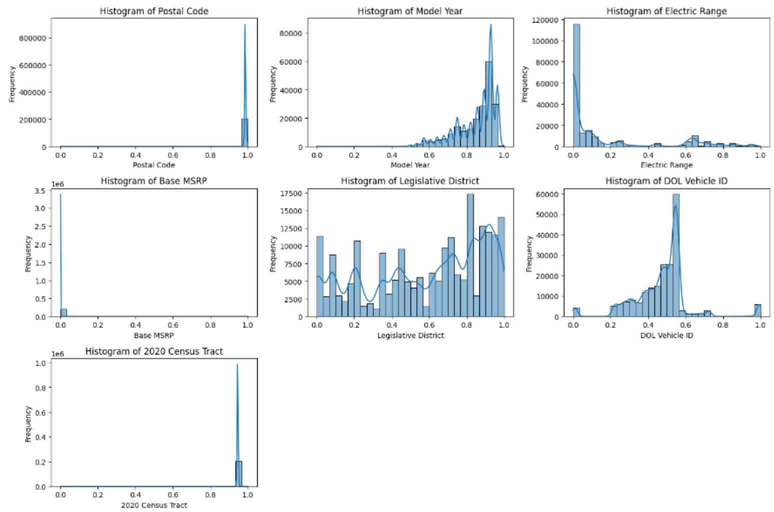

Figure 17:skewness plots for Numerical features

10. Comparative Visualization:

The Figure below shows the Distribution of EVs by City, we can see that Aberdeen have the maximum number of EVs, also the plot for Distribution of EVs by county we notice that Asotion Have the maximum number of EVs

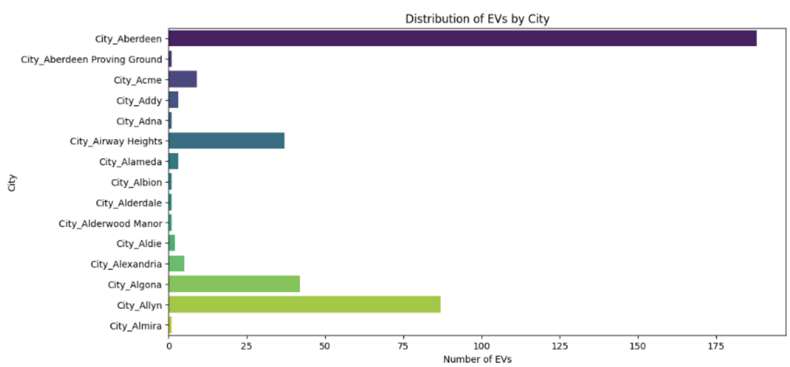

Figure 18:Distribution of EVs By city

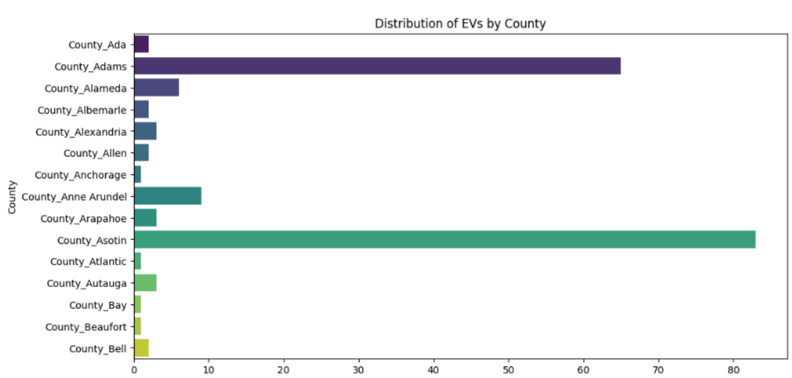

Figure 19: Distribution of EVs By

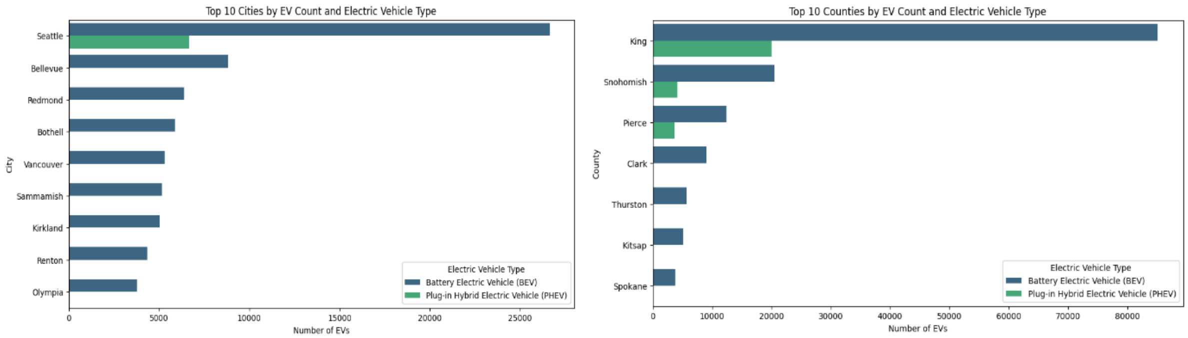

Figure 20: Top 10 Counties and Cities by EV Count and Electric Vehicle Type

11. Temporal Analysis (Optional):

The two figures below shows the Model Popularity by year, we can notice the increases and decreases of each model by years.

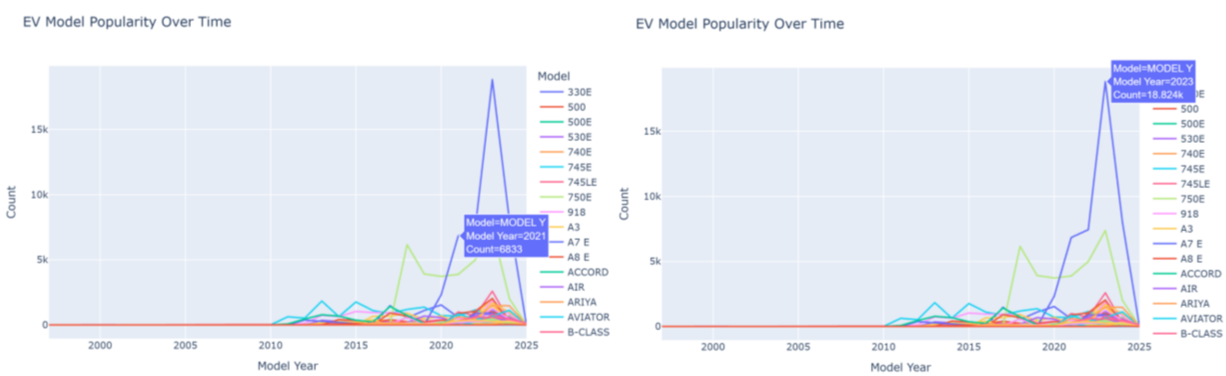

Figure 21:Ev Model Over Time

The plot below shows the Electric Vehicle Type over time ,we can see the high increase in manufactured Battery Electric Vehicle in 2023.

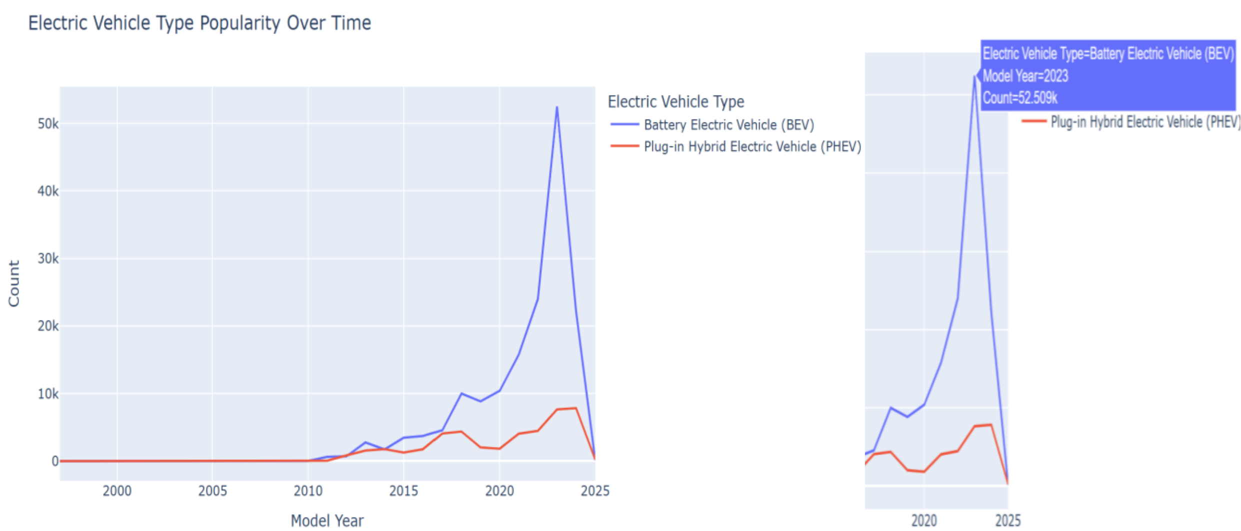

Figure 22: Electric Vehicle Type over time

The figure below shows the growth rate of EV models over time

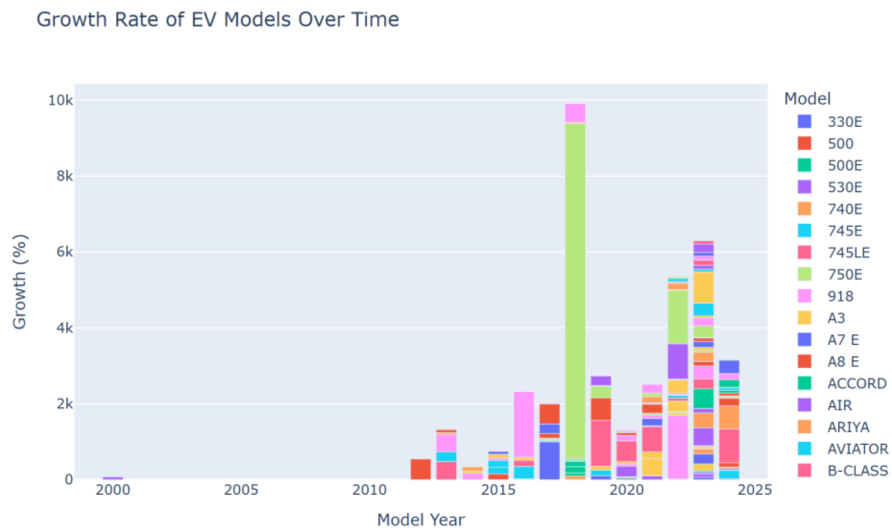

Figure 23:Growth rate of EV models over time