BIRZEIT UNIVERSITY

Faculty of Engineering & Technology

Electrical & Computer Engineering Department

ENCS5341

MACHINE LEARNING AND DATA SCIENCE

Report Assignment 1

Student Name: Hamza Al Shaer

Student ID: 1211162

Section: 3

Instructor: Dr. Ismail Khater

Date: 30/10/2024

❖ Document Missing Values

```
Terminal    Local ×  +  ⌄
PS C:\Users\hamza\Desktop\4th first symmary\ML\Assigment 1> python main.py
<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
PS C:\Users\hamza\Desktop\4th first symmary\ML\Assigment 1>
```

Figure 1: Data Set Information

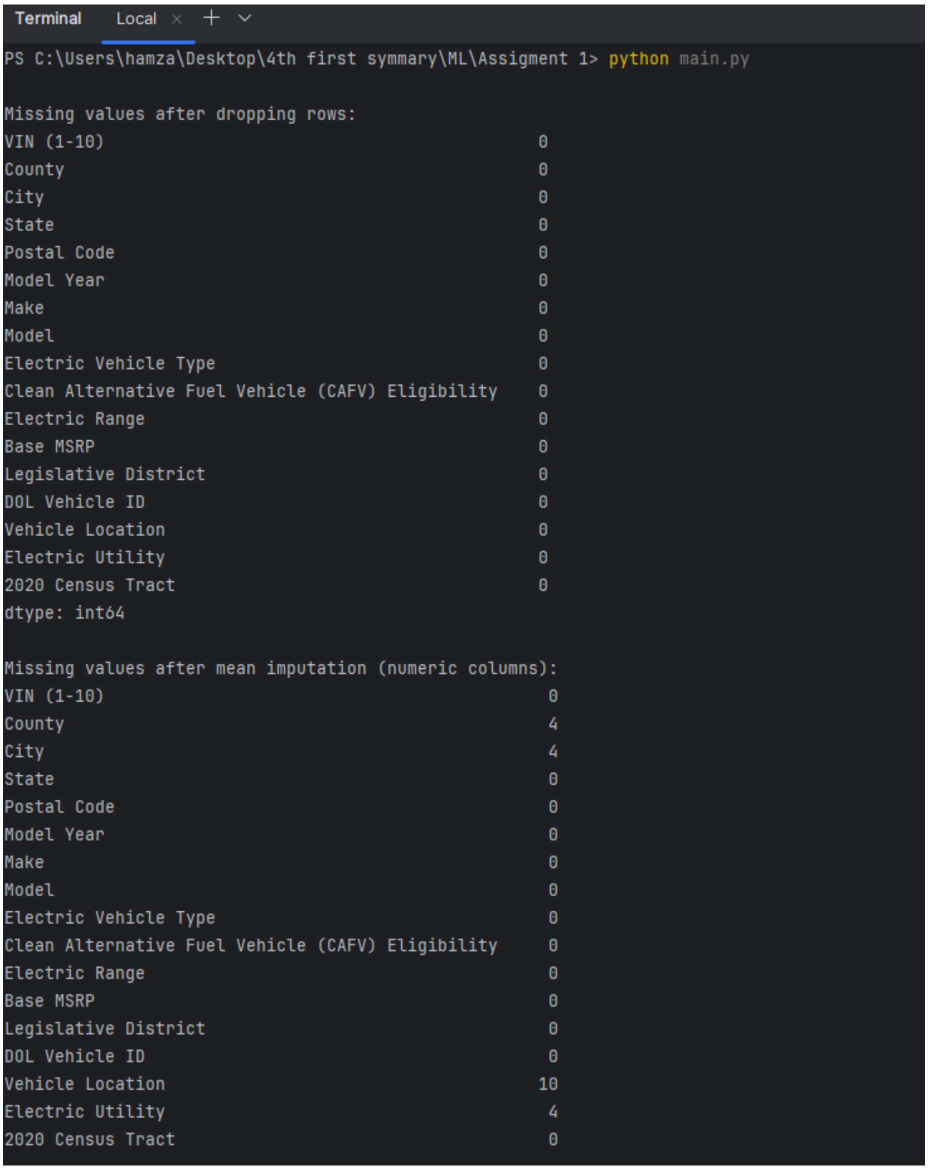
```
Terminal    Local ×  +  ⌄
PS C:\Users\hamza\Desktop\4th first symmary\ML\Assigment 1> python main.py

Missing Values Report:
                      Missing Values  Percentage
County                             4    0.001903
City                               4    0.001903
Postal Code                        4    0.001903
Electric Range                     5    0.002379
Base MSRP                          5    0.002379
Legislative District             445    0.211738
Vehicle Location                  10    0.004758
Electric Utility                   4    0.001903
2020 Census Tract                  4    0.001903
PS C:\Users\hamza\Desktop\4th first symmary\ML\Assigment 1>
```

Figure 2: Document Missing Value Output

This function identifies and documents any missing values across the dataset's attributes. Its purpose is to assess data completeness by highlighting variables that contain null or NaN values. By quantifying these missing values, the function allows for better data preprocessing decisions, such as handling imputation, deletion, or ignoring certain records. As show in result this function represents number of missing values in each column in Data Set.

❖ Missing Value Strategies



```
Terminal    Local × + ⌄
PS C:\Users\hamza\Desktop\4th first symmary\ML\Assigment 1> python main.py

Missing values after dropping rows:
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                                 0
DOL Vehicle ID                                       0
Vehicle Location                                     0
Electric Utility                                     0
2020 Census Tract                                    0
dtype: int64

Missing values after mean imputation (numeric columns):
VIN (1-10)                                            0
County                                                4
City                                                  4
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                                  0
DOL Vehicle ID                                        0
Vehicle Location                                     10
Electric Utility                                      4
2020 Census Tract                                     0
```

Figure 3: results Missing Values after Stratygis Misiing Value

This Strategies apply only on numeric columns, in first Strategy "dropping column" in this approach, any column with a high percentage of missing values was completely removed from the dataset. Then in imputation Strategy used to fill in these gaps with estimated values, preserving the dataset's overall structure and completeness. Various imputation methods (e.g., mean, median, or mode) were applied based on the nature of the data. The outcomes show how each method impacts data quality and model performance, highlighting whether dropping or imputing values was more effective in retaining data utility while addressing missingness.

❖ Feature Encoding

```
Terminal   Local
PS C:\Users\hamza\Desktop\4th first symmary\ML\Assigment 1> python main.py
The Feature Encoding for Model
        VIN (1-10)     County              City State  Postal Code  Model Year     Make                    Electric Vehicle Type  ... Model_WRANGLER  Model_X3  Model_X5  Model_XC40  Model_XC60 Model_XC90 Model_XM  Model_ZDX
0       5UXTA6C0XM     Kitsap           Seabeck    WA      98380.0        2021      BMW  Plug-in Hybrid Electric Vehicle (PHEV)  ...            0.0       0.0       1.0         0.0         0.0        0.0      0.0        0.0
1       5YJ3E1EB1J     Kitsap           Poulsbo    WA      98370.0        2018    TESLA           Battery Electric Vehicle (BEV)  ...            0.0       0.0       0.0         0.0         0.0        0.0      0.0        0.0
2       WP0AD2A73G  Snohomish           Bothell    WA      98012.0        2016  PORSCHE  Plug-in Hybrid Electric Vehicle (PHEV)  ...            0.0       0.0       0.0         0.0         0.0        0.0      0.0        0.0
3       5YJ3E1EB5J     Kitsap         Bremerton    WA      98310.0        2018    TESLA           Battery Electric Vehicle (BEV)  ...            0.0       0.0       0.0         0.0         0.0        0.0      0.0        0.0
4       1N4AZ1CP3K       King           Redmond    WA      98052.0        2019   NISSAN           Battery Electric Vehicle (BEV)  ...            0.0       0.0       0.0         0.0         0.0        0.0      0.0        0.0
...            ...        ...               ...   ...          ...         ...      ...                                      ...  ...            ...       ...       ...         ...         ...        ...      ...        ...
210160  KNDCM3LD8J      Grant        Moses Lake    WA      98837.0        2018      KIA  Plug-in Hybrid Electric Vehicle (PHEV)  ...            0.0       0.0       0.0         0.0         0.0        0.0      0.0        0.0
210161  JM3KKBHA0R     Kitsap         Bremerton    WA      98312.0        2024    MAZDA  Plug-in Hybrid Electric Vehicle (PHEV)  ...            0.0       0.0       0.0         0.0         0.0        0.0      0.0        0.0
210162  JTMAB3FV3P     Pierce  University Place    WA      98466.0        2023   TOYOTA  Plug-in Hybrid Electric Vehicle (PHEV)  ...            0.0       0.0       0.0         0.0         0.0        0.0      0.0        0.0
210163  5YJ3E1EA1M       King          Kirkland    WA      98034.0        2021    TESLA           Battery Electric Vehicle (BEV)  ...            0.0       0.0       0.0         0.0         0.0        0.0      0.0        0.0
210164  7SAYGDEE0P    Spokane           Spokane    WA      99208.0        2023    TESLA           Battery Electric Vehicle (BEV)  ...            0.0       0.0       0.0         0.0         0.0        0.0      0.0        0.0

[210165 rows x 169 columns]
[210165 rows x 169 columns]
```

Figure 4: Feture Encoding Result on Attribute 'Model'

```
The Feature Encoding for Make
        VIN (1-10)     County              City State  Postal Code  Model Year       Model  ... Make_TESLA Make_TH!NK  Make_TOYOTA  Make_VINFAST  Make_VOLKSWAGEN  Make_VOLVO Make_WHEEGO ELECTRIC CARS
0       5UXTA6C0XM     Kitsap           Seabeck    WA      98380.0        2021          X5  ...        0.0        0.0          0.0           0.0              0.0         0.0                      0.0
1       5YJ3E1EB1J     Kitsap           Poulsbo    WA      98370.0        2018     MODEL 3  ...        1.0        0.0          0.0           0.0              0.0         0.0                      0.0
2       WP0AD2A73G  Snohomish           Bothell    WA      98012.0        2016    PANAMERA  ...        0.0        0.0          0.0           0.0              0.0         0.0                      0.0
3       5YJ3E1EB5J     Kitsap         Bremerton    WA      98310.0        2018     MODEL 3  ...        1.0        0.0          0.0           0.0              0.0         0.0                      0.0
4       1N4AZ1CP3K       King           Redmond    WA      98052.0        2019        LEAF  ...        0.0        0.0          0.0           0.0              0.0         0.0                      0.0
...            ...        ...               ...   ...          ...         ...         ...  ...        ...        ...          ...           ...              ...         ...                      ...
210160  KNDCM3LD8J      Grant        Moses Lake    WA      98837.0        2018        NIRO  ...        0.0        0.0          0.0           0.0              0.0         0.0                      0.0
210161  JM3KKBHA0R     Kitsap         Bremerton    WA      98312.0        2024       CX-90  ...        0.0        0.0          0.0           0.0              0.0         0.0                      0.0
210162  JTMAB3FV3P     Pierce  University Place    WA      98466.0        2023  RAV4 PRIME  ...        0.0        0.0          1.0           0.0              0.0         0.0                      0.0
210163  5YJ3E1EA1M       King          Kirkland    WA      98034.0        2021     MODEL 3  ...        1.0        0.0          0.0           0.0              0.0         0.0                      0.0
210164  7SAYGDEE0P    Spokane           Spokane    WA      99208.0        2023     MODEL Y  ...        1.0        0.0          0.0           0.0              0.0         0.0                      0.0

[210165 rows x 59 columns]
```

Figure 5:Feture Encoding Result on Attribute 'Make'

The goal of feature encoding is to transform categorical variables into a format that machine learning algorithms can utilize. Take an attribution and convert this attribution to (0,1), This conversion is done by counting all possible options for this attribute and then placing 1 When you achieve this option and the rest of the options set 0. This done by use "One-Hot Encoding" This technique creates binary (0 or 1) columns for each unique category in the specified feature.

❖ Normalization

```
Terminal   Local
PS C:\Users\hamza\Desktop\4th first symmary\ML\Assigment 1> python main.py
 Before Min-Max Normalized Dataset:

         Postal Code     Model Year  Electric Range      Base MSRP  Legislative District  DOL Vehicle ID  2020 Census Tract
count  210161.000000  210165.000000   210160.000000  210160.000000         209720.000000    2.101650e+05       2.101610e+05
mean    98178.209406    2021.048657       50.602241     897.676889             28.929954    2.290774e+08       5.297929e+10
std      2445.429402       2.988941       86.973210    7653.588604             14.908392    7.115519e+07       1.551466e+09
min      1731.000000    1999.000000        0.000000       0.000000              1.000000    4.469000e+03       1.001020e+09
25%     98052.000000    2019.000000        0.000000       0.000000             17.000000    1.948816e+08       5.303301e+10
50%     98125.000000    2022.000000        0.000000       0.000000             32.000000    2.405164e+08       5.303303e+10
75%     98374.000000    2023.000000       42.000000       0.000000             42.000000    2.629758e+08       5.305307e+10
max     99577.000000    2025.000000      337.000000  845000.000000             49.000000    4.792548e+08       5.602100e+10
 After Min-Max Normalized Dataset:

         Postal Code     Model Year  Electric Range      Base MSRP  Legislative District  DOL Vehicle ID  2020 Census Tract
count  210161.000000  210165.000000   210160.000000  210160.000000         209720.000000   210165.000000      210161.000000
mean        0.985704       0.848025        0.150155       0.001062              0.581874        0.477982           0.944716
std         0.024993       0.114959        0.258081       0.009058              0.310592        0.148472           0.028198
min         0.000000       0.000000        0.000000       0.000000              0.000000        0.000000           0.000000
25%         0.984414       0.769231        0.000000       0.000000              0.333333        0.406629           0.945693
50%         0.985160       0.884615        0.000000       0.000000              0.645833        0.501850           0.945693
75%         0.987705       0.923077        0.124629       0.000000              0.854167        0.548714           0.946057
max         1.000000       1.000000        1.000000       1.000000              1.000000        1.000000           1.000000
PS C:\Users\hamza\Desktop\4th first symmary\ML\Assigment 1>
```

Figure 6: Reults Befoer and After make Normliation on DataSet

Normalization is Strategy use for fit and transform data in Data Set, here use Min Max Scaler normalization The Min Max Scaler scales each numerical feature to a [0, 1] range. This transformation is applied only to the numerical columns, so categorical or text features remain unaffected. Before Normalization: provides summary statistics (like min, max, mean) for the dataset before scaling, which helps to understand the original range of values. After Normalization: The statistics after normalization will confirm that all numerical features are now within the [0, 1] range.

❖ Descriptive Statistics

```
Terminal    Local ×  +  ⌄
PS C:\Users\hamza\Desktop\4th first symmary\ML\Assigment 1> python main.py
Mean values:
Postal Code                98178.21
Model Year                  2021.05
Electric Range                50.60
Base MSRP                    897.68
Legislative District          28.93
DOL Vehicle ID          229077434.80
2020 Census Tract     52979294365.52
dtype: float64

Median values:
Postal Code                98125.00
Model Year                  2022.00
Electric Range                 0.00
Base MSRP                      0.00
Legislative District          32.00
DOL Vehicle ID          240516391.00
2020 Census Tract     53033030101.00
dtype: float64

Standard Deviation values:
Postal Code                2445.43
Model Year                    2.99
Electric Range               86.97
Base MSRP                  7653.59
Legislative District         14.91
DOL Vehicle ID          71155185.13
2020 Census Tract     1551466456.12
dtype: float64
```

Figure 7: Descriptive Statistics (Mean,Median,STdv) Reslult

Here calculate mean and median and standard deviation by call function as attribute for data set, the Mean and Median are statistical measures that are only applicable to numeric data. It doesn't make sense to compute the mean or median for non-numeric data types (like strings or categories).

❖ Spatial Distribution

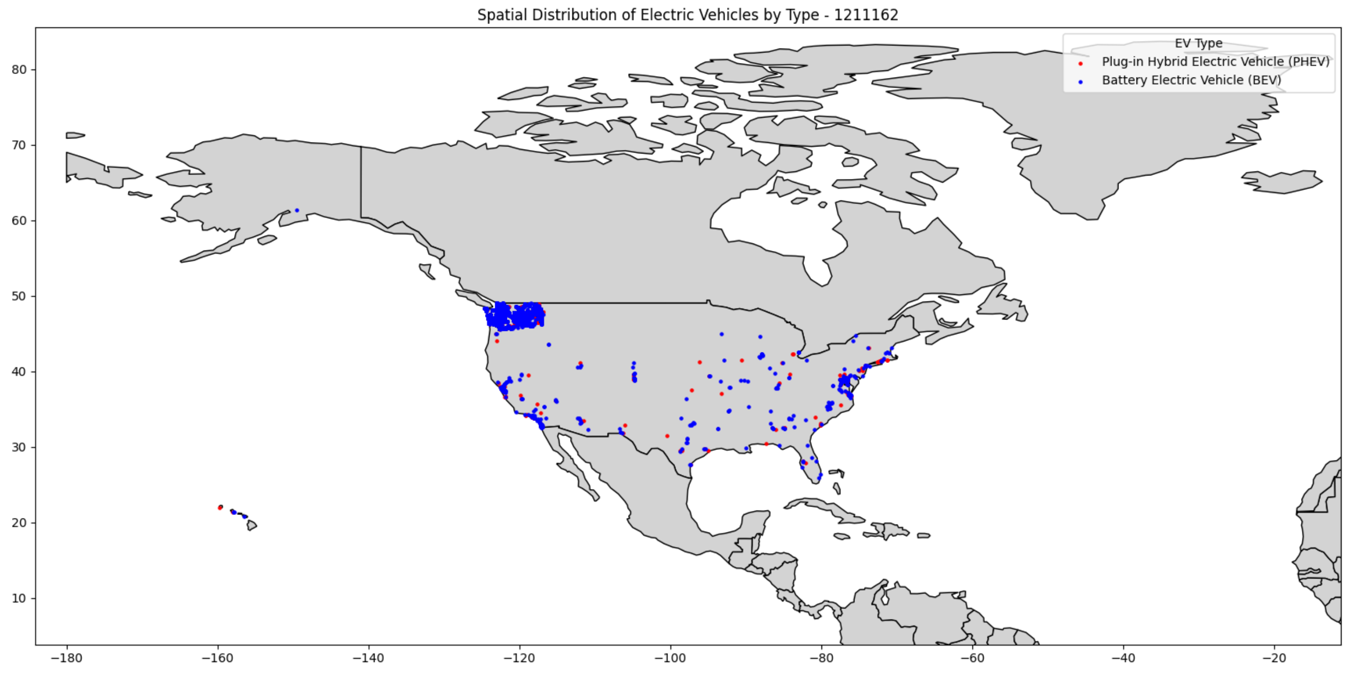


Figure 8: Spatial Distribution Result

The goal of the spatial distribution map is to visualize where electric vehicles (EVs) are located geographically based on latitude and longitude data, then import file "ne_110m_admin_0_countries.shp" to plot a world map and determine location on it. I can use html file for world map but this method faster.

❖ Model Popularity

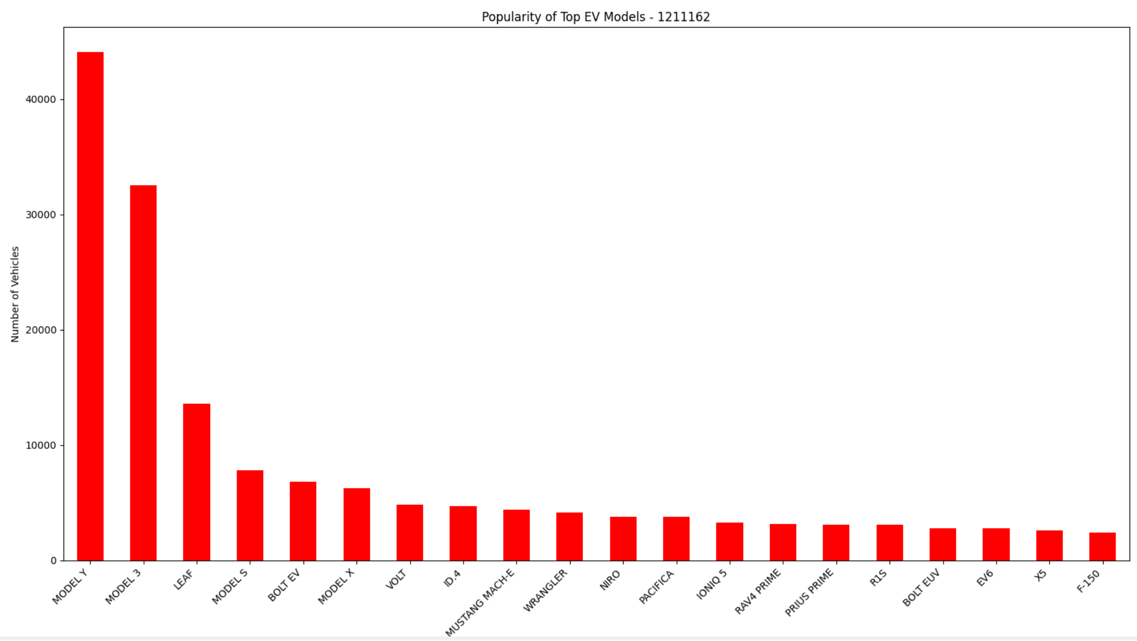


Figure 9: Model Popularity Result

This analysis ranks the top 20 car models based on popularity metrics. The purpose is to identify which models are most favored in the market and understand their appeal.

❖ Correlation (relationship between every pair of numeric features)

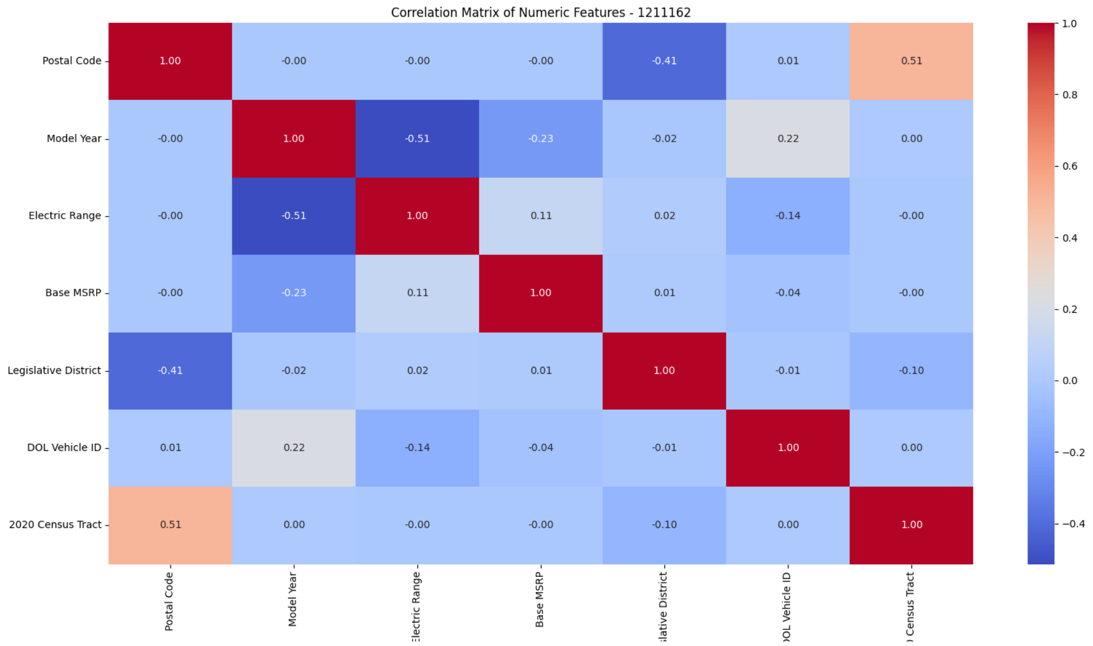


Figure 10:Correlation Matrix

This is Correlation matrix visualization for numerical features, Values close to +1 indicate a strong positive correlation, meaning that as one variable increases, the other tends to increase as well. Values close to -1 indicate a strong negative correlation, meaning that as one variable increases, the other tends to decrease. Values near 0 indicate little to no linear relationship between the variables.

❖ Data Exploration Visualizations

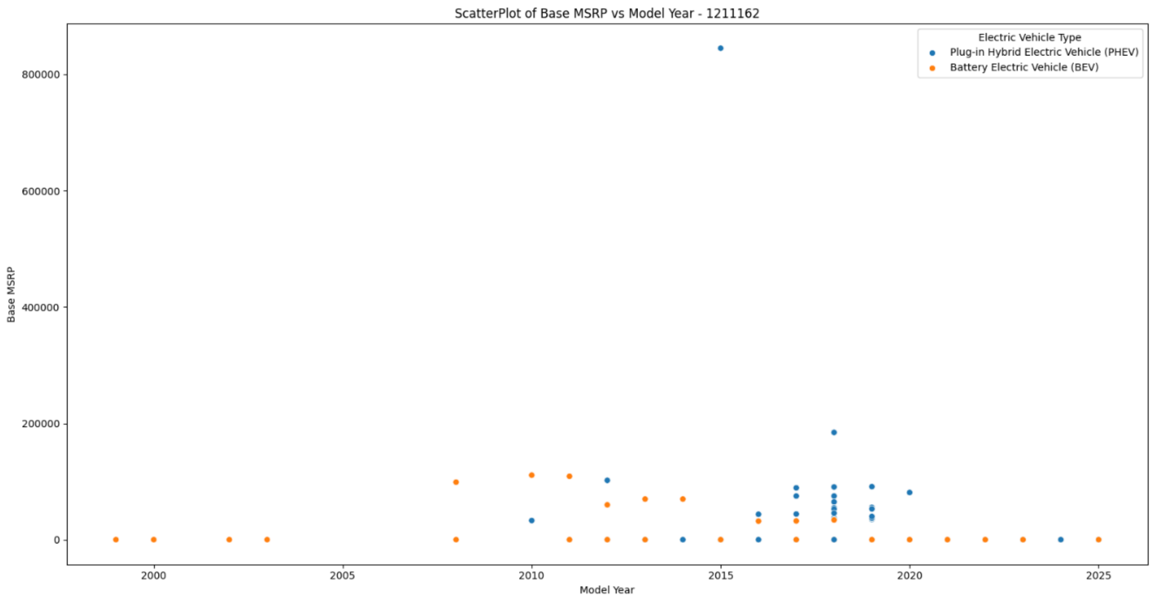


Figure 11: Exploration Visualization (ScatterPlot)

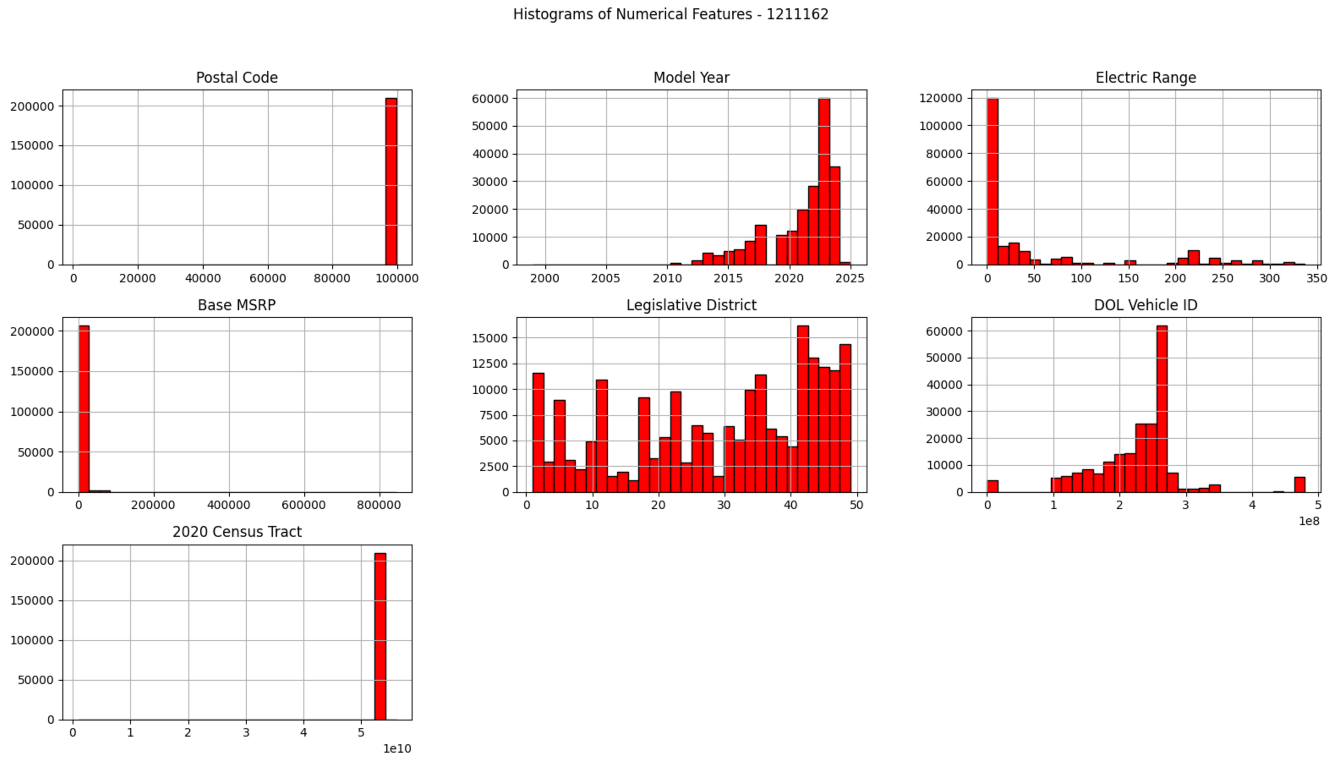


Figure 12:Exploration Visualization (Histograms)

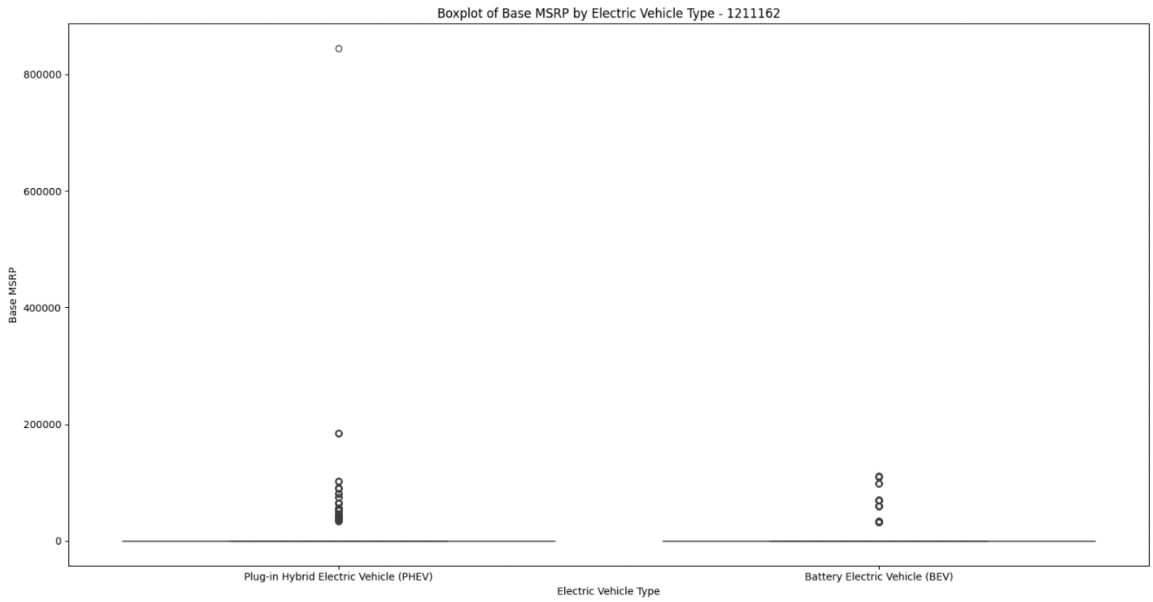
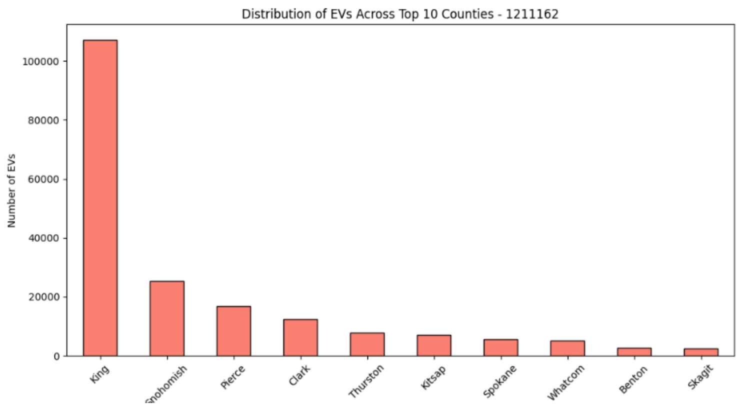
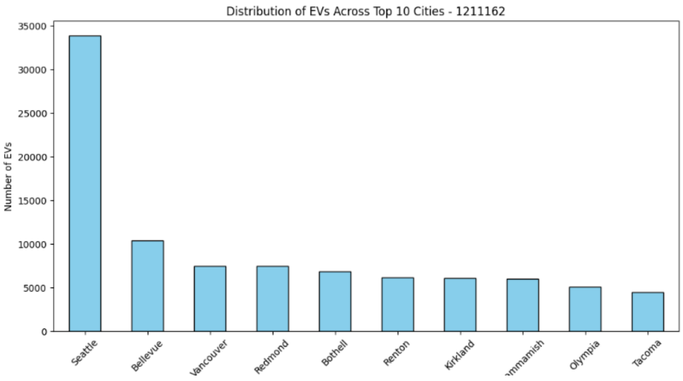


Figure 13:Exploration Visualization (boxplots)

The Histograms result show the distribution of individual numerical features, helping identify patterns such as skewness, multimodality, or outliers. Scatter plot explores the relationship between Model Year and Base MSRP, with color encoding for Electric Vehicle Type. Boxplots provide a visual summary of the distribution of Base MSRP across different categories of Electric Vehicle Type.

❖ Comparative Visualization (the distribution of EVs across different locations)

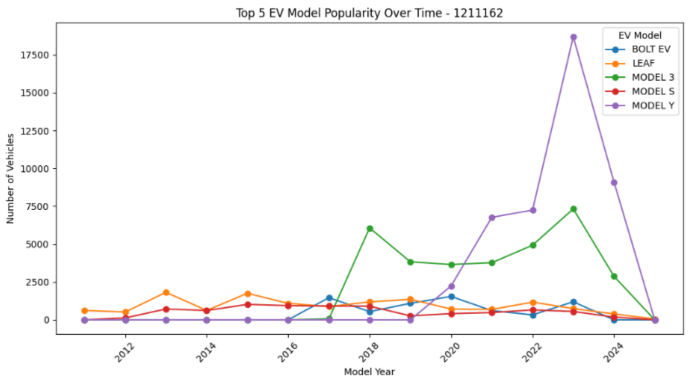


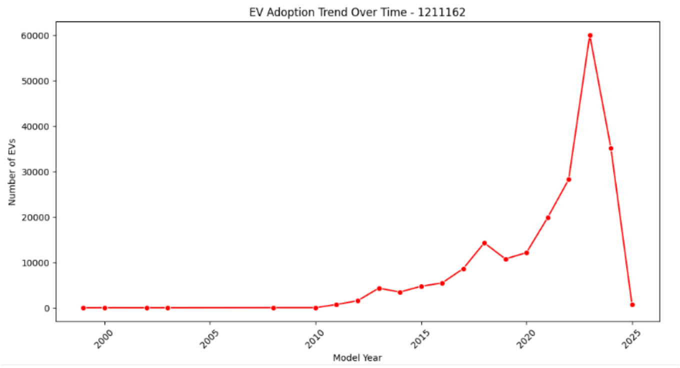


- "Distribution of EVs Across Top 10 Cities" shows that Seattle has the highest number of EVs by a significant margin, followed by cities like Bellevue, Vancouver, and Redmond.
- "Distribution of EVs Across Top 10 Counties" highlights that King County has the highest EV count, far surpassing others like Snohomish and Pierce.

❖ Temporal Analysis (Optional)

❖





- EV Adoption Trend Over Time: This analysis tracks the number of electric vehicles introduced each year, illustrating the growth trajectory in EV adoption. The line plot generated for the adoption trend shows yearly counts of EV models from the dataset's start year to the present. Peaks and trends in the graph may reflect key market developments.
- Model Popularity Over Time: Focusing on the five most popular EV models, this part of the analysis charts each model's yearly presence. By filtering the dataset to include only these top models and plotting their popularity by model year, the analysis reveals fluctuations in each model's demand. The resulting line plot highlights specific years when certain models surged in popularity, potentially due to new releases or enhancements.