**Faculty of Engineering and Technology Department of Electrical and Computer Engineering**
**Machine Learning and Data Science**

Malak Moqbel 1210608
Omar Hussain 1212739

**Instructor's name:** Dr.Yazan Abu Farha
**Section: 1**
**Submission date:** 30.10.2024

---

The dataset provides information on electric vehicles, covering details such as model, make, range, and pricing. It highlights trends in vehicle specifications and model years, mainly focusing on recent releases. Some fields, like range and pricing, have missing values, which may affect analysis completeness.

1. **Document Missing Values: Check for missing values and document their frequency and distribution across features.**

```
County: missing values = 4 || (0.001903%)
City: missing values = 4 || (0.001903%)
Postal Code: missing values = 4 || (0.001903%)
Electric Range: missing values = 5 || (0.002379%)
Base MSRP: missing values = 5 || (0.002379%)
Legislative District: missing values = 445 || (0.211738%)
Vehicle Location: missing values = 10 || (0.004758%)
Electric Utility: missing values = 4 || (0.001903%)
2020 Census Tract: missing values = 4 || (0.001903%)
```

The data is mostly complete, with very few missing values across features. Most columns have only tiny gaps, but the "Legislative District" column has a few more missing entries, which may need extra attention.

2. **Missing Value Strategies: If missing values are present, apply multiple strategies (e.g., mean/median imputation, dropping rows) and compare their impact on the analysis.**

```
Original Data Statistics:
         Postal Code     Model Year  Electric Range      Base MSRP  \
count  210161.000000  210165.000000   210160.000000  210160.000000
mean    98178.209406    2021.048657       50.602241     897.676889
std      2445.429402       2.988941       86.973210    7653.588604
min      1731.000000    1999.000000        0.000000       0.000000
25%     98052.000000    2019.000000        0.000000       0.000000
50%     98125.000000    2022.000000        0.000000       0.000000
75%     98374.000000    2023.000000       42.000000       0.000000
max     99577.000000    2025.000000      337.000000  845000.000000

       Legislative District  DOL Vehicle ID  2020 Census Tract
count         209720.000000    2.101650e+05       2.101610e+05
mean              28.929954    2.290774e+08       5.297929e+10
std               14.908392    7.115519e+07       1.551466e+09
min                1.000000    4.469000e+03       1.001020e+09
25%               17.000000    1.948816e+08       5.303301e+10
50%               32.000000    2.405164e+08       5.303303e+10
75%               42.000000    2.629758e+08       5.305307e+10
max               49.000000    4.792548e+08       5.602100e+10

Imputed Data Statistics:
         Postal Code     Model Year  Electric Range      Base MSRP  \
count  210161.000000  210165.000000   210165.000000  210165.000000
mean    98178.209406    2021.048657       50.601037     897.676889
std      2445.429402       2.988941       86.972525    7653.497560
min      1731.000000    1999.000000        0.000000       0.000000
25%     98052.000000    2019.000000        0.000000       0.000000
50%     98125.000000    2022.000000        0.000000       0.000000
75%     98374.000000    2023.000000       42.000000       0.000000
max     99577.000000    2025.000000      337.000000  845000.000000

       Legislative District  DOL Vehicle ID  2020 Census Tract
count         209720.000000    2.101650e+05       2.101610e+05
mean              28.929954    2.290774e+08       5.297929e+10
std               14.908392    7.115519e+07       1.551466e+09
min                1.000000    4.469000e+03       1.001020e+09
25%               17.000000    1.948816e+08       5.303301e+10
50%               32.000000    2.405164e+08       5.303303e+10
75%               42.000000    2.629758e+08       5.305307e+10
max               49.000000    4.792548e+08       5.602100e+10
```

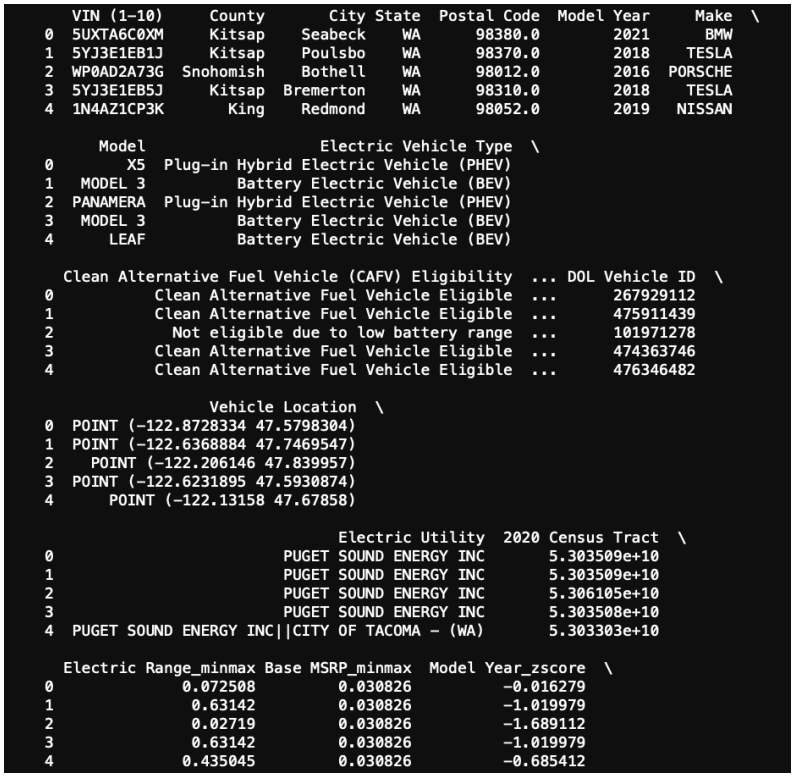
This table compares how the dataset looks before and after addressing missing values. Initially, some columns like `Electric Range` and `Base MSRP` had missing entries. To handle this, we used two techniques: filling in missing `Base MSRP` values with the column's mean (average) and filling in missing `Electric Range` values with the median (middle value). These choices helped maintain the overall distribution and balance of the data Additionally, we tried removing rows with any missing data entirely, and observed that this did not greatly impact the average or spread of the remaining data.

**3. Feature Encoding: Encode categorical features (e.g., Make, Model) using techniques like one-hot encoding.**

In this feature encoding process, we applied one-hot encoding to the categorical columns "Make" and "Model." One-hot encoding is a method of converting categorical data into a binary format that can be used in analysis or machine learning. For each unique value in "Make" and "Model," a new column is created, with a value of 1 indicating the presence of that specific make or model for a given entry, and 0 otherwise. This process allows us to transform non-numeric data into a format that can be interpreted by algorithms without introducing numerical biases. As a result, the dataset expanded, adding new columns for each unique make and model, making the data more machine-readable while preserving the categorical information.

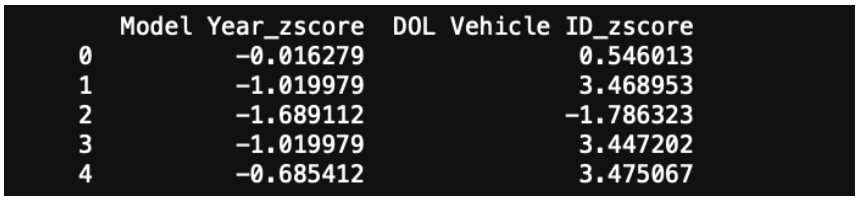
```
    VIN (1-10)     County       City State  Postal Code  Model Year  \
0  5UXTA6C0XM     Kitsap    Seabeck    WA      98380.0        2021
1  5YJ3E1EB1J     Kitsap    Poulsbo    WA      98370.0        2018
2  WP0AD2A73G  Snohomish    Bothell    WA      98012.0        2016
3  5YJ3E1EB5J     Kitsap  Bremerton    WA      98310.0        2018
4  1N4AZ1CP3K       King    Redmond    WA      98052.0        2019

                     Electric Vehicle Type  \
0  Plug-in Hybrid Electric Vehicle (PHEV)
1          Battery Electric Vehicle (BEV)
2  Plug-in Hybrid Electric Vehicle (PHEV)
3          Battery Electric Vehicle (BEV)
4          Battery Electric Vehicle (BEV)

  Clean Alternative Fuel Vehicle (CAFV) Eligibility  Electric Range  \
0           Clean Alternative Fuel Vehicle Eligible            30.0
1           Clean Alternative Fuel Vehicle Eligible           215.0
2             Not eligible due to low battery range            15.0
3           Clean Alternative Fuel Vehicle Eligible           215.0
4           Clean Alternative Fuel Vehicle Eligible           150.0

    Base MSRP  ...  Model_VOLT  Model_WHEEGO Model_WRANGLER Model_X3  Model_X5  \
0         0.0  ...       False         False         False    False      True
1         0.0  ...       False         False         False    False     False
2         0.0  ...       False         False         False    False     False
3         0.0  ...       False         False         False    False     False
4         0.0  ...       False         False         False    False     False

   Model_XC40  Model_XC60  Model_XC90  Model_XM  Model_ZDX
0       False       False       False     False      False
1       False       False       False     False      False
2       False       False       False     False      False
3       False       False       False     False      False
4       False       False       False     False      False

[5 rows x 209 columns]
```

**4. Normalization: Normalize numerical features if necessary for chosen analysis methods.**



| | VIN (1-10) | County | City | State | Postal Code | Model Year | Make | Model | Electric Vehicle Type | Clean Alternative Fuel Vehicle (CAFV) Eligibility | ... | DOL Vehicle ID | Vehicle Location | Electric Utility | 2020 Census Tract | Electric Range_minmax | Base MSRP_minmax | Model Year_zscore |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 5UXTA6C0XM | Kitsap | Seabeck | WA | 98380.0 | 2021 | BMW | X5 | Plug-in Hybrid Electric Vehicle (PHEV) | Clean Alternative Fuel Vehicle Eligible | ... | 267929112 | POINT (-122.8728334 47.5798304) | PUGET SOUND ENERGY INC | 5.303509e+10 | 0.072508 | 0.030826 | -0.016279 |
| 1 | 5YJ3E1EB1J | Kitsap | Poulsbo | WA | 98370.0 | 2018 | TESLA | MODEL 3 | Battery Electric Vehicle (BEV) | Clean Alternative Fuel Vehicle Eligible | ... | 475911439 | POINT (-122.6368884 47.7469547) | PUGET SOUND ENERGY INC | 5.303509e+10 | 0.63142 | 0.030826 | -1.019979 |
| 2 | WP0AD2A73G | Snohomish | Bothell | WA | 98012.0 | 2016 | PORSCHE | PANAMERA | Plug-in Hybrid Electric Vehicle (PHEV) | Not eligible due to low battery range | ... | 101971278 | POINT (-122.206146 47.839957) | PUGET SOUND ENERGY INC | 5.306105e+10 | 0.02719 | 0.030826 | -1.689112 |
| 3 | 5YJ3E1EB5J | Kitsap | Bremerton | WA | 98310.0 | 2018 | TESLA | MODEL 3 | Battery Electric Vehicle (BEV) | Clean Alternative Fuel Vehicle Eligible | ... | 474363746 | POINT (-122.6231895 47.5930874) | PUGET SOUND ENERGY INC | 5.303508e+10 | 0.63142 | 0.030826 | -1.019979 |
| 4 | 1N4AZ1CP3K | King | Redmond | WA | 98052.0 | 2019 | NISSAN | LEAF | Battery Electric Vehicle (BEV) | Clean Alternative Fuel Vehicle Eligible | ... | 476346482 | POINT (-122.13158 47.67858) | PUGET SOUND ENERGY INC\|\|CITY OF TACOMA - (WA) | 5.303303e+10 | 0.435045 | 0.030826 | -0.685412 |

The normalization process applied here uses Min-Max scaling, which adjusts the values of the numerical features (`Electric Range` and `Base MSRP`) to fall within a range of 0 to 1. For example, if the `Electric Range` in miles ranges from 15 to 337 across vehicles, Min-Max scaling would convert the minimum value (15 miles) to 0 and the maximum (337 miles) to 1, with all other values proportionally adjusted in between. This helps put different scales on a common basis, making it easier to compare features like `Electric Range` and `Base MSRP` without one feature dominating due to its larger numeric range. In the results, we see scaled values like `0.030826` for the `Base MSRP_minmax` feature, indicating the value has been normalized between the minimum and maximum MSRP values in the dataset. This approach is useful for algorithms that are sensitive to data ranges, like clustering or neural networks.



| | Model Year_zscore | DOL Vehicle ID_zscore |
|---|---|---|
| 0 | -0.016279 | 0.546013 |
| 1 | -1.019979 | 3.468953 |
| 2 | -1.689112 | -1.786323 |
| 3 | -1.019979 | 3.447202 |
| 4 | -0.685412 | 3.475067 |

Z-score normalization is a technique that transforms each data point based on the average (mean) and variability (standard deviation) of that feature. This method helps us understand how far each value is from the center (the mean) in a standardized way.

After applying Z-score normalization, each value tells us how many standard deviations it is above or below the mean.

For instance, if the average Model Year is 2021, and a specific model year has a Z-score of -1.69, this means it's about 1.69 standard deviations below the 2021 average. If a Z-score is positive, the value is above the mean; if it's negative, it's below. The spread of all

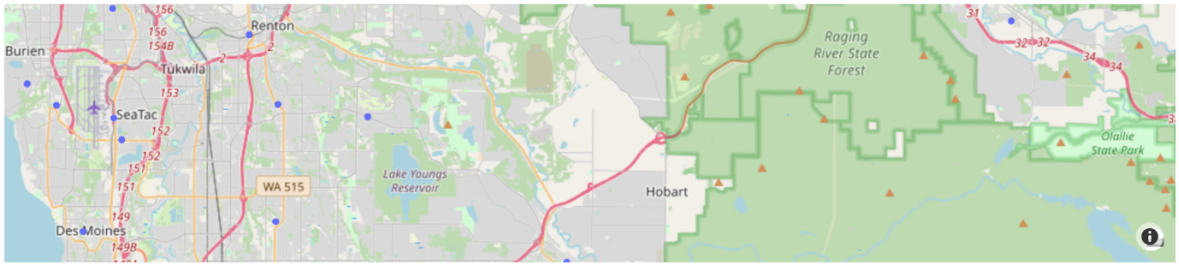
transformed values will now be around 0 (the mean), with most values within a range of about -3 to +3. This standardization is particularly useful when comparing features that originally had different units (like years and ID numbers), as it scales them to a common baseline.

5. **Descriptive Statistics: Calculate summary statistics (mean, median, standard deviation) for numerical features.**

| | Mean | Median | Standard Deviation |
|---|---|---|---|
| Electric Range | 116.217155 | 75.0 | 98.728475 |
| Base MSRP | 57012.926866 | 59900.0 | 22829.647001 |

We calculated descriptive statistics for the Electric Range and Base MSRP to summarize and understand the data. We computed the mean by adding all values for each feature and dividing by the total number, giving us an average `Electric Range` of about 116 miles and an average Base MSRP of approximately $57,013. The median was found by sorting each feature's values and locating the middle point, giving us a Base MSRP median of $59,000, which indicates that half of the prices are below this value. Lastly, the standard deviation was calculated by measuring how spread out the values are from the mean, resulting in 98 for Electric Range and 22,830 for Base MSRP. This high deviation in both features shows a diverse range in electric vehicle capabilities and prices.

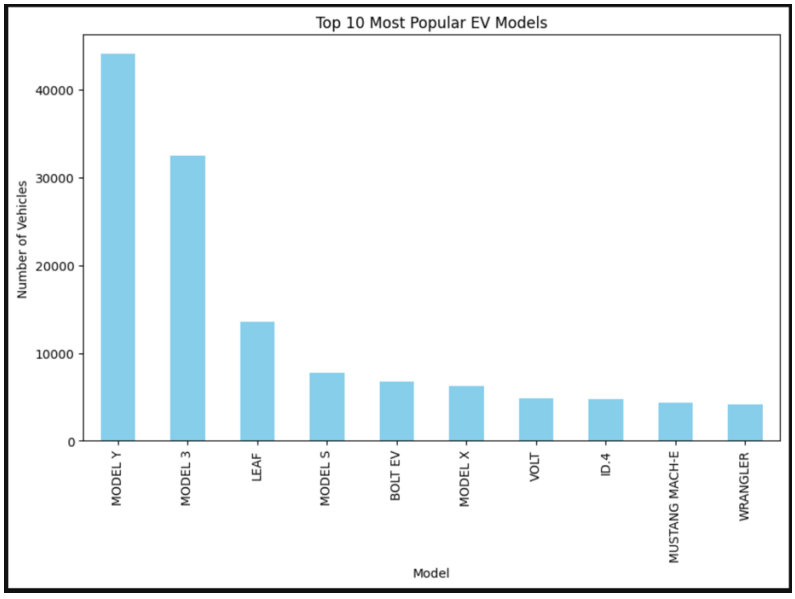
6. **Spatial Distribution: Visualize the spatial distribution of EVs across locations (e.g., maps).**




The map shows the locations of electric vehicles (EVs) based on their recorded GPS points. By extracting the latitude and longitude from each vehicle's location data, we've placed them on an interactive map. Each dot represents an EV's position, allowing us to visually assess where EVs are concentrated in the area. This helps in identifying patterns in EV distribution across different regions.

7. **Model Popularity: Analyze the popularity of different EV models (categorical data) and identify any trends.**

```
Top 10 Most Popular EV Models:
Model
MODEL Y           44038
MODEL 3           32520
LEAF              13606
MODEL S            7795
BOLT EV            6780
MODEL X            6239
VOLT               4815
ID.4               4716
MUSTANG MACH-E     4363
WRANGLER           4116
Name: count, dtype: int64
```




The analysis of EV model popularity highlights the ten most frequently registered electric vehicle models. Model Y and Model 3 dominate in terms of popularity, with registrations of 44,038 and 32,520, respectively. This suggests a strong consumer preference for these models compared to others, such as the LEAF and Model S, which, while still popular, have significantly lower counts. This trend may reflect factors like brand appeal, specific features, or affordability. The bar chart further illustrates this, with Model Y and Model 3 clearly leading, underscoring Tesla's prominent role in the electric vehicle market.

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

To investigate relationships among numeric features, we created a **correlation matrix** that shows how each feature relates to every other feature in the dataset. In this matrix, values range between -1 and 1, where values closer to 1 indicate a strong positive relationship, values near -1 show a strong negative relationship, and values around 0 suggest little to no relationship.

```
Correlation Matrix:
                      Postal Code  Model Year  Legislative District  \
Postal Code              1.000000   -0.000693              0.018969
Model Year              -0.000693    1.000000             -0.016191
Legislative District     0.018969   -0.016191              1.000000
DOL Vehicle ID           0.005389    0.215703             -0.010241
2020 Census Tract        0.521067    0.005152              0.074912
Electric Range_minmax   -0.000932   -0.513540              0.018689
Base MSRP_minmax        -0.004581   -0.230651              0.010036
Model Year_zscore       -0.000693    1.000000             -0.016191
DOL Vehicle ID_zscore    0.005389    0.215703             -0.010241
Longitude               -0.717461   -0.003913             -0.227499
Latitude                 0.392953    0.000225              0.207366

                      DOL Vehicle ID  2020 Census Tract  \
Postal Code                 0.005389           0.521067
Model Year                  0.215703           0.005152
Legislative District       -0.010241           0.074912
DOL Vehicle ID              1.000000           0.002982
2020 Census Tract           0.002982           1.000000
Electric Range_minmax      -0.140696          -0.000443
Base MSRP_minmax           -0.039503          -0.001326
Model Year_zscore           0.215703           0.005152
DOL Vehicle ID_zscore       1.000000           0.002982
Longitude                  -0.001431          -0.408546
Latitude                   -0.008433           0.527542

                      Electric Range_minmax  Base MSRP_minmax  \
Postal Code                       -0.000932         -0.004581
Model Year                        -0.513540         -0.230651
Legislative District               0.018689          0.010036
DOL Vehicle ID                    -0.140696         -0.039503
2020 Census Tract                 -0.000443         -0.001326
Electric Range_minmax              1.000000          0.114157
Base MSRP_minmax                   0.114157          1.000000
Model Year_zscore                 -0.513540         -0.230651
DOL Vehicle ID_zscore             -0.140696         -0.039503
Longitude                          0.000655          0.004485
Latitude                           0.003437         -0.001146

                      Model Year_zscore  DOL Vehicle ID_zscore  Longitude  \
Postal Code                   -0.000693               0.005389  -0.717461
Model Year                     1.000000               0.215703  -0.003913
Legislative District          -0.016191              -0.010241  -0.227499
DOL Vehicle ID                 0.215703               1.000000  -0.001431
2020 Census Tract              0.005152               0.002982  -0.408546
Electric Range_minmax         -0.513540              -0.140696   0.000655
Base MSRP_minmax              -0.230651              -0.039503   0.004485
Model Year_zscore              1.000000               0.215703  -0.003913
DOL Vehicle ID_zscore          0.215703               1.000000  -0.001431
Longitude                     -0.003913              -0.001431   1.000000
Latitude                       0.000225              -0.008433  -0.472864

                      Latitude
Postal Code           0.392953
Model Year            0.000225
Legislative District  0.207366
DOL Vehicle ID       -0.008433
2020 Census Tract     0.527542
Electric Range_minmax 0.003437
Base MSRP_minmax     -0.001146
Model Year_zscore     0.000225
```

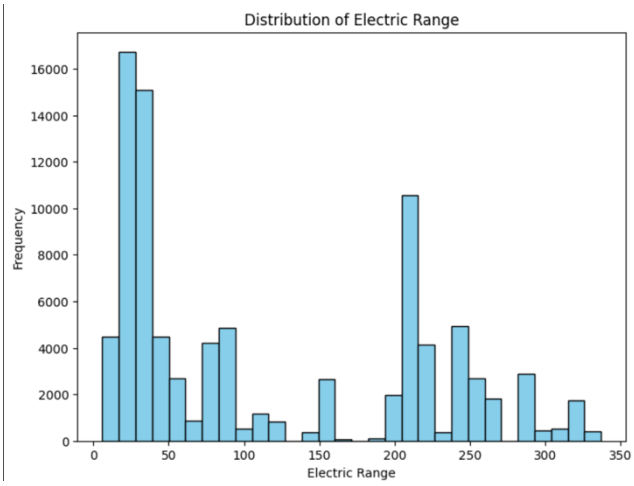
In our dataset, most correlations are close to zero, suggesting that there aren't strong linear relationships between many features. However, a few notable relationships do appear:

- **Longitude and Latitude** have a weak negative correlation (-0.47), meaning as one value increases, the other tends to decrease slightly, but this is not a strong pattern.

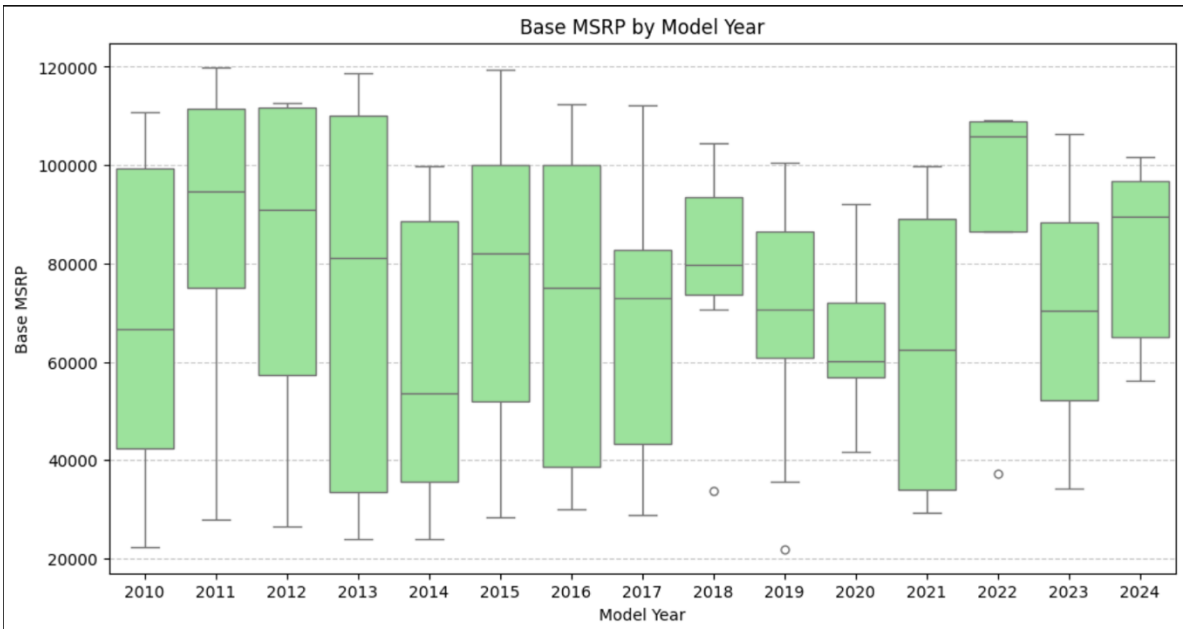
- **Model Year and Base MSRP** have a weak negative correlation (-0.23), indicating that newer models tend to have a slightly lower MSRP on average, though the relationship is not strong.

- **Electric Range and Model Year** show a negative correlation (-0.51), suggesting newer models might have a broader range variability.

**9. Data Exploration Visualizations: Create various visualizations (e.g., histograms, scatter plots, boxplots) to explore the relationships between features.**

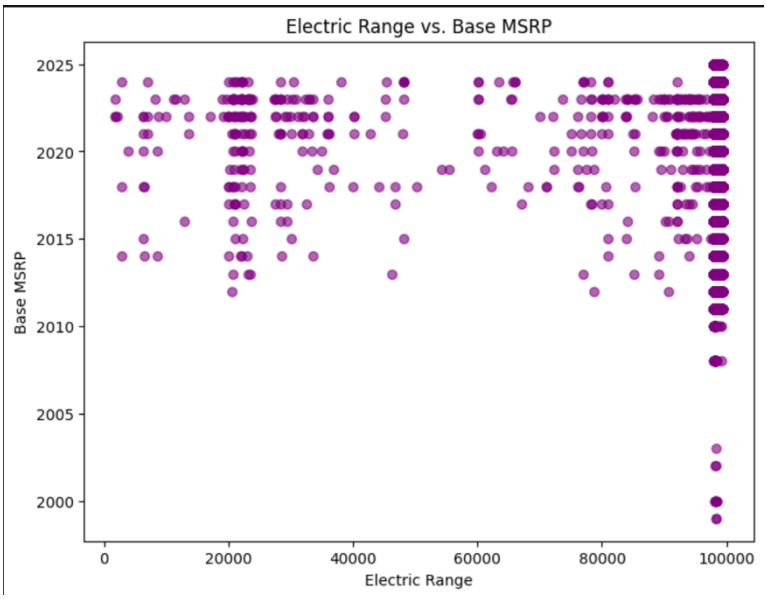



A histogram is a type of bar chart that represents the distribution of numerical data by grouping values into bins or intervals. Each bar in the histogram shows the frequency of data points that fall within that interval, giving a visual sense of how data is spread out. In this histogram, the Electric Range of electric vehicles is divided into ranges, and each bar's height indicates the number of vehicles that fall within that specific range. The high bar on the left shows that many vehicles have an electric range under 50 miles, while another noticeable peak around 200 miles shows a second common range for EVs, highlighting diversity in vehicle capabilities.




A box plot visually shows how data is spread out and whether it's skewed. The main part of the plot, the box, represents the middle 50% of the data, from the 25th percentile (lower quartile) to the 75th percentile (upper quartile). The line inside the box marks the median, or middle value, of the data. Extending from the box, the "whiskers" show the overall range, covering data within 1.5 times the box's height (the interquartile range). Points outside the whiskers, shown as circles, are outliers, meaning they are unusually high or low compared to the rest of the values.

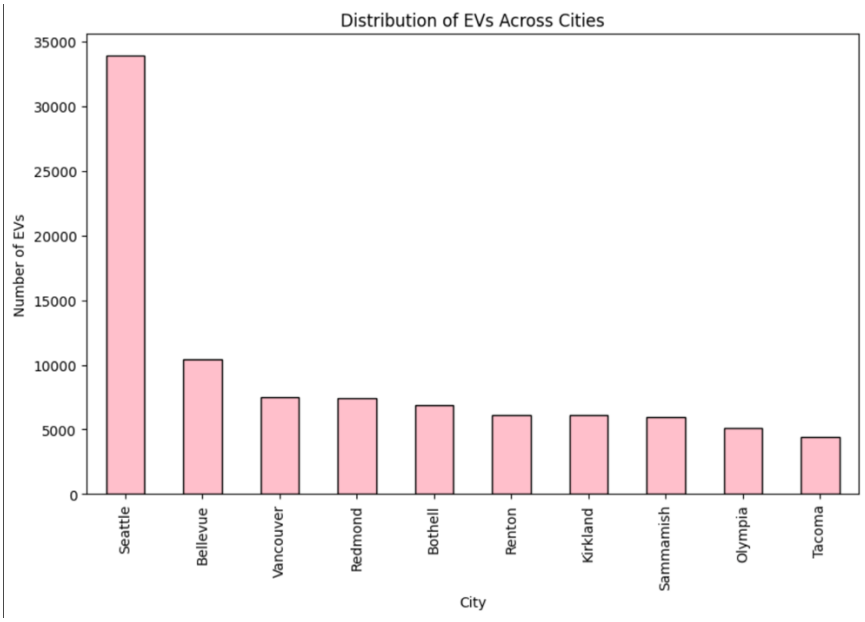
The plot reveals that some years have a wider price range (e.g., 2014, 2016), while others, like 2021, have more consistent prices. This helps us see how EV prices have changed over time and which years had more price variety.




This scatter plot displays the relationship between Electric Range and Base MSRP of electric vehicles, with each dot representing one vehicle. The x-axis shows the Electric Range (how far the vehicle can travel on a single charge), and the y-axis shows the Base MSRP (price).

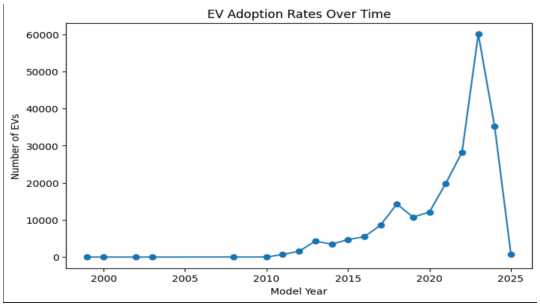
In scatter plots like this, the position of each dot provides insight into the distribution and correlation between the two variables. If the dots follow a trend, it indicates a correlation. In this case, the plot shows a spread of prices across all ranges, with no clear trend, suggesting there isn't a strong relationship between electric range and price. There are clusters of prices in certain areas, particularly for mid-range prices and ranges under 50,000.

**10. Comparative Visualization: Compare the distribution of EVs across different locations (cities, counties) using bar charts or stacked bar charts.**

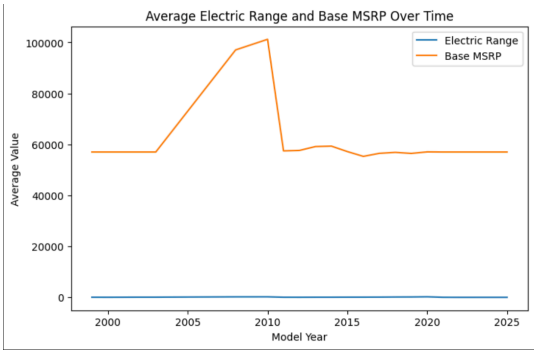



The bar chart shows that Seattle has the most electric vehicles 9EVs) by far, with over 35,000, making it the main hub for EVs among the top cities. Bellevue comes next with under 15,000 EVs, and other cities like Vancouver and Redmond have even fewer. This highlights Seattle as the leading city for EV adoption, possibly due to better infrastructure or local incentives.

**11. Temporal Analysis : If the dataset includes data across multiple time points, analyze the temporal trends in EV adoption rates and model popularity.**




EV releases surged from the 2010s onward, with a sharp rise in recent years, indicating growing popularity and production.




While the number of EVs increased, the average price (MSRP) stayed mostly steady, and the average range hasn't changed much, suggesting stable battery capacity and pricing.