Assignment #1 - Lana Musaffer||1210455 - Hala Jebreel||1210606

October 31, 2024

1. Data Cleaning and Feature Engineering

1.1. Document Missing Values & Missing Value Strategies

[58]:
```
df.isna().sum()
```

VIN (1-10)	0
County	4
City	4
State	0
Postal Code	4
Model Year	0
Make	0
Model	0
Electric Vehicle Type	0
Clean Alternative Fuel Vehicle (CAFV) Eligibility	0
Electric Range	5
Base MSRP	5
Legislative District	445
DOL Vehicle ID	0
Vehicle Location	10
Electric Utility	4
2020 Census Tract	4

The features that have missing values are: Country, City, Postal Code, Electric Range, Base MSRP, Legislative District, Vehicle Location, Electric Utility, and 2020 Census Tract.

Cleaning Data Types

[59]:
```
df.info()
```

VIN (1-10)	0
County	4
City	4
State	0
Postal Code	4
Model Year	0
Make	0
Model	0
Electric Vehicle Type	0
Clean Alternative Fuel Vehicle (CAFV) Eligibility	0
Electric Range	5
Base MSRP	5
Legislative District	445
DOL Vehicle ID	0
Vehicle Location	10
Electric Utility	4
2020 Census Tract	4

A. Converting Postal Code to Float

```
Series name: Postal Code
Non-Null Count   Dtype
--------------   -----
210161 non-null  float64
```

B. Converting Legislative District to Integer

```
Series name: Legislative District
Non-Null Count   Dtype
--------------   -----
210165 non-null  int64
```

1.1.1. Filling the missing values in Country

[62]:
```
most_frequent_county = df['County'].
    ↪mode()[0]
print("Most frequent county:",␣
    ↪most_frequent_county)
```

```
Most frequent county: King
```

Since the County feature is object type, we will fill the missing values using mode. In the above cell you can see the most frequent value in this feature which is ‘King’.

[63]:
```
df[df['County'] == 'King']
```

[63]:

107115 rows × 17 columns

Note that the data with counties equal to King are 107115 before filling the null values in this feature.

[64]:
```
df['County'] = df['County'].
    ↪fillna(most_frequent_county) #␣
    ↪Fill null values in 'County'␣
    ↪column with the most frequent␣
    ↪value

print("Null values in 'County'␣
    ↪column after fill:", df['County'].
    ↪isnull().sum()) #Check to confirm␣
    ↪null values are filled
```

```
Null values in 'County' column after␣
    ↪fill: 0
```

[65]:
```
df[df['County'] == 'King']
```

[65]:

107119 rows × 17 columns

After filling the null values with “King,” we observed that the number of rows where the county is “King” increased to 107119. Initially, there were 4 null values in the County column. Before filling, "King" was the most frequent value with 107,115 occurrences. After filling, the count of "King" rows increased to 107119.

1.1.2. Filling Missing Values in City

```
Most frequent county: Seattle
```

Since the City feature is object type, we will fill the missing values using mode. In the above cell you can see the most frequent value in this feature which is 'Seattle'.

After filling the null values with "Seattle," we observed that the number of rows where the city is "Seattle" increased to 33858. Initially, there were 4 null values in the City column. Before filling, "Seattle" was the most frequent value with 33854 occurrences. After filling, the count of "Seattle" rows increased to 33858.

1.1.3. Filling Missing Values in Postal Code

To handle missing values in the Postal Code column, we filled them with the median postal code for each respective county.

```
[72]: df['Postal Code'] = df.
 ↪groupby('County')['Postal Code'].
 ↪transform(lambda x: x.fillna(x.
 ↪median()))

df['Postal Code'].isna().sum()

df[df['Postal Code'].isnull()]
```

```
[72]: Empty DataFrame
Index: []
```

After filling the missing Postal Code values with the median postal code for each county, we checked to see if any null values remained. The line df['Postal Code'].isna().sum() shows how many missing values are left, if any. Displaying df[df['Postal Code'].isnull()] helps us find any rows where Postal Code is still empty, making sure our data is complete.

1.1.4. Filling Missing Values in Electric Range

To handle missing values in the Electric Range column, we filled them with the median Electric Range for each respective county.

```
[73]: Empty DataFrame
Index: []
```

After filling the missing Electric Range values with the median range for each county, we checked to see if any null values remained. The line `df['Electric Range'].isna().sum()` shows how many missing values are left, if any. Displaying `df[df['Electric Range'].isnull()]` helps us find any rows where Electric Range is still empty, ensuring our data is complete.

1.1.5. Filling Missing Values in Base MSRP

To handle missing values in the Base MSRP column, we filled them with the median Base MSRP for each respective county.

```
[16]: Empty DataFrame
Index: []
```

After filling the missing Base MSRP values with the median range for each county, we checked to see if any null values remained. The line df['Base MSRP'].isna().sum() shows how many missing values are left, if any. Displaying df[df['Base MSRP'].isnull()] helps us find any rows where Base MSRP is still empty, ensuring our data is complete.

1.1.6. Filling Missing Values in Vehicle Location

```
Most frequent county: POINT (-122.
 ↪13158 47.67858)
```

Since the Vehicle Location feature is object type, we will fill the missing values using mode. In the above cell you can see the most frequent value in this feature which is 'POINT (-122.13158 47.67858)'.

```
Null values in 'Vehicle Location'␣
 ↪column after fill: 0
```

1.1.7. Filling Missing Values in Electric Utility

```
Most frequent county: PUGET SOUND␣
 ↪ENERGY INC||CITY OF TACOMA - (WA)
```

```
Null values in 'Electric Utility'␣
 ↪column after fill: 0
```

Since the Electric Utility feature is object type, we will fill the missing values using mode. In the above cell you can see the most frequent value in this feature which is 'PUGET SOUND ENERGY INC||CITY OF TACOMA - (WA)'.

1.1.8. Filling Missing Values in 2020 Census Tract

To handle missing values in the 2020 Census Tract column, we filled them with the median 2020 Census Tract for each respective county.

```
[22]: Empty DataFrame
Index: []
```

After filling the missing 2020 Census Tract values with the median range for each county, we checked to see if any null values remained. The line df['2020 Census Tract'].isna().sum() shows how many missing values are left, if any. Displaying df[df['2020 Census Tract'].isnull()] helps us find any rows where 2020 Census Tract is still empty, ensuring our data is complete.

The results after filling all the missing values:

```
[23]: df.isna().sum()
```

```
[23]:
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
```

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

In this section, we applied one-hot encoding to the categorical features (VIN (1-10), County, City, etc.) to convert them into numerical format. Using pd.get_dummies, each category in these columns is transformed into separate binary columns, making the data suitable for machine learning models. The resulting df_encoded DataFrame shows the encoded data with the first few rows displayed using head(). Here's a snippet of the result:

VIN (1-10)_1C4JJXN60P	VIN (1-10)_1C4JJXN61P	VIN (1-10)_1C4JJXN62P	...	Electric Utility_PORTLAND GENERAL ELECTRIC CO	Electric Utility_PUD NO 1 OF CHELAN COUNTY	Electric Utility_PUD NO 1 OF DOUGLAS COUNTY	Electric Utility_PUD NO 1 OF OKANOGAN COUNTY	Electric Utility_PUD NO 1 OF PEND OREILLE COUNTY	Electric Utility_PUD NO 1 OF WHATCOM COUNTY	Electric Utility_PUD NO 2 OF GRANT COUNTY	Electric Utility_PUGET SOUND ENERGY INC
False	False	False	...	False	False	False	False	False	False	False	True
False	False	False	...	False	False	False	False	False	False	False	True
False	False	False	...	False	False	False	False	False	False	False	True
False	False	False	...	False	False	False	False	False	False	False	True
False	False	False	...	False	False	False	False	False	False	False	False

After applying one-hot encoding to the categorical features, each unique category within these columns has been transformed into a separate binary column. In the encoded DataFrame, a value of 1 indicates the presence of that specific category for a row, while 0 indicates its absence. For example, in the Electric Utility columns, Electric Utility_PUGET SOUND ENERGY INC being True shows that this utility serves the corresponding vehicle.

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

In this section, we used Min-Max Scaling to normalize the numerical features (Model Year, Electric Range, Base MSRP, etc.), transforming their values to a range between 0 and 1. This step helps ensure that all numerical features are on a similar scale, which can improve model performance and convergence for certain algorithms. The normalized values are stored back in df_encoded, and we display the first few rows to verify the changes.

	Postal Code	Model Year	Electric Range	Base MSRP	Legislative District	DOL Vehicle ID	2020 Census Tract
0	98380.0	0.846154	0.089021	0.0	0.714286	0.559050	0.945730
1	98370.0	0.730769	0.637982	0.0	0.469388	0.993024	0.945730
2	98012.0	0.653846	0.044510	0.0	0.020408	0.212763	0.946202
3	98310.0	0.730769	0.637982	0.0	0.469388	0.989794	0.945730
4	98052.0	0.769231	0.445104	0.0	0.918367	0.993932	0.945693

Note that after the normalization the numerical features became between 0 to 1.

2. Exploratory Data Analysis

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

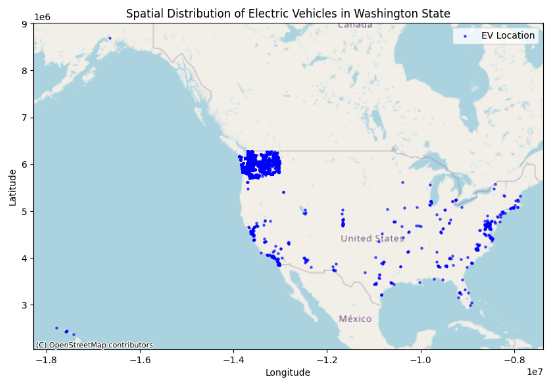
	Mean	Median	Standard Deviation
Model Year	2.021049e+03	2.022000e+03	2.988941e+00
Electric Range	5.060107e+01	0.000000e+00	8.697251e+01
Base MSRP	8.976769e+02	0.000000e+00	7.653589e+03
Legislative District	2.886870e+01	3.200000e+01	1.495185e+01
DOL Vehicle ID	2.290774e+08	2.405164e+08	7.115519e+07
2020 Census Tract	5.297929e+10	5.303303e+10	1.551466e+09

- Model Year: The mean model year is around 2021, with a standard deviation of approximately 2.99, indicating that most vehicles are from recent years, but there is some variation.
- Electric Range: The mean electric range is about 50 miles, with a high standard deviation (86.97), suggesting a wide variation in range. The median of 0 indicates that many entries may have missing or zero values for this feature.
- Base MSRP: The mean base MSRP is approximately $897.67, with a standard deviation of

765.36, indicating a substantial range in vehicle prices.

- Legislative District: The mean and standard deviation for legislative districts suggest a fairly even spread across districts.
- DOL Vehicle ID and 2020 Census Tract: These identifiers have large values due to the nature of IDs, and they have high standard deviations, which is expected as they are unique identifiers rather than typical numeric variables.

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



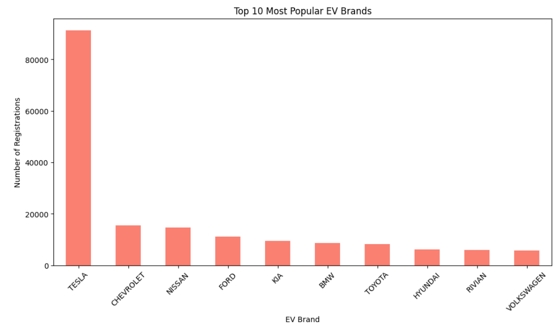
The map reveals a dense concentration of electric vehicles (EVs) primarily in the western region of Washington State, particularly around the Puget Sound area. This region includes major cities like Seattle, Tacoma, and Bellevue, which are known for having well-established infrastructure and policies supporting EV adoption. The high density of points here suggests that urban and suburban areas in western Washington, especially near Seattle, lead in EV registrations. In contrast, more rural or eastern parts of Washington show fewer data points, indicating a lower EV adoption rate in those areas. This pattern may be due to the availability of charging infrastructure, population density, and local incentives supporting EV use in more urbanized areas of the state.

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

2.3.1. Analyze Model Popularity

To see the most popular models we used ‘value$_{counts}()$‘

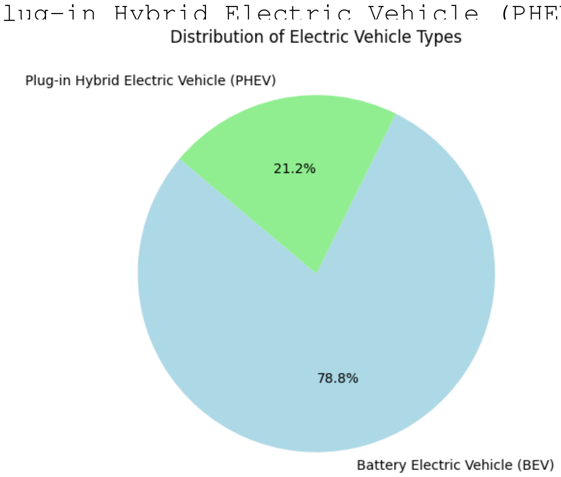
```
Top 10 EV Models by Popularity:
Model
MODEL Y            44038
MODEL 3            32520
LEAF               13606
MODEL S             7795
BOLT EV             6780
MODEL X             6239
VOLT                4815
ID.4                4716
MUSTANG MACH-E      4363
WRANGLER            4116
```



The bar chart highlights the top 10 EV brands by registration count, with Tesla leading by a large margin at over 91,000 registrations. This dominance underscores Tesla’s strong position in the EV market, driven by its technological innovations, brand loyalty, and extensive charging network. Following Tesla, Chevrolet and Nissan rank second and third, with 15,419 and 14,721 registrations, respectively, reflecting their established presence in the EV space through models like the Chevrolet Bolt and Nissan Leaf. Other brands, such as Ford, Kia, BMW, and Toyota, have moderate representation, indicating growing competition in the market but still trailing significantly behind Tesla. Newer entrants like Rivian also appear in the top 10, showing emerging interest in alternative EV brands.

2.3.3. Analyze Popularity by Electric Vehicle Type

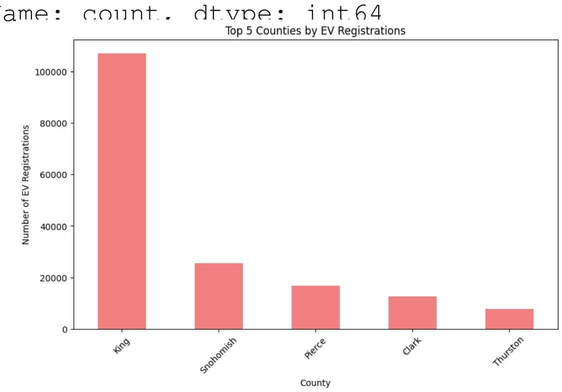
```
Electric Vehicle Type Distribution:
Electric Vehicle Type
Battery Electric Vehicle (BEV)                 165554
Plug-in Hybrid Electric Vehicle (PHEV)          44611
```



The pie chart shows the distribution of electric vehicle types, with Battery Electric Vehicles (BEVs) making up the majority at 78.8% and Plug-in Hybrid Electric Vehicles (PHEVs) accounting for 21.2%. This indicates a strong preference for fully electric vehicles over hybrids among consumers, likely due to advancements in battery technology, longer range, and the expansion of charging infrastructure, which make BEVs more practical. The smaller share of PHEVs suggests that while they offer flexibility with both fuel and electric power, they are less popular as more consumers shift towards fully electric options.

2.3.4. Analyze Regional Popularity

```
Top 5 Counties by EV Registrations:
County
King          107119
Snohomish      25392
Pierce         16677
Clark          12537
Thurston        7719
Name: count, dtype: int64
```

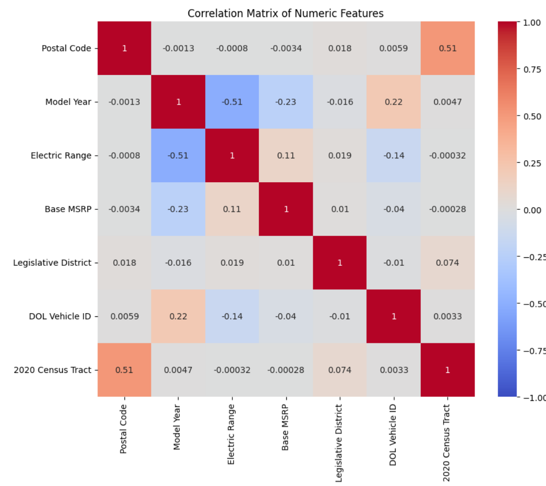


The bar chart shows the top 5 counties in Washington by electric vehicle (EV) registrations. King County leads by a substantial margin, with over 107,000 registrations, far surpassing other counties. This high concentration is likely due to King County's urban population, which includes Seattle, as well as robust EV infrastructure and incentives that support EV adoption. Snohomish and Pierce counties follow, with around 25,000 and 16,000 registrations respectively, indicating moderate adoption in neighboring regions. Clark and Thurston counties have fewer registrations but still show a notable presence of EVs. This distribution suggests that urban and suburban areas with developed infrastructure and population density drive higher EV adoption rates in the state.

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

Note that correlation values range from -1 to 1, where values close to 1 indicate a strong positive correlation and values close to -1 indicate a strong negative correlation. Specifically, values between 0 and 0.3 (or 0 and -0.3) suggest a weak correlation, indicating little to no linear relationship between the variables. Values between 0.3 and 0.7 (or -0.3 and -0.7) indicate a moderate correlation, which suggests a noticeable, though not perfect, linear relationship. Finally, values between 0.7 and 1.0 (or -0.7 and -1.0) signify a strong correlation, where one variable is highly predictive of the other.

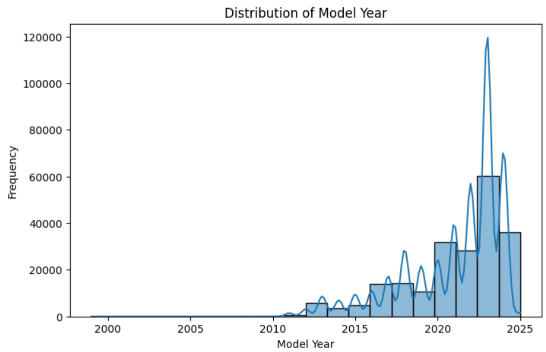
```
Correlation Matrix:
                      Postal Code  Model Year  Electric Range  Base MSRP  \
Postal Code              1.000000   -0.001291       -0.000800  -0.003409
Model Year              -0.001291    1.000000       -0.513540  -0.230651
Electric Range          -0.000800   -0.513540        1.000000   0.114155
Base MSRP               -0.003409   -0.230651        0.114155   1.000000
Legislative District     0.017785   -0.016191        0.018689   0.010035
DOL Vehicle ID           0.005861    0.215703       -0.140695  -0.039501
2020 Census Tract        0.508744    0.004710       -0.000324  -0.000283

                      Legislative District  DOL Vehicle ID  2020 Census Tract
Postal Code                       0.017785        0.005861           0.508744
Model Year                       -0.016191        0.215703           0.004710
Electric Range                    0.018689       -0.140695          -0.000324
Base MSRP                         0.010035       -0.039501          -0.000283
Legislative District              1.000000       -0.010241           0.074487
DOL Vehicle ID                   -0.010241        1.000000           0.003347
2020 Census Tract                 0.074487        0.003347           1.000000
```



- Model Year and Electric Range: There is a moderate negative correlation (-0.51), suggesting that older EV
- models generally have a shorter electric range compared to newer models. This reflects advancements in battery technology over time.
- Model Year and DOL Vehicle ID: A moderate positive correlation (0.22) exists here, as expected, since newer vehicles likely have higher ID numbers, reflecting the sequential nature of registrations.
- Model Year and Base MSRP: There is a weak negative correlation (-0.23), indicating that newer models might be slightly less expensive on average, which could be due to decreasing EV costs over time.
- Postal Code and 2020 Census Tract: There is a moderate positive correlation (0.51), suggesting some alignment between these two geographic identifiers. This is logical since census tracts often correspond to specific postal codes.
- Electric Range and Base MSRP: A weak positive correlation (0.11) is observed, implying that EVs with a higher range tend to have a slightly higher price, but the relationship is not strong.

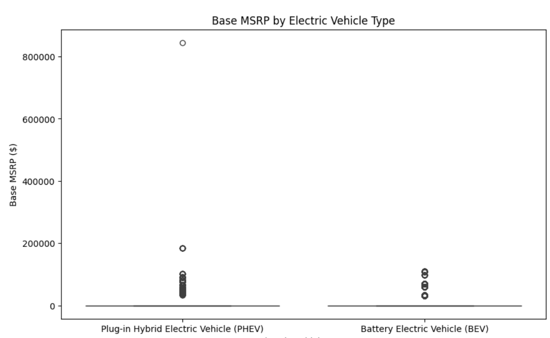
Most of the other relationships show very weak or near-zero correlations, indicating little to no linear relationship between those feature pairs. Overall, the matrix highlights a few moderate correlations related to model year, range, and geographic identifiers, which are in line with expectations for these types of features in an EV dataset.

3.1. VISUALIZATION



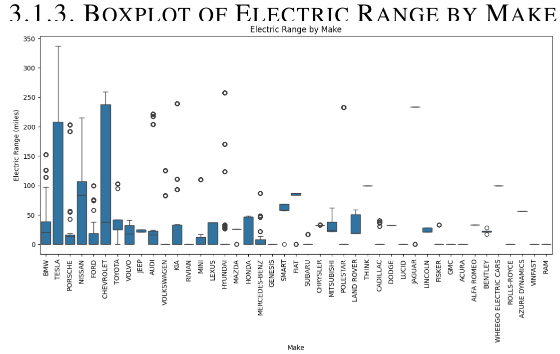
The histogram shows the distribution of electric vehicle (EV) model years in the dataset. There is a clear upward trend, with very few registrations for older models (before 2010) and a sharp increase in frequency starting around 2015. This trend indicates that most EVs in the dataset are relatively recent models, with a peak around 2022. The significant increase in newer models likely reflects growing adoption of EVs in recent years due to advancements in technology, increased availability, and possibly more incentives and support for EV purchases. The small drop-off toward 2025 may reflect incomplete data for that year, as newer models may still be in the process of registration.

3.1.2. BOXPLOT OF BASE MSRP BY ELECTRIC



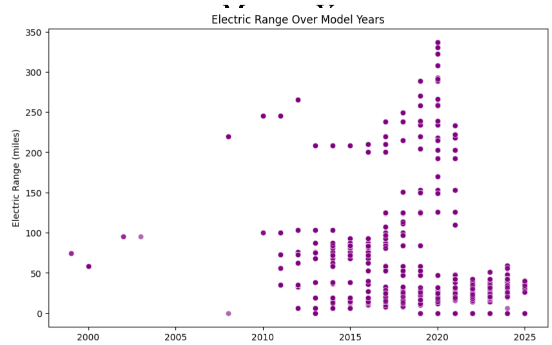
The boxplot compares the base MSRP of Plug-in Hybrid Electric Vehicles (PHEVs) and Battery Electric Vehicles (BEVs). Most vehicles in both categories have base MSRPs clustered toward the lower end of the scale, indicating that the majority of EVs are relatively affordable. However, there are a few notable outliers, particularly among BEVs, with prices reaching up to $800,000. This suggests that while the typical price for both vehicle types is within a similar range, some BEVs are significantly more expensive, likely due to high-end or luxury models. The plot highlights that BEVs have a wider range in pricing, reflecting a diversity of models and price points in the fully electric category.

3.1.3. BOXPLOT OF ELECTRIC RANGE BY MAKE



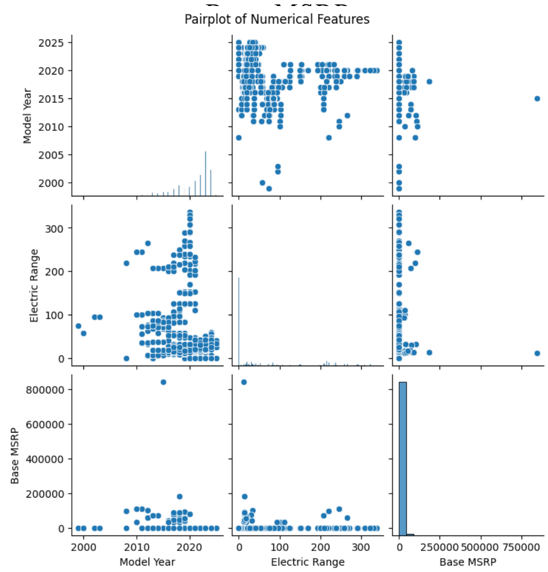
The boxplot displays the distribution of electric range by vehicle make. Tesla stands out with a high median range and a wide range of values, reaching up to 300 miles, highlighting its focus on long-range electric vehicles. Other brands like Chevrolet, Ford, and BMW also show notable ranges, though their distributions are generally lower than Tesla's, with fewer high-range outliers. Many brands, such as Fiat, Hyundai, and Volkswagen, have lower median ranges and tighter distributions, indicating that their models typically offer shorter ranges. This plot illustrates that Tesla dominates in terms of electric range, while other brands provide a mix of shorter-range options, possibly targeting different segments of the market. The presence of outliers in several brands suggests a few models with extended ranges, though these are not as common.

3.1.4. SCATTER PLOT FOR ELECTRIC RANGE



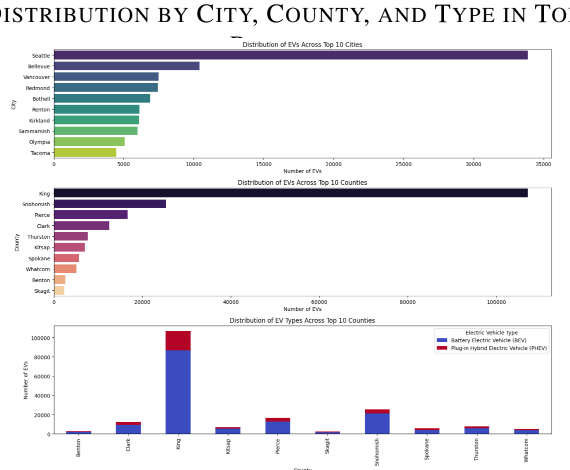
The scatter plot illustrates the relationship between electric vehicle (EV) model year and electric range. There is a clear upward trend in electric range over time, especially from around 2015 onward. Newer model years tend to have higher ranges, reflecting advancements in battery technology and increased focus on longer-range EVs. The clustering of data points below 100 miles for earlier model years suggests that older EVs generally had shorter ranges. In contrast, recent years show more models achieving ranges above 200 miles, with some reaching up to 300 miles or more. This trend highlights the progress made in EV technology to meet consumer demand for greater range.

3.1.5. Pairplot for Exploring Relationships Between Model Year, Electric Range, and Base MSRP



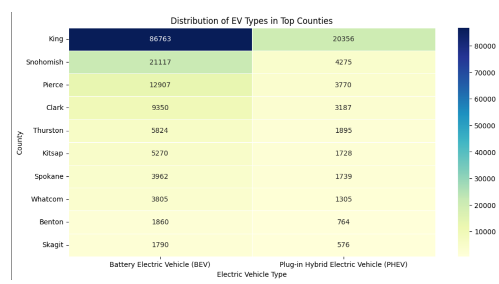
This pairplot shows the relationships between Model Year, Electric Range, and Base MSRP for electric vehicles. Newer model years tend to have higher electric ranges, indicating advancements in battery technology and a focus on extended range in recent EV models. While the Base MSRP slightly increases with newer model years, this relationship is weak, possibly due to competitive pricing strategies. The scatter between Electric Range and Base MSRP suggests that vehicles with longer ranges generally cost more, as they likely require more advanced batteries. The individual feature distributions reveal that most EVs are concentrated in recent years (post-2015), have ranges below 100 miles, and are priced at lower MSRPs, with a few high-priced outliers likely representing luxury models.

3.1.6. Stacked Bar Plot for Electric Vehicle Distribution by City, County, and Type in Top



This figure shows the distribution of electric vehicles (EVs) across the top 10 cities and counties in Washington, along with the breakdown of EV types in these counties. Seattle and King County lead by a large margin, with Seattle having over 30,000 EVs and King County exceeding 100,000, indicating high EV adoption in urbanized areas. Other cities like Bellevue and counties like Snohomish and Pierce also have notable EV counts but lag significantly behind Seattle and King County. The stacked bar chart reveals a strong preference for Battery Electric Vehicles (BEVs) over Plug-in Hybrid Electric Vehicles (PHEVs) in these regions, particularly in King County. This distribution suggests that Washington's urban centers are the primary drivers of EV adoption, with a clear inclination towards fully electric models.

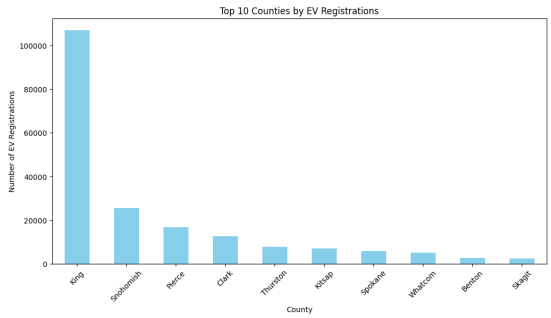
3.1.7. Heatmap for Electric Vehicle Type vs. County



This heatmap shows the distribution of electric vehicle (EV) types—Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs)—across the top counties in Washington by EV registrations. King County has the highest numbers, with 86,763 BEVs and 20,356 PHEVs, reflecting a strong preference for fully electric vehicles in this urbanized region. Snohomish and Pierce counties also have significant BEV counts (21,117 and 12,907, respectively) with smaller PHEV numbers, indicating a similar preference pattern. Other counties like Clark and Thurston have fewer EVs overall but maintain a higher proportion of BEVs. The heatmap reveals that BEVs are more popular than PHEVs across these top counties, particularly in urban centers where infrastructure and incentives likely support fully electric vehicles.

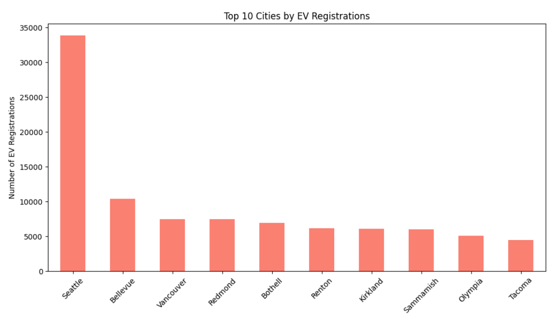
3.2. Comparative Visualization: Compare the distribution of EVs across different locations (cities, counties) using bar charts or stacked bar charts.

3.2.1. Distribution of EVs by County



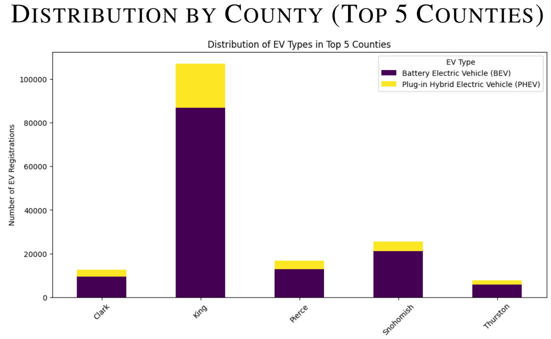
The bar chart shows that King County has the highest number of EV registrations in Washington, with over 100,000, far surpassing other counties. Snohomish and Pierce follow with much lower numbers, indicating that EV adoption is concentrated in King County, likely due to its urban infrastructure. Counties like Clark, Thurston, and Kitsap show moderate EV adoption, while the rest have relatively low counts, positioning King County as a central hub for EV ownership in the state.

3.2.2. Distribution of EVs by City



The bar chart shows that Seattle leads in EV registrations in Washington with over 35,000, followed by Bellevue with significantly fewer. Cities like Vancouver, Redmond, and Bothell show moderate EV adoption, while Renton, Kirkland, Sammamish, Olympia, and Tacoma have lower but notable counts.

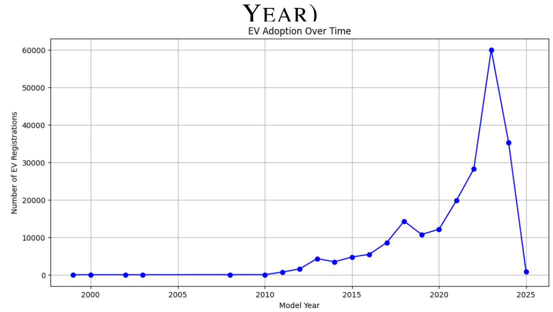
3.2.3. Stacked Bar Chart for EV Type Distribution by County (Top 5 Counties)



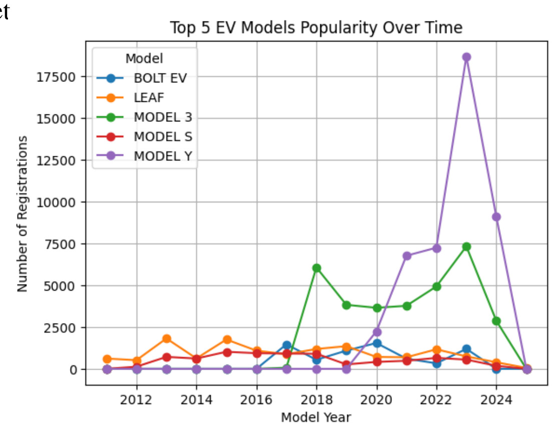
The bar chart shows that in the top 5 counties for EV registrations, there is a strong preference for Battery Electric Vehicles (BEVs) over Plug-in Hybrid Electric Vehicles (PHEVs), with King County leading in registrations. Snohomish and Pierce counties follow a similar pattern, while Clark and Thurston have lower overall numbers but maintain a consistent BEV-to-PHEV ratio.

4. Temporal Analysis

4.1. EV Adoption Over Time (using Model Year)



The plot shows how the number of electric vehicle (EV) registrations has changed over time by model year. Starting around 2010, EV registrations gradually increased, with a big jump after 2018, reaching a peak in recent years. This growth likely reflects more people choosing EVs, better technology, and support for EVs through incentives. The drop at the end may be because data for the latest models or 2025 isn't fully recorded yet



The plot shows the popularity of the top 5 electric vehicle (EV) models over time by model year. Tesla's Model Y leads with a sharp increase around 2023, followed by Model 3, which also saw a peak in recent years. Nissan Leaf and Chevrolet Bolt EV have more stable but lower numbers, indicating steady but less dominant popularity. The Tesla Model S shows moderate popularity, increasing around 2018 but remaining below Model Y and Model 3 levels. This trend highlights Tesla's strong hold on the EV market in recent years, especially with Model Y and Model 3. The decline at the end likely reflects incomplete data for the latest model years.