BIRZEIT UNIVERSITY

Electrical and Computer Engineering Department
Machine Learning and Data Science - ENCS5341
Assignment #1

Prepared By: Tala Abahra 1201002, Sondos Shahin 1200166
Instructor: Dr. Yazan Abu Farha Date: Oct 30, 2024
Topic: Machine Learning Assignment: 1
https://github.com/tabahra2015/Electric_Vehicle_Population_Data

---

Overview

Data preprocessing involves several essential steps that prepare raw data for analysis and modeling. The first step is **data collection,** gathered from various sources such as databases, spreadsheets, or online repositories. Next, **data cleaning** is performed to handle missing values by identifying and managing them through deletion, imputation, or other strategies. This step also includes removing duplicates to maintain data integrity and correcting errors such as typos or formatting inconsistencies.

This dataset pertains to electric vehicles and includes various features that provide insights into their specifications and classifications. The missing values and frequency and percentage for each feature are documented, highlighting the extent of data gaps.

Data cleaning

1. Document Missing Values:

The initial step in data cleaning is documenting missing values. This involves identifying and quantifying missing data in the dataset. Understanding the extent and distribution of missing values is essential for evaluating data quality. By classifying the types of missing data researchers can determine the best strategies for imputation or removal, ensuring the dataset is ready for analysis or modeling.

In the dataset on electric vehicles, several features exhibit missing values, which could impact data quality and analysis. The **Legislative District** feature has the highest number of missing values at 445 (0.211738%), making it crucial for understanding local policies' impact on electric vehicle adoption. **Electric Range** and **Base MSRP** each have 5 missing values (0.002379%), and their absence could distort comparisons and hinder pricing assessments. The **Vehicle Location** feature has 10 missing values (0.004758%), which limits insights into the geographic distribution of electric vehicles. Other features like **County**, **City**, **Postal Code**, and **Electric Utility** each have 4 missing values (0.001903%), potentially affecting regional demographic analysis. Overall, addressing these missing values through imputation or removal is essential to ensure data integrity and reliability in subsequent analyses.

2.Missing Value Strategies and Descriptive Statistics:

When comparing the impact of different missing value strategies on the analysis of numerical data, each method can influence the results in distinct ways:

Mean Imputation: Replacing missing values with the mean can be effective when the data is normally distributed, as it maintains the overall average of the dataset. However, this approach can distort the analysis for skewed data by pulling the central tendency toward the mean, potentially masking the presence of outliers. For example, using the mean for **Postal Code** (98178) and **Electric Range** (51) could lead to unrealistic data points if there are significant outliers or cluster values around extremes.

Median Imputation: This method is more robust against outliers, making it suitable for skewed data. In cases like **Electric Range** and **Base MSRP**, where the median is 0, median imputation ensures that the dataset reflects the central tendency of most records without being affected by extreme values. However, it may underestimate the variability in the data, leading to less nuanced insights, especially if outliers are important.

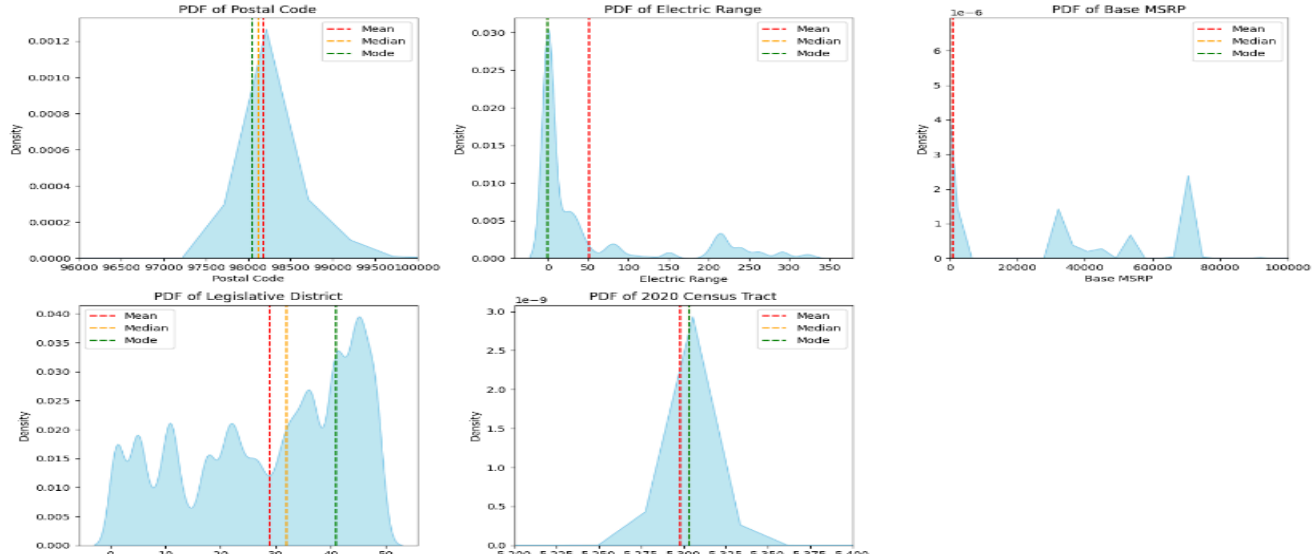
Mode Imputation: Using the most frequent value is effective for fields that behave like categorical variables, even if they are numerical. For example, the mode for the **Legislative District** is 41, and for the **Postal Code**, it's 98052. While this approach can be useful for filling in common values, it risks oversimplifying the dataset by reducing variability, especially in fields where a range of values is expected. In unique identifiers like **Postal Code**, mode imputation could obscure geographic diversity in the data.

Overall, the choice of strategy has a significant impact on the analysis. **Mean imputation** can lead to biased results in the presence of outliers, **median imputation** provides stability but may hide important variations, and **mode imputation** simplifies the data, potentially obscuring key distinctions. Researchers should select the method that best fits the data's distribution and the importance of preserving its variability to ensure accurate, meaningful analysis.

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

Missing Value Strategies:
Summary Statistics for Columns with Missing Values:
                                Mean        Median          Mode  \
Postal Code                    98178         98125         98052
Electric Range                    51             0             0
Base MSRP                        898             0             0
Legislative District              29            32            41
2020 Census Tract     52979294366  53033030101  53033028200

                      Standard Deviation
Postal Code                         2445
Electric Range                        87
Base MSRP                           7654
Legislative District                  15
2020 Census Tract             1551466456
```



The analysis of **Non-numerical** missing values in the **County**, **City**, **Vehicle Location**, and **Electric Utility** columns shows that using mode imputation (replacing missing values with the most common one) had little impact. For example, the missing value percentage in **County** changed by just 0.0009%, and **City** by 0.0016%. Even for **Vehicle Location**, the change was only 0.0046%. Mode imputation helped fill in missing data without significantly altering the dataset. When rows with missing values were dropped, the impact was also small. The percentage of dropped rows was only 0.0019% for **County** and **City**, and 0.0048% for **Vehicle Location**.

```
Impact of Missing Value Handling on 'County' Column:
Missing Values Before:  4
Mode Value:  King
Percentage Before:  50.96806733884974
Percentage After:  50.96900054718911
different :  0.0009332083393687185
Dropped Rows Percentage:  0.001903266481098185

Impact of Missing Value Handling on 'City' Column:
Missing Values Before:  4
Mode Value:  Seattle
Percentage Before:  16.10860245240554
Percentage After:  16.110199129255584
different :  0.0015966768500454975
Dropped Rows Percentage:  0.001903266481098185

Impact of Missing Value Handling on 'Vehicle Location' Column:
Missing Values Before:  10
Mode Value:  POINT (-122.13158 47.67858)
Percentage Before:  2.501011158430682
Percentage After:  2.5056503223657605
different :  0.004639163935078461
Dropped Rows Percentage:  0.004758166202745461

Impact of Missing Value Handling on 'Electric Utility' Column:
Missing Values Before:  4
Mode Value:  PUGET SOUND ENERGY INC||CITY OF TACOMA - (WA)
Percentage Before:  36.45110177435395
Percentage After:  36.45231127923299
different :  0.0012095048790357055
Dropped Rows Percentage:  0.001903266481098185
```

Comparing mode imputation and dropping rows shows minimal impact on the dataset. **Mode** imputation filled missing values without reducing the dataset size, keeping it complete for analysis. **Dropping rows** slightly reduced the dataset, which can lead to loss of data. Mode imputation is often better for preserving the full dataset, while dropping rows may be useful when missing data is significant.

3.Feature Encoding :

To prepare the data for machine learning, categorical features such as **Make**, **Model**, **Electric Vehicle Type**, **Clean Alternative Fuel Vehicle (CAFV) Eligibility**, and **Electric Utility** were encoded using one-hot encoding. This technique transforms each category within these features into separate binary columns, facilitating their integration into predictive models. For instance, the **Make** feature contains **42 unique manufacturers**, while the **Model** feature has **152 unique models**. The **Electric Vehicle Type** has **1 category**, and **CAFV Eligibility** includes **2 categories** indicating eligibility status. Additionally, the **Electric Utility** feature comprises **73 unique utility providers**.

```
One-Hot Encoding for Clean Alternative Fuel Vehicle (CAFV) Eligibility:
        Clean Alternative Fuel Vehicle (CAFV) Eligibility_Eligibility unknown
as battery range has not been researched  \
0                                                 False
1                                                 False
2                                                 False
3                                                 False
4                                                 False
…                                                   …
210160                                            False
210161                                            False
210162                                            False
210163                                             True
210164                                             True

        Clean Alternative Fuel Vehicle (CAFV) Eligibility_Not eligible due to
low battery range
0                                                 False
1                                                 False
2                                                  True
3                                                 False
4                                                 False
…                                                   …
210160                                             True
210161                                             True
210162                                            False
210163                                            False
210164                                            False

[210165 rows x 2 columns]
Number of columns after encoding for Clean Alternative Fuel Vehicle (CAFV)
```

Applying one-hot encoding is crucial because many machine learning algorithms cannot directly interpret categorical variables, which can lead to inaccurate predictions or model performance issues. By converting categorical data into a numerical format, one-hot encoding allows models to capture the relationships and patterns inherent in the data, ultimately improving predictive accuracy and enabling more effective analysis of electric vehicle trends and behaviors.

4.Normalization:

The dataset contains two features: Electric Range and Base MSRP, with a total of five instances. Normalization is deemed necessary due to the highly skewed distribution of the Electric Range, which clusters most values near the lower end, and the wide range of Base MSRP, which includes extreme outliers.

After capping the values, the Electric Range has a maximum of 337.0 and a minimum of 0.0, while the Base MSRP ranges from a maximum of 845,000.0 to a minimum of 0.0.

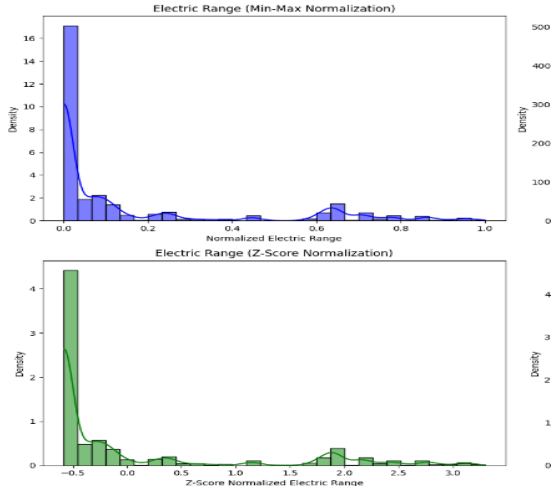
Z-score normalization was applied to both features, yielding the following formulas: for Electric Range, the Z-score is calculated as Z=(Electric Range−50.60)/86.97, and for Base MSRP, it is Z=(Base MSRP−897.68)/7653.59 A snapshot of the normalized data illustrates the Electric Range and Base MSRP values alongside their normalized counterparts, providing a clearer view of the data distribution.

```
Max and Min Values after Capping:
Electric Range - Max: 337.0, Min: 0.0
Base MSRP - Max: 845000.0, Min: 0.0

Z-score Calculations for Electric Range and Base MSRP:
Electric Range Z-score: z = (Electric Range - 50.60) / 86.97
Base MSRP Z-score: z = (Base MSRP - 897.68) / 7653.59
```

	Electric Range	Base MSRP	Electric Range MinMax	Base MSRP MinMax
0	30.0	0.0	0.089021	0.0
1	215.0	0.0	0.637982	0.0
2	15.0	0.0	0.044510	0.0
3	215.0	0.0	0.637982	0.0
4	150.0	0.0	0.445104	0.0

	Electric Range ZScore	Base MSRP ZScore
0	-0.236880	-0.117288
1	1.890211	-0.117288
2	-0.409347	-0.117288
3	1.890211	-0.117288
4	1.142855	-0.117288



Exploratory Data Analysis

Exploratory analysis approach is usually done on datasets that we don't have a clear understanding about, in order to find unknown relationships in the data. It is usually done using visualization techniques, by plotting the data in different ways to try and find patterns in it.

Descriptive Statistics:

Numerical features in the data set can be described statistically using the mean, median and the standard deviation. These metrics provide information about the distribution of the data in the dataset, and are used in summarizing and analyzing the data. Computing these metrics provides a center which the data is distributed around, in addition to how far away the data is distributed about that center.

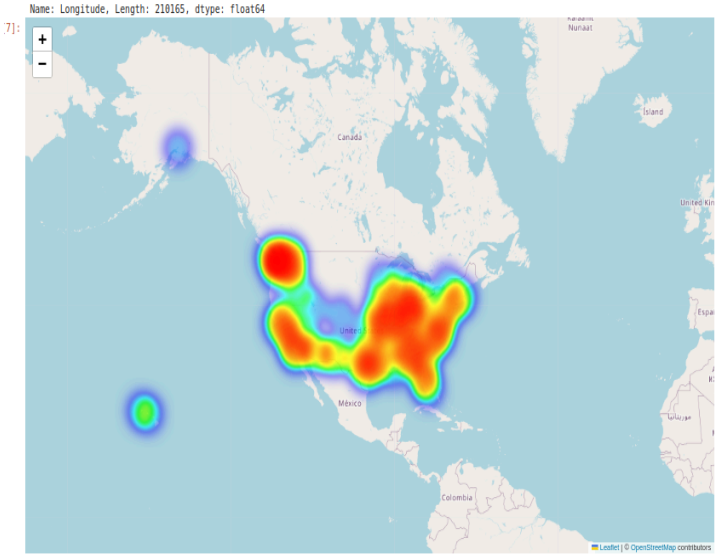
Summary Statistics:

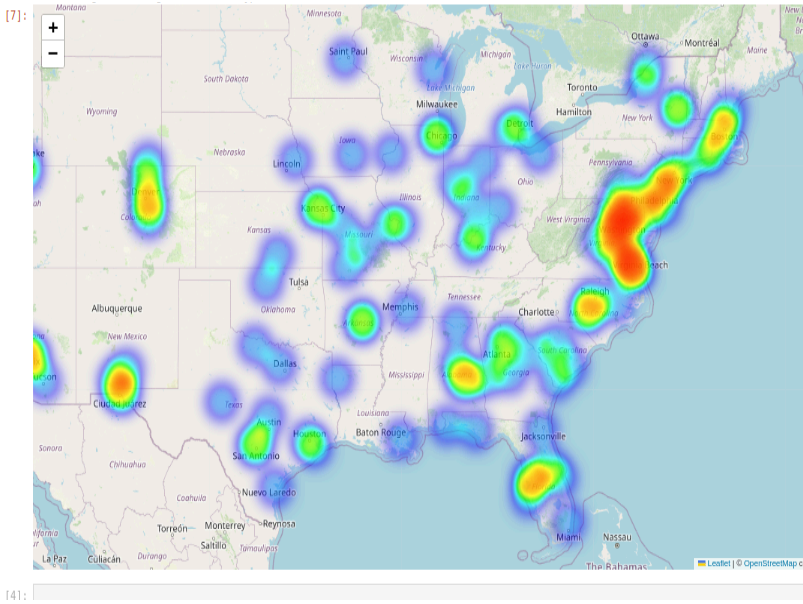
	Mean	Median	Mode	Standard Deviation
Model Year	2021	2022	2023	3
Electric Range	51	0	0	87
Base MSRP	898	0	0	7654
Legislative District	29	32	41	15
2020 Census Tract	-2147483648	-2147483648	-2147483648	1551466456

Mean, median, mode and standard deviation values were calculated for the numerical features in our data set.

Spatial Distribution:

Spatial visualization such as geo maps are used to visualize the distribution of data in relation to their geographical area. In our dataset, the Vehicle Location feature can be visualized on a map to show the distribution of different electrical vehicle models in different locations.



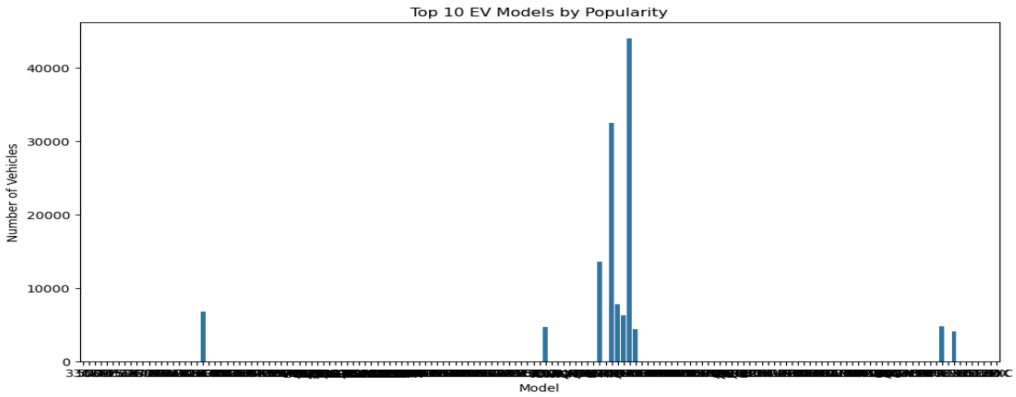


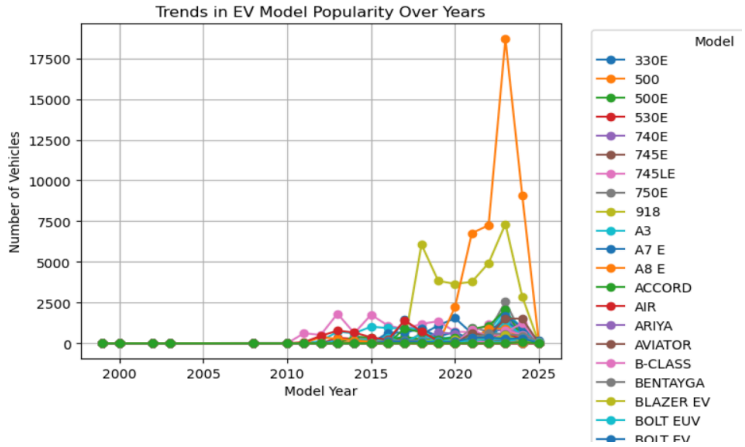
[4]:

Model Popularity:

The first analysis technique to find trends in the data from the dataset was to count the number of electric vehicles from each model, and compare them to find the most popular models over years. The top 10 models over all years were displayed on a bar chart for simplicity. A bar chart was chosen since the analysis is of a categorical feature (Model) and a numerical one (number of vehicles).

	Model	Count
0	MODEL Y	44038
1	MODEL 3	32520
2	LEAF	13606
3	MODEL S	7795
4	BOLT EV	6780
..	...	...
148	750E	2
149	745LE	2
150	FLYING SPUR	2
151	VF 8	1
152	918	1

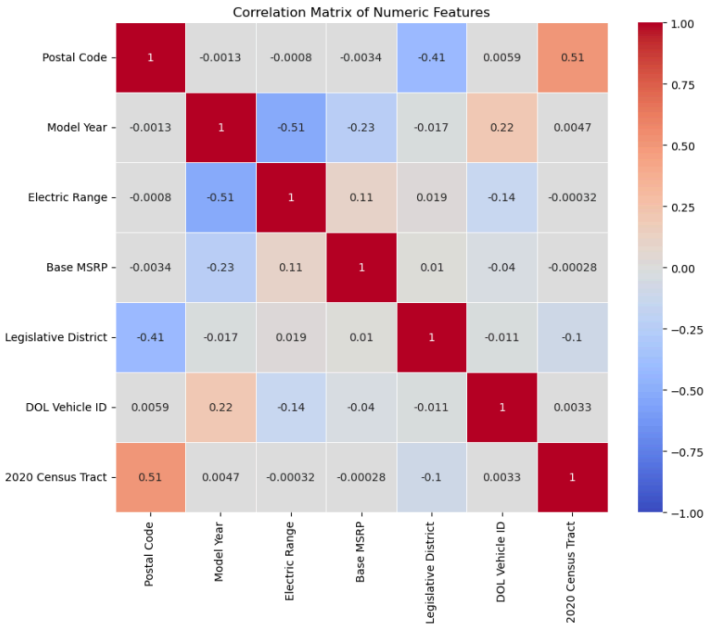




Another form of finding trends in the data is by visualizing the change in the number of vehicles for each model over years in a line plot.

Correlation:

Correlation is a statistical analysis that is used to measure and describe the strength and direction of the relationship between two numerical features.Strength indicates how closely two variables are related to each other, and direction indicates how one variable would change its value as the value of the other variable changes. If the correlation is positive, it means that both features are changing in the same direction (both increasing or both decreasing), whereas if it is negative, it means that features are changing in different directions (if one is increasing the other is decreasing).

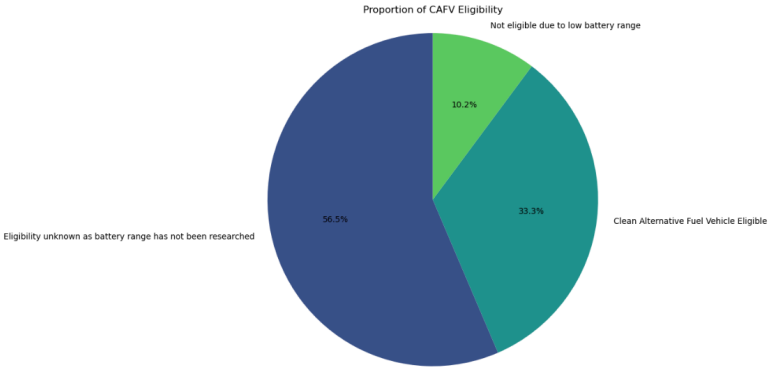


As shown in the correlation matrix figure, there is neither very strong nor strong correlation between any two numerical features, since there are no values in the ranges (0.6, 1) and (-0.6 , -1). There is some correlation between features such as Model Year and Electric Range features, Postal Code and 2020 Census Tract, and between Postal Code and Legistivate District.

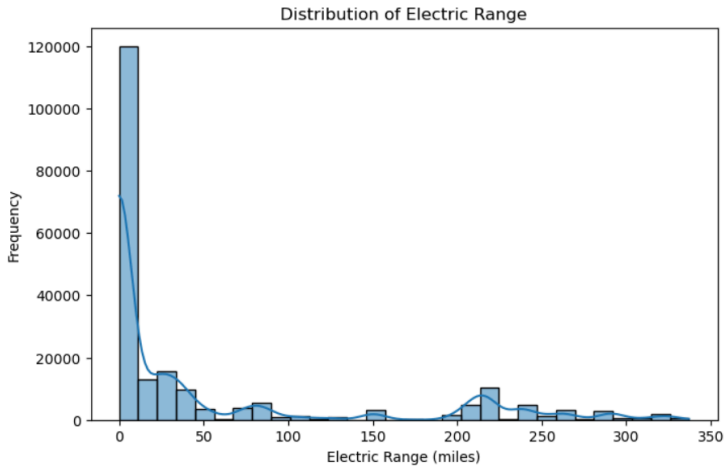
Visualization:

As mentioned before, visualization is done in order to find patterns in the data, and highlight relationships and correlation between different features. Since the dataset contains both categorical and numerical features, different types of plots are needed to find all possible relations.

A pie chart was used to visualize the Clean Alternative Fuel Vehicle (CAFV) Eligibility feature, since it is a categorical feature, and we only visualize this one feature to find how it is distributed and not trying to find the relation between it and another feature.

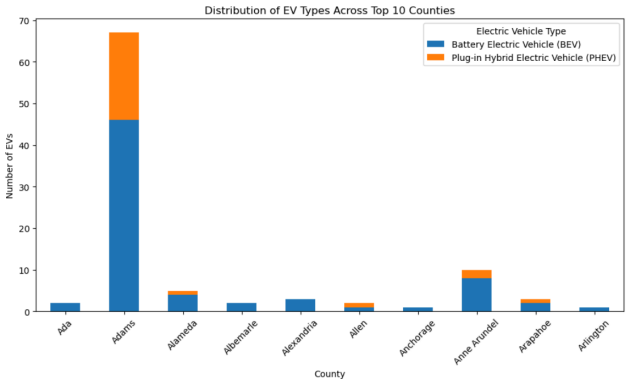


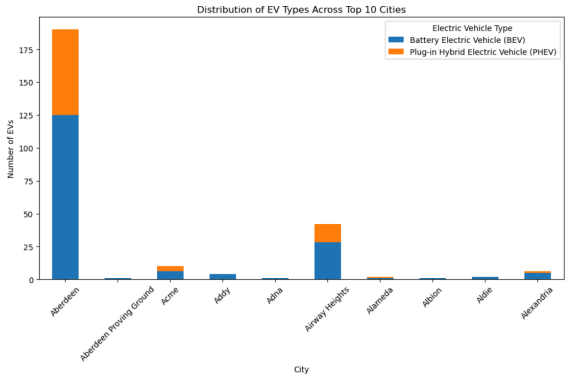
A pie chart was used to visualize the Clean Alternative Fuel Vehicle (CAFV) Eligibility feature, since it is a categorical feature, and we only visualize this one feature to find how it is distributed and not trying to find the relation between it and another feature.

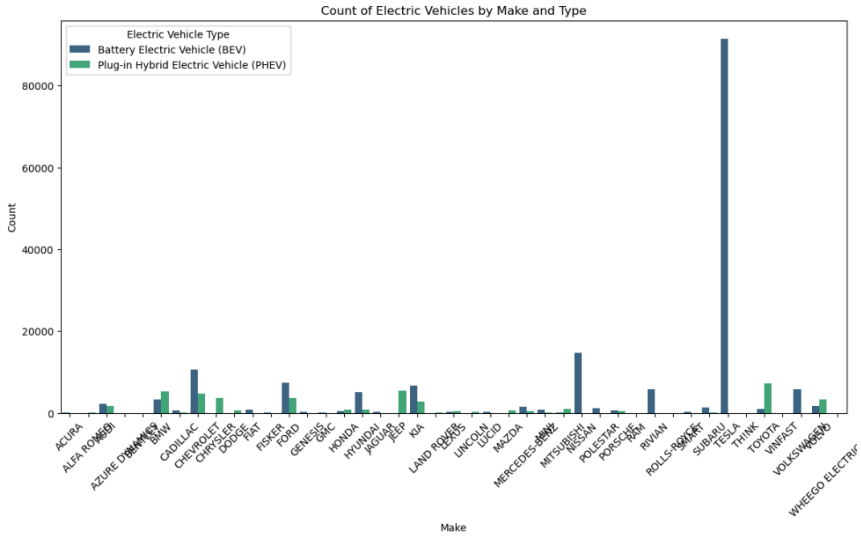


Histograms can be used to find relations between numerical features. But since it does not make sense to find the corresponding value on the y-axis for each value/bin on the x-axis, the x-axis is cut into ranges of values for more readability, simplicity and meaning. An example is a histogram that shows the frequency of the electric range feature.

Another commonly used way of visualizing data is by using a bar chart. It can be used when we want to visualize a categorical feature against a numerical one. A stacked bar chart is the same as the normal bar chart, but is used when the data we visualize has more than 1 type. An example is when we want to visualize the number of electrical vehicles (numerical) in different countries and cities (categorical), but the electrical vehicles can either be battery or plug-in hybrid vehicles.







Moreover, instead of using a stacked bar, we can have two bars for each element of the categorical feature, and each bar represents a different type of the electric vehicles. An example of this is visualizing the number of electric vehicles based on the make feature, and each type has a different bar with a different color.

More visualization examples are in the submitted notebook.