Faculty of Engineering & Technology
Electrical & Computer Engineering Department

ENCS5341 Machine Learning and Data Science

Assignment #1

Prepared by:

Amany Hmidan	1200255
Hiba Jaouni	1201154

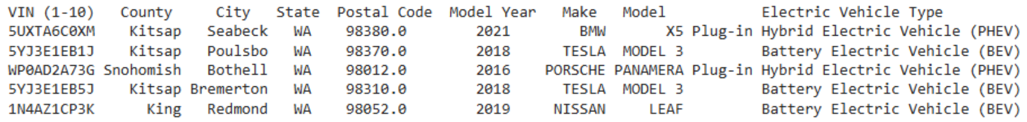
Instructor: Dr. Ismail Khater

Section: 3

Date: Oct 29th, 2024

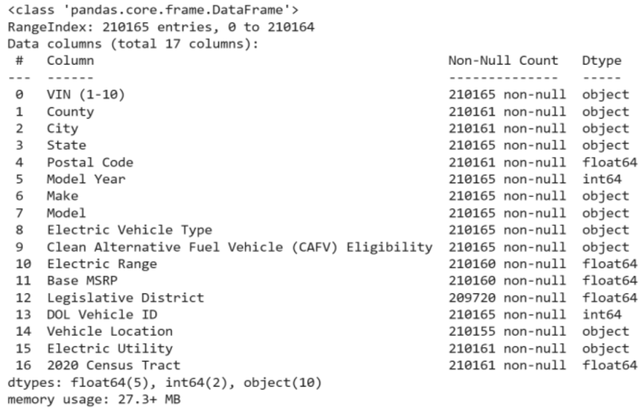
Part 1: Data Cleaning and Feature Engineering

First, the dataset was loaded into a data frame, and here is a sample of the data to ensure that it was loaded correctly.

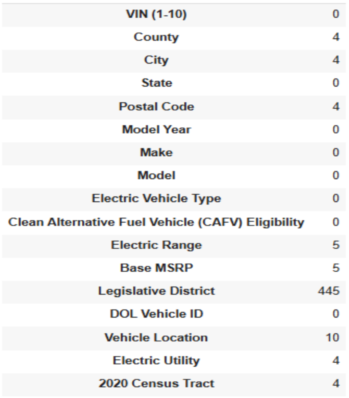


```
VIN (1-10)  County     City  State  Postal Code  Model Year    Make   Model           Electric Vehicle Type
5UXTA6C0XM   Kitsap  Seabeck    WA     98380.0       2021       BMW        X5 Plug-in Hybrid Electric Vehicle (PHEV)
5YJ3E1EB1J   Kitsap  Poulsbo    WA     98370.0       2018     TESLA  MODEL 3          Battery Electric Vehicle (BEV)
WP0AD2A73G Snohomish  Bothell   WA     98012.0       2016   PORSCHE PANAMERA Plug-in Hybrid Electric Vehicle (PHEV)
5YJ3E1EB5J   Kitsap Bremerton   WA     98310.0       2018     TESLA  MODEL 3          Battery Electric Vehicle (BEV)
1N4AZ1CP3K     King  Redmond    WA     98052.0       2019    NISSAN     LEAF          Battery Electric Vehicle (BEV)
```

Next, we have extracted the dataset's number of features and examples.



```
<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
```



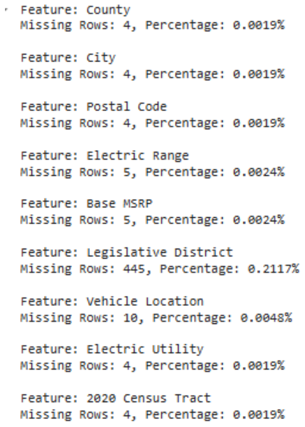
VIN (1-10)	0
County	4
City	4
State	0
Postal Code	4
Model Year	0
Make	0
Model	0
Electric Vehicle Type	0
Clean Alternative Fuel Vehicle (CAFV) Eligibility	0
Electric Range	5
Base MSRP	5
Legislative District	445
DOL Vehicle ID	0
Vehicle Location	10
Electric Utility	4
2020 Census Tract	4

As shown in the figure, there are 17 features and 210165 samples. The feature types include object, float64, and int64. Based on these features, we can conclude what is the dataset for, which shows that it contains information about electric vehicles.

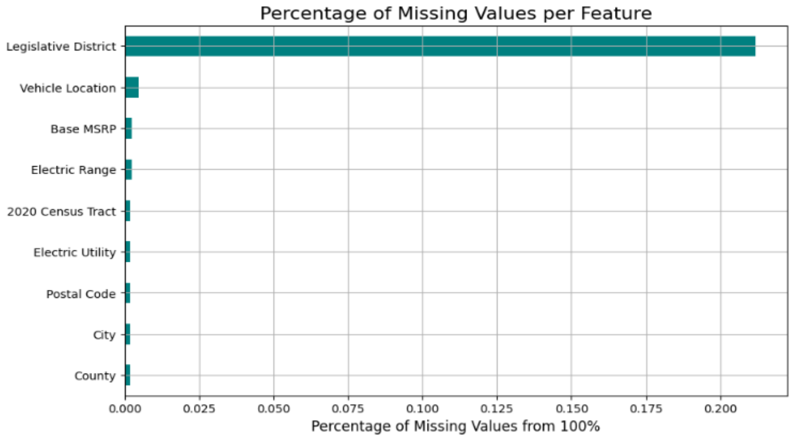
Document Missing Values:

We can observe missing values (NAN) in several features, including County, City, Postal Code, Electric Range, Base MSRP, Legislative District, Vehicle Location, Electric Utility, and 2020 Census Tract. See figure 2.

Then we calculated and displayed the percentage of missing values in each feature relative to the total number of rows in the dataset.



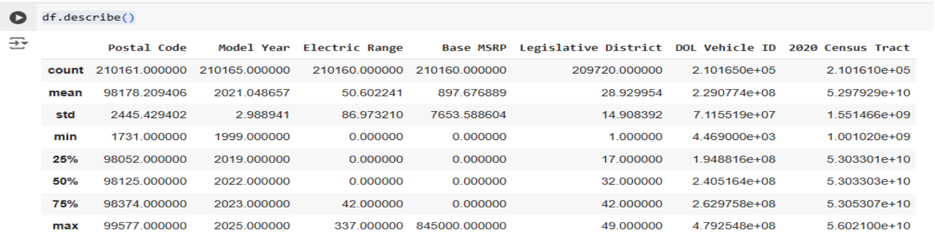
```
Feature: County
Missing Rows: 4, Percentage: 0.0019%

Feature: City
Missing Rows: 4, Percentage: 0.0019%

Feature: Postal Code
Missing Rows: 4, Percentage: 0.0019%

Feature: Electric Range
Missing Rows: 5, Percentage: 0.0024%

Feature: Base MSRP
Missing Rows: 5, Percentage: 0.0024%

Feature: Legislative District
Missing Rows: 445, Percentage: 0.2117%

Feature: Vehicle Location
Missing Rows: 10, Percentage: 0.0048%

Feature: Electric Utility
Missing Rows: 4, Percentage: 0.0019%

Feature: 2020 Census Tract
Missing Rows: 4, Percentage: 0.0019%
```



Based on the printed percentages and the graph, we can observe that the percentage of missing values for most features is almost zero. Therefore, we can safely drop these features without impacting the analysis. However, the feature *Legislative District* stands out with 445 missing rows, accounting for 0.21% of the dataset, which means that 0.21% of the rows in that feature have missing values. Although this percentage is still small, we will attempt to fill these missing values.

Missing Value Strategies:

In this part, the missing values will be handled. But before that, here is a summary of the key statistical measures for each numerical column in the data frame.



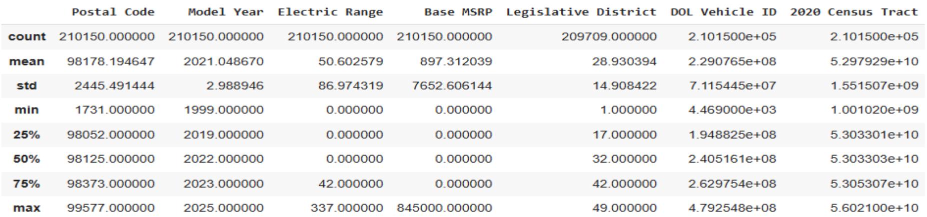
```
df.describe()
```

	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

The count shows that there are missing values in features with a count less than 210165.

As mentioned in the previous part, we will drop the rows containing missing values in all features except for the *Legislative District*. Their missing values are negligible, so filling them would have minimal effect.

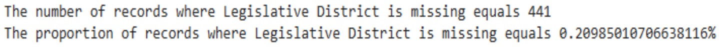
Statistical measures after dropping the rows:



	Postal Code	Model Year	Electric Range	Base MSRP	Legislative District	DOL Vehicle ID	2020 Census Tract
count	210150.000000	210150.000000	210150.000000	210150.000000	209709.000000	2.101500e+05	2.101500e+05
mean	98178.194647	2021.048670	50.602579	897.312039	28.930394	2.290765e+08	5.297929e+10
std	2445.491444	2.988946	86.974319	7652.606144	14.908422	7.115445e+07	1.551507e+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	1.948825e+08	5.303301e+10
50%	98125.000000	2022.000000	0.000000	0.000000	32.000000	2.405161e+08	5.303303e+10
75%	98373.000000	2023.000000	42.000000	0.000000	42.000000	2.629754e+08	5.305307e+10
max	99577.000000	2025.000000	337.000000	845000.000000	49.000000	4.792548e+08	5.602100e+10

Notice that the descriptive statistics for the updated features stayed almost the same because the number of dropped rows is negligible compared to the total number of rows.

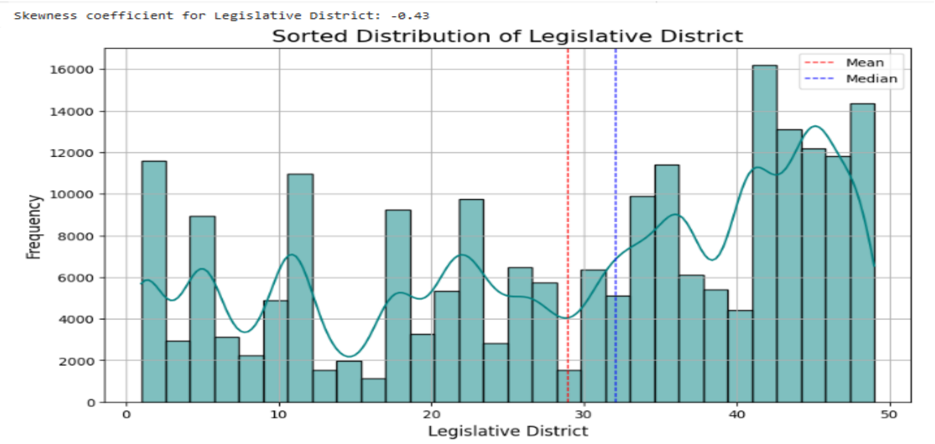
Sometimes, a single record may have multiple missing features, which means that the number of empty records for the *Legislative District* feature might remain unchanged or only be slightly reduced after the previous cleaning step.



```
The number of records where Legislative District is missing equals 441
The proportion of records where Legislative District is missing equals 0.20985010706638116%
```

The number of records containing missing values for the *Legislative District* feature has decreased from 445 to 441.

To address missing values in the Legislative District feature, we can use central tendency measures like the mean, median, or mode. To decide between mean and median, we'll examine the feature's skewness coefficient to select the most representative measure.



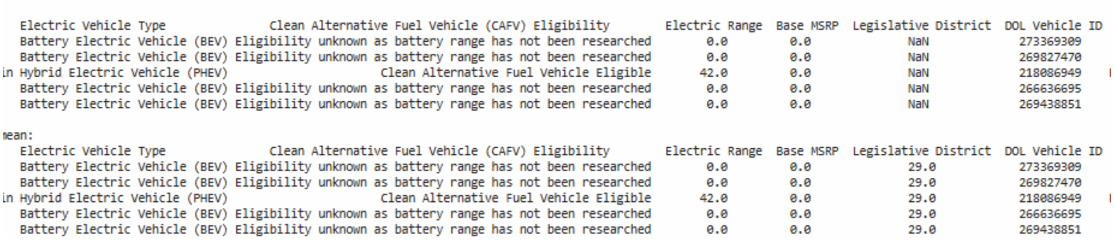
```
Skewness coefficient for Legislative District: -0.43
```

We can observe that the *Legislative District* is negatively skewed since the skewness coefficient is negative and equal to -0.43.

We computed the mean value which equals 28.9 and the median equals 32.

When missing values were filled using the median, the absolute value of the skewness coefficient increased by 0.01. However, when the mean was used for imputation, the skewness coefficient remained unchanged at -0.43. So, we chose to fill them with mean value.

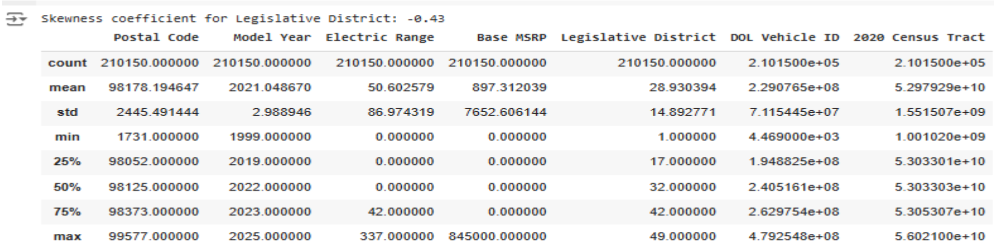
We will round the mean value of 28.9 to 29, and that's because *Legislative District* is often represented as an integer value, as it corresponds to numbered districts indicating the specific political or geographic area where the vehicle is.



```
   Electric Vehicle Type             Clean Alternative Fuel Vehicle (CAFV) Eligibility      Electric Range  Base MSRP  Legislative District  DOL Vehicle ID
   Battery Electric Vehicle (BEV) Eligibility unknown as battery range has not been researched      0.0        0.0                NaN         273369309
   Battery Electric Vehicle (BEV) Eligibility unknown as battery range has not been researched      0.0        0.0                NaN         269827470
in Hybrid Electric Vehicle (PHEV)                         Clean Alternative Fuel Vehicle Eligible     42.0        0.0                NaN         218086949
   Battery Electric Vehicle (BEV) Eligibility unknown as battery range has not been researched      0.0        0.0                NaN         266636695
   Battery Electric Vehicle (BEV) Eligibility unknown as battery range has not been researched      0.0        0.0                NaN         269438851

nean:
   Electric Vehicle Type             Clean Alternative Fuel Vehicle (CAFV) Eligibility      Electric Range  Base MSRP  Legislative District  DOL Vehicle ID
   Battery Electric Vehicle (BEV) Eligibility unknown as battery range has not been researched      0.0        0.0               29.0         273369309
   Battery Electric Vehicle (BEV) Eligibility unknown as battery range has not been researched      0.0        0.0               29.0         269827470
in Hybrid Electric Vehicle (PHEV)                         Clean Alternative Fuel Vehicle Eligible     42.0        0.0               29.0         218086949
   Battery Electric Vehicle (BEV) Eligibility unknown as battery range has not been researched      0.0        0.0               29.0         266636695
   Battery Electric Vehicle (BEV) Eligibility unknown as battery range has not been researched      0.0        0.0               29.0         269438851
```

The figure shows some of the rows that had a missing value and how they were replaced by the value 29.

Statistical measures after filling missing spaces with mean:



Skewness coefficient for Legislative District: -0.43

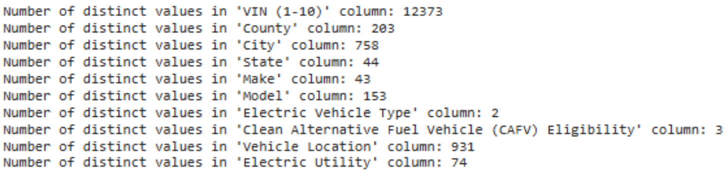
	Postal Code	Model Year	Electric Range	Base MSRP	Legislative District	DOL Vehicle ID	2020 Census Tract
count	210150.000000	210150.000000	210150.000000	210150.000000	210150.000000	2.101500e+05	2.101500e+05
mean	98178.194647	2021.048670	50.602579	897.312039	28.930394	2.290765e+08	5.297929e+10
std	2445.491444	2.988946	86.974319	7652.606144	14.892771	7.115445e+07	1.551507e+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	1.948825e+08	5.303301e+10
50%	98125.000000	2022.000000	0.000000	0.000000	32.000000	2.405161e+08	5.303303e+10
75%	98373.000000	2023.000000	42.000000	0.000000	42.000000	2.629754e+08	5.305307e+10
max	99577.000000	2025.000000	337.000000	845000.000000	49.000000	4.792548e+08	5.602100e+10

However, Statistical measures when filled with median will be almost the same as mean because the number of rows with missing values are small compared to total number of rows.

Feature Encoding:

In this part, we will encode some categorical features into numerical data, because it is better for fitting the data in machine learning models.

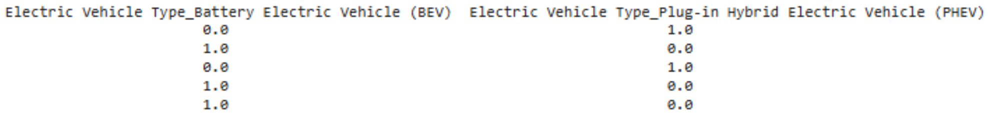
To decide which encoding strategy is the best, we need to observe how many distinct values are presented in each feature.



```
Number of distinct values in 'VIN (1-10)' column: 12373
Number of distinct values in 'County' column: 203
Number of distinct values in 'City' column: 758
Number of distinct values in 'State' column: 44
Number of distinct values in 'Make' column: 43
Number of distinct values in 'Model' column: 153
Number of distinct values in 'Electric Vehicle Type' column: 2
Number of distinct values in 'Clean Alternative Fuel Vehicle (CAFV) Eligibility' column: 3
Number of distinct values in 'Vehicle Location' column: 931
Number of distinct values in 'Electric Utility' column: 74
```

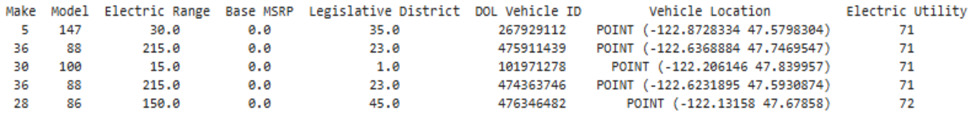
One-hot encoding adds a separate column for each category in the feature. For *Electrical Vehicle Type*, and *Clean Alternative Fuel Vehicle Eligibility*, One-Hot encoding will be used. This is because the number of distinct categories in them is small, so adding that number of columns won't cause a significant overhead.

In One-hot encoding, a value of 1 is placed in the column corresponding to the category, while all other columns are set to 0.



```
Electric Vehicle Type_Battery Electric Vehicle (BEV)  Electric Vehicle Type_Plug-in Hybrid Electric Vehicle (PHEV)
                         0.0                                                  1.0
                         1.0                                                  0.0
                         0.0                                                  1.0
                         1.0                                                  0.0
                         1.0                                                  0.0
```

However, the number of distinct categories in other features is big, so One-hot encoding won't be the most appropriate method. We will use label encoding for *Make*, *Model* and *Electric Utility*.



```
Make  Model  Electric Range  Base MSRP  Legislative District  DOL Vehicle ID         Vehicle Location          Electric Utility
  5    147         30.0         0.0             35.0             267929112    POINT (-122.8728334 47.5798304)         71
 36     88        215.0         0.0             23.0             475911439    POINT (-122.6368884 47.7469547)         71
 30    100         15.0         0.0              1.0             101971278      POINT (-122.206146 47.839957)         71
 36     88        215.0         0.0             23.0             474363746    POINT (-122.6231895 47.5930874)         71
 28     86        150.0         0.0             45.0             476346482        POINT (-122.13158 47.67858)         72
```

Label Encoding keeps the feature as a single column, and assigns a unique numerical value to each category within a feature.

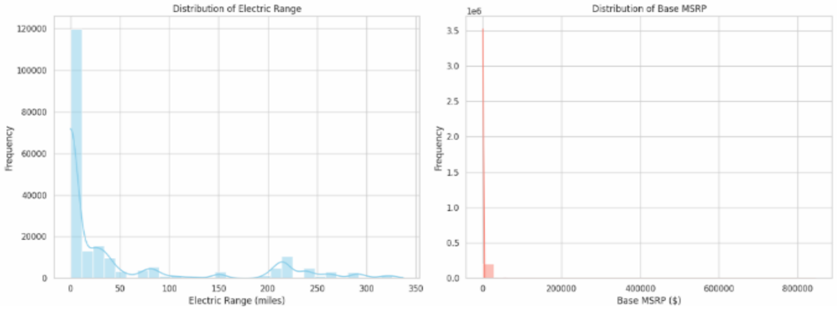
Normalization:

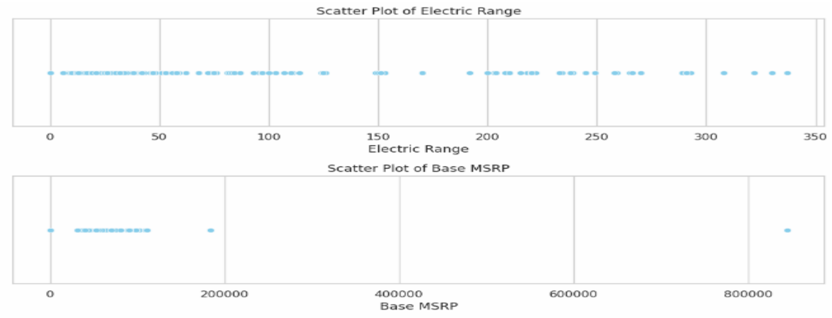
Normalization is generally recommended for features that have a wide range of values, especially if they are going to be used in machine learning models that rely on distance metrics like KNN.

We studied each feature to decide whether it needs normalization or not, *Base MSRP* represents the recommended selling price of a vehicle, which will include a wide range of prices because vehicles models vary from budget priced to luxury ones that are very expensive, so normalizing it will be a good option.

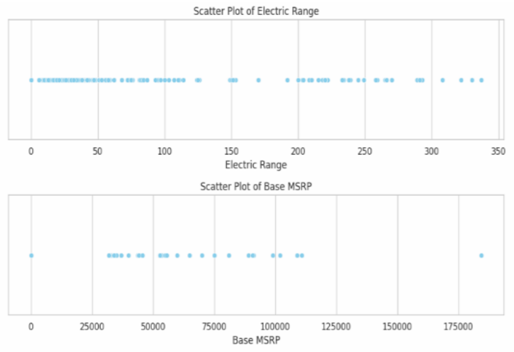
Moreover, *Electric Range* is directly related to the vehicle's performance, so normalization would be beneficial.

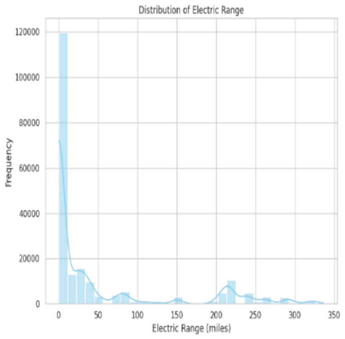
However, we didn't see that other features require normalization, because they aren't directly related to performance or numeric calculations. For example, *Make, Model, Electric Vehicle Type, CAFV Eligibility, Electric Utility* features were encoded to either label encoding or One-Hot encoding so no need to normalize them. Additionally, *VIN (1-10) and DOL Vehicle ID* are identifiers which are unique and do not contain ordinal or interval information, so they do not need normalization. Moreover, Vehicle Location was split into *Longitude* and *Latitude* and used in spatial distribution.

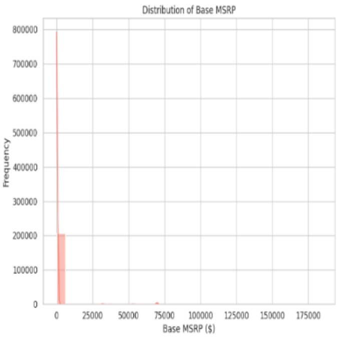




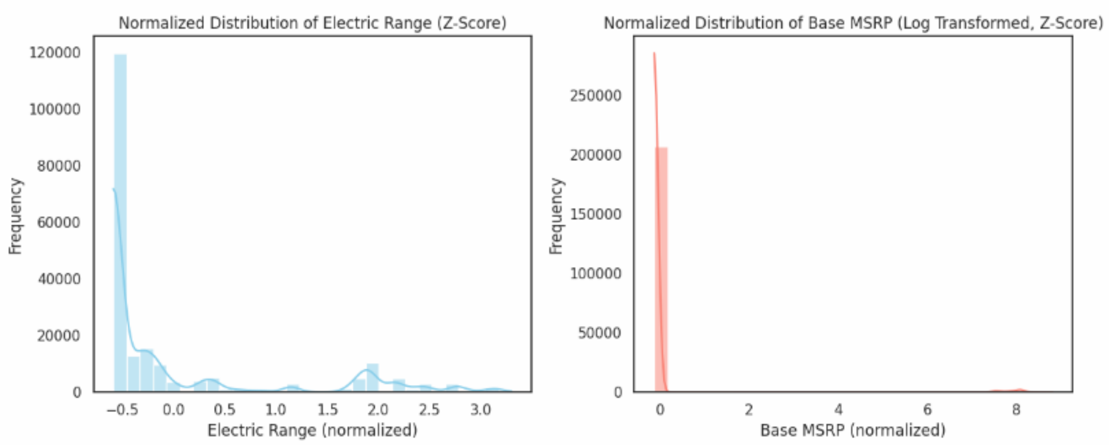
While plotting the distributions of *Electric Range* and *Base MSRP*, we noticed an outlier around the value of 800000, and it was only one row with this value so we dropped it.







After removing the outlier, we applied Z-Score normalization.

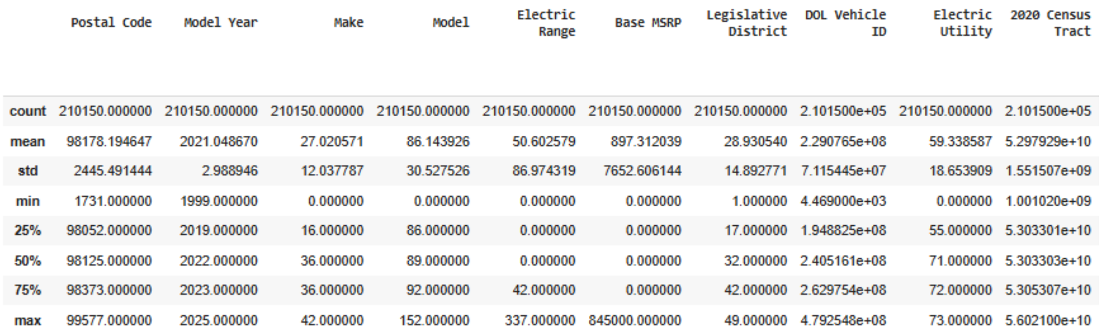


Notice that the range of values became smaller and easier to deal with, but there were still outliers, even though we handled the most extreme value above.

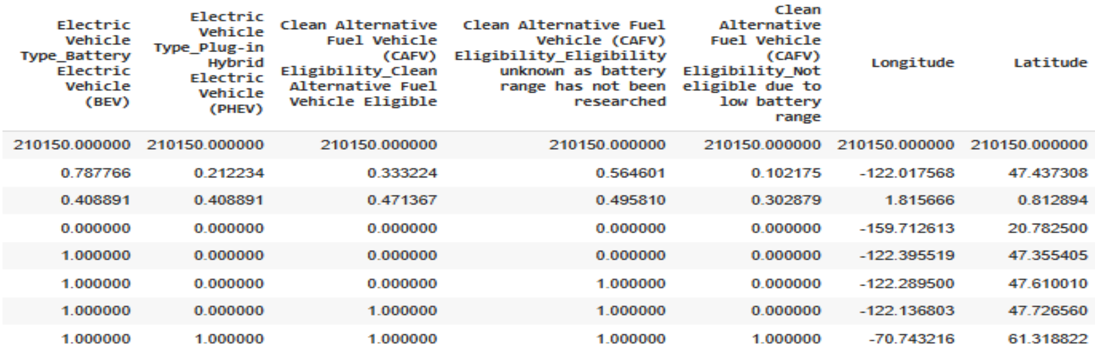
Part 2: Exploratory Data Analysis & Visualizations

Descriptive Statistics:

Here is summary statistics (mean, median, standard deviation) for numerical features:



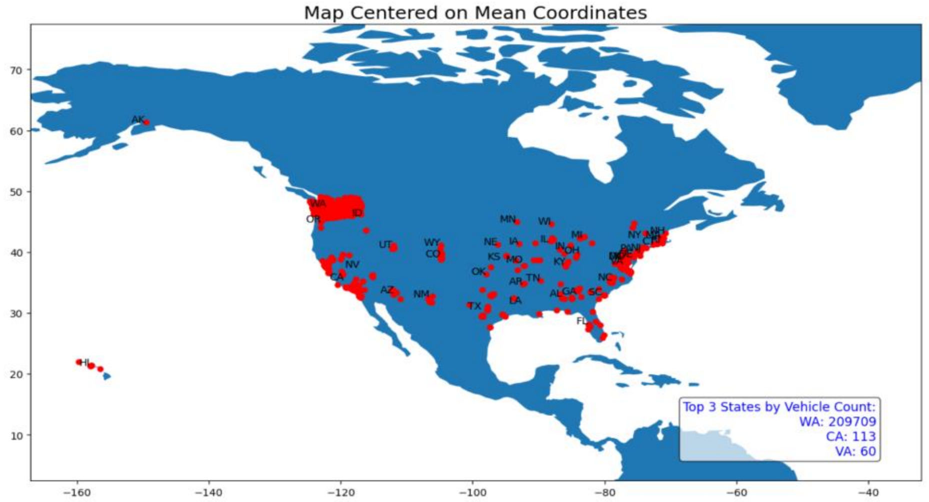
	Postal Code	Model Year	Make	Model	Electric Range	Base MSRP	Legislative District	DOL Vehicle ID	Electric Utility	2020 Census Tract
count	210150.000000	210150.000000	210150.000000	210150.000000	210150.000000	210150.000000	210150.000000	2.101500e+05	210150.000000	2.101500e+05
mean	98178.194647	2021.048670	27.020571	86.143926	50.602579	897.312039	28.930540	2.290765e+08	59.338587	5.297929e+10
std	2445.491444	2.988946	12.037787	30.527526	86.974319	7652.606144	14.892771	7.115445e+07	18.653909	1.551507e+09
min	1731.000000	1999.000000	0.000000	0.000000	0.000000	0.000000	1.000000	4.469000e+03	0.000000	1.001020e+09
25%	98052.000000	2019.000000	16.000000	86.000000	0.000000	0.000000	17.000000	1.948825e+08	55.000000	5.303301e+10
50%	98125.000000	2022.000000	36.000000	89.000000	0.000000	0.000000	32.000000	2.405161e+08	71.000000	5.303303e+10
75%	98373.000000	2023.000000	36.000000	92.000000	42.000000	0.000000	42.000000	2.629754e+08	72.000000	5.305307e+10
max	99577.000000	2025.000000	42.000000	152.000000	337.000000	845000.000000	49.000000	4.792548e+08	73.000000	5.602100e+10



Electric Vehicle Type_Battery Electric Vehicle (BEV)	Electric Vehicle Type_Plug-in Hybrid Electric Vehicle (PHEV)	Clean Alternative Fuel Vehicle (CAFV) Eligibility_Clean Alternative Fuel Vehicle Eligible	Clean Alternative Fuel Vehicle (CAFV) Eligibility_Eligibility unknown as battery range has not been researched	Clean Alternative Fuel Vehicle (CAFV) Eligibility_Not eligible due to low battery range	Longitude	Latitude
210150.000000	210150.000000	210150.000000	210150.000000	210150.000000	210150.000000	210150.000000
0.787766	0.212234	0.333224	0.564601	0.102175	-122.017568	47.437308
0.408891	0.408891	0.471367	0.495810	0.302879	1.815666	0.812894
0.000000	0.000000	0.000000	0.000000	0.000000	-159.712613	20.782500
1.000000	0.000000	0.000000	0.000000	0.000000	-122.395519	47.355405
1.000000	0.000000	0.000000	1.000000	0.000000	-122.289500	47.610010
1.000000	0.000000	1.000000	1.000000	0.000000	-122.136803	47.726560
1.000000	1.000000	1.000000	1.000000	1.000000	-70.743216	61.318822

Spatial Distribution:

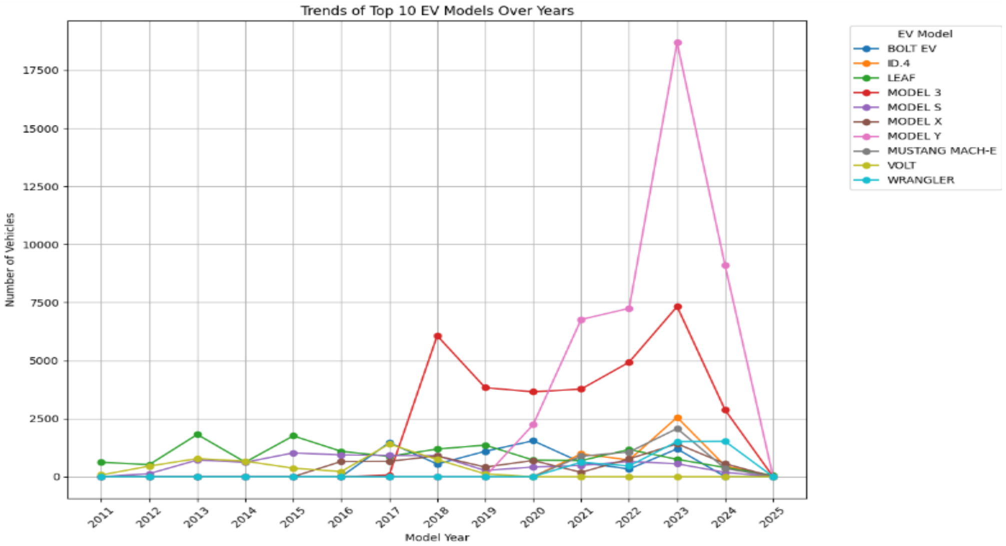
In this part, we will visualize the spatial distribution of EVs across locations, but we first split the *Vehicle Location* feature into two separate columns: *Longitude* and *Latitude*, to enhance its usability.



The map shows the states in which EVs are located. Note that the majority of EVs are located in Washington state.

Model Popularity & Temporal Analysis:

In this part, we will analyze the popularity of different EV models and identify any trends.

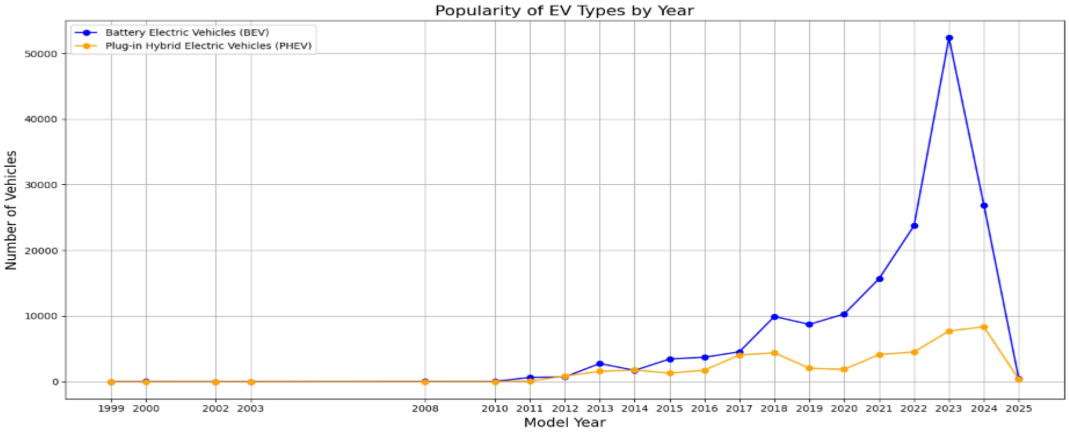


We observe that there is a sharp increase in the popularity of Model Y around the year of 2022, where its count reached the highest peak of all models at year 2023. However, there is a drop-off in the following year.

In 2018, Model 3 became more popular than others, and its popularity was relatively stable, though with smaller peaks than Model Y.

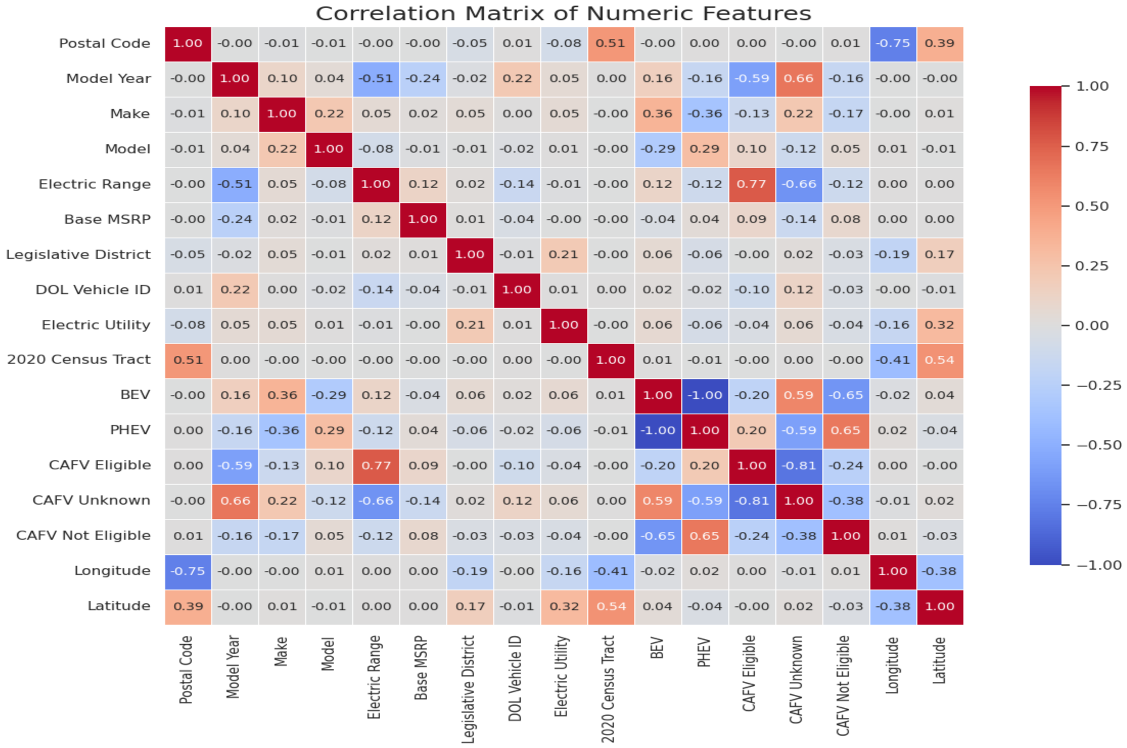
Some models like VOLT had less presence over the years, it almost disappeared from year 2019. Others show a gradual increase over year but without significant spikes.

Models Like LEAF, BOLT EV, Model S and ID.4 show a steady presence over time, though with smaller counts compared to Model Y and Model 3. These models appear to have consistent market presence but without significant spikes.



The graph shows a gap between Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) in terms of popularity, especially in recent years. This is normal as consumer awareness about environmental impact grows, there has been a noticeable shift toward fully electric vehicles and many countries are encouraging them.

Correlations between features:



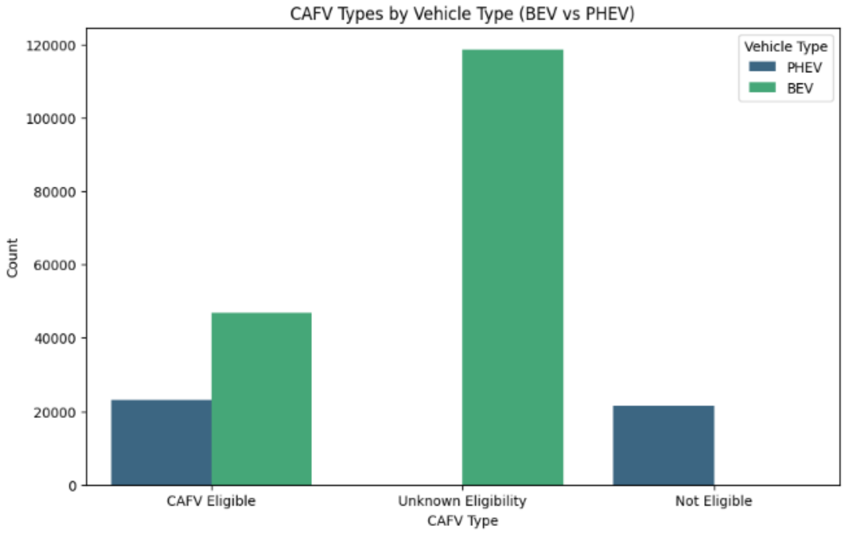
We observe that there is a positive correlation between Electric Range and Base MSRP, indicating that vehicles with a higher electric range tend to have a higher MSRP. Moreover, notice one of the features that was encoded by One-Hot encoding, which is *Clean Alternative Fuel Vehicle (CAFV) Eligibility*, is related to BEV such that there is a negative correlation between BEV and CAFV Eligible.

Additionally, Code and Longitude exhibit a strong negative correlation, implying that areas with lower postal codes tend to be located further west. 2020 Census Tract also has a strong positive correlation with Postal Code and Longitude, indicating geographic clustering in the data.

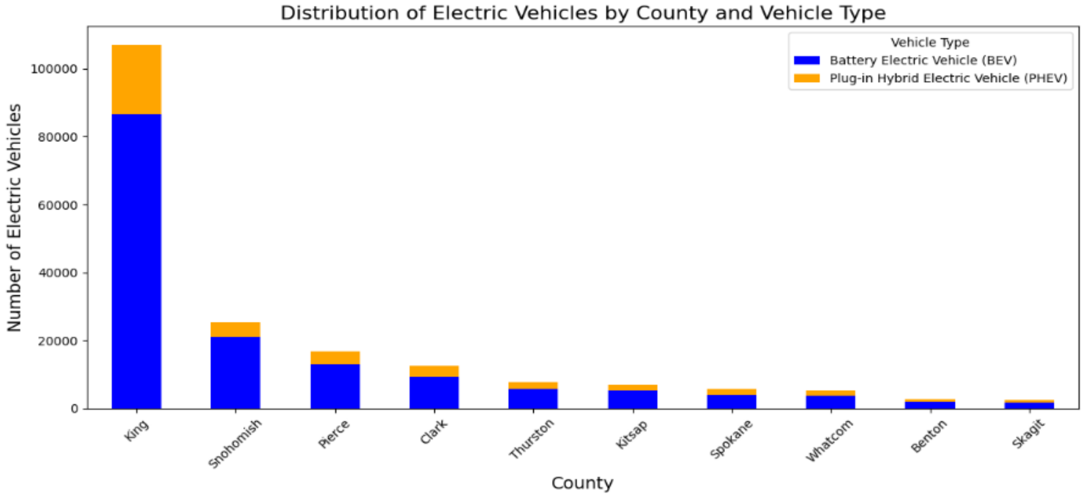
Vehicle Characteristics: Make and Model Year show a positive correlation with each other, which could mean that newer models are associated with certain manufacturers more than others. Electric Range and BEV have a positive correlation, as expected, given that BEVs generally offer a significant electric range compared to plug-in hybrid electric vehicles (PHEVs).

Another important observation is that the correlation between the vehicle types BEV and PHEV is -1, which means they are highly correlated but negatively, which is normal because the characteristics of the two types are different than each other.

Visualization:



This bar chart ensures what was showed in the correlation part, which shows CAFV Unknown has a high positive correlation with BEV. On the other hand, PHEV is more correlated with CAFV Eligible.



This bar chart shows the top 10 counties with highest EVs distribution, and the biggest number of EVS is in King. It also shows that battery EVs are a lot more popular than Hybrid, this is because they have less negative impact on the environment so people in these counties are choosing them.