Faculty of Engineering and Technology

Department of Electrical and Computer Engineering

Machine Learning and Data Science - ENCS5341

Assignment #1

Data Science

Prepared by:

- Ali Shaikh Qasem ID: 1212171
- Abdalrahman Juber ID: 1211769

Instructors: Dr. Ismail Khater, Dr. Yazan Abu Farha.

Sections: 1,3.

Date: 30/10/2024

Birzeit University

2024-2025

Table of Contents

List of Figures

List of Tables

Data set description

We brought "Electric Vehicle Population Data" dataset from:
https://catalog.data.gov/dataset/electric-vehicle-population-data

Provided by the State of Washington, this dataset displays information about battery electric vehicles (BEVs) and plug-in hybrid electric vehicles (PHEVs) currently registered through the Washington State Department of Licensing.

Data is separated into 17 different columns, showing each vehicle's VIN, county and city of registration, make and model, electric type and electric range. Vehicle model years range from 2013 to the current year, with metadata being routinely updated by the Washington government.

By using **info ()** method form Panda's library, the number of examples(rows) and features(column) is (210165, 17).

Type of features:

Table 1: Types of data

Object Features	Numerical Features
VIN, County, City, Make, Model, Electric Vehicle Type, Clean Alternative Fuel Vehicle (CAFV) Eligibility, Vehicle Location, and Electric Utility.	Postal Code, Model Year, Electric Range, Base MSRP, Legislative District, DOL Vehicle ID and 2020 Census Tract.

The context specified that the dataset is useful for analyzing the electric vehicle population across Washington state and the cars models.

Data Cleaning and Feature Engineering:

Document Missing Values:

For our dataset we may find missing data. It can occur for a variety of reasons, such as human error, equipment malfunction, or deliberate omission.

And it's important to handle missing data because missing data can have a significant impact on the quality and accuracy of a dataset. If not handled properly, it can lead to biased results and inaccurate conclusions.

We used isNull () and sum () methods from panda's library to find the missing values and sum their frequency for each column.

Table 2: Sum of missing values

Features	VIN	County	City	State	Postal Code	Model Year	Make	Model	Electric Utility
Number of missing values	0	4	4	0	0	0	0	0	4
Features	Electric Vehicle Type	CAFV	Electric Range	Base MSRP	Legislative District	DOL Vehicle ID	Vehicle Location	2020 Census Tract	
Number of missing values	0	0	5	5	445	0	10	4	

Also, we find the number of missing values for each column by using isnull (). any (). sum () which represents: 456

And by dividing the number of rows to this value we got the percentage 0.22 %.

Missing Value Strategies:

Handling Missing Data Through Dropping:

First, we check if we have an empty record. Luckly, we don't have any empty record to drop all the rows that have an empty record.

Then removes all rows that contain at least one missing value (NaN) from the DataFrame.

By using dropna () method. We now have 209709 rows from 210165 rows.

Mean/Median Imputation:

We apply median and mean methods for numerical columns to full out the missing data.

Analysis:

Original Data Summary:

	Postal Code	Model Year	...	DOL Vehicle ID	2020 Census Tract
count	210161.000000	210165.000000	...	2.101650e+05	2.101610e+05
mean	98178.209406	2021.048657	...	2.290774e+08	5.297929e+10
std	2445.429402	2.988941	...	7.115519e+07	1.551466e+09
min	1731.000000	1999.000000	...	4.469000e+03	1.001020e+09
25%	98052.000000	2019.000000	...	1.948816e+08	5.303301e+10
50%	98125.000000	2022.000000	...	2.405164e+08	5.303303e+10
75%	98374.000000	2023.000000	...	2.629758e+08	5.305307e+10
max	99577.000000	2025.000000	...	4.792548e+08	5.602100e+10

[8 rows x 7 columns]

Mean Imputed Summary:

	Postal Code	Model Year	...	DOL Vehicle ID	2020 Census Tract
count	210165.000000	210165.000000	...	2.101650e+05	2.101650e+05
mean	98178.209406	2021.048657	...	2.290774e+08	5.297929e+10
std	2445.406130	2.988941	...	7.115519e+07	1.551452e+09
min	1731.000000	1999.000000	...	4.469000e+03	1.001020e+09
25%	98052.000000	2019.000000	...	1.948816e+08	5.303301e+10
50%	98125.000000	2022.000000	...	2.405164e+08	5.303303e+10
75%	98374.000000	2023.000000	...	2.629758e+08	5.305307e+10
max	99577.000000	2025.000000	...	4.792548e+08	5.602100e+10

[8 rows x 7 columns]

Dropped Rows Summary:

	Postal Code	Model Year	...	DOL Vehicle ID	2020 Census Tract
count	209709.000000	209709.000000	...	2.097090e+05	2.097090e+05
mean	98266.123528	2021.049478	...	2.290932e+08	5.303999e+10
std	307.778373	2.989276	...	7.116107e+07	1.637504e+07
min	98001.000000	1999.000000	...	4.469000e+03	5.300195e+10
25%	98052.000000	2019.000000	...	1.949064e+08	5.303301e+10
50%	98125.000000	2022.000000	...	2.405281e+08	5.303303e+10
75%	98374.000000	2023.000000	...	2.629829e+08	5.305307e+10
max	99403.000000	2025.000000	...	4.792548e+08	5.307794e+10

Median Imputed Summary:

	Postal Code	Model Year	...	DOL Vehicle ID	2020 Census Tract
count	210165.000000	210165.000000	...	2.101650e+05	2.101650e+05
mean	98178.208393	2021.048657	...	2.290774e+08	5.297930e+10
std	2445.406141	2.988941	...	7.115519e+07	1.551452e+09
min	1731.000000	1999.000000	...	4.469000e+03	1.001020e+09
25%	98052.000000	2019.000000	...	1.948816e+08	5.303301e+10
50%	98125.000000	2022.000000	...	2.405164e+08	5.303303e+10
75%	98374.000000	2023.000000	...	2.629758e+08	5.305307e+10
max	99577.000000	2025.000000	...	4.792548e+08	5.602100e+10

Figure 1: Analysis of different missing value strategies

As we can see in the figure above the drop strategy reduced the dataset size.

Maintains dataset size but reduces variability by filling missing values with the mean.

Like mean imputation, but better for skewed data or when outliers are present.

Feature Encoding:

First, we will use one hot encoding for the Make feature, which will divide our feature to 41 columns.

Then, we will do the same for the Model feature.

```
   Make_ACURA  Make_ALFA ROMEO  Make_AUDI  ...  Model_XC90  Model_XM  Model_ZDX
0         0.0              0.0        0.0  ...         0.0       0.0        0.0
1         0.0              0.0        0.0  ...         0.0       0.0        0.0
2         0.0              0.0        0.0  ...         0.0       0.0        0.0
3         0.0              0.0        0.0  ...         0.0       0.0        0.0
4         0.0              0.0        0.0  ...         0.0       0.0        0.0

[5 rows x 196 columns]
```

Figure 2: One hot encoding

Normalization:

We are going to use MinMax scalar for normalization on Electric Range feature. We got.

```
0    0.089021
1    0.637982
2    0.044510
3    0.637982
4    0.445104
Name: Electric Range, dtype: float64
0     30.0
1    215.0
2     15.0
3    215.0
4    150.0
Name: Electric Range, dtype: float64
```

Figure 3: Min Maz normalization

As we can see above our values are now ranged from 0 to 1.

Exploratory Data Analysis

Descriptive Statistics

Descriptive analysis is a method for quantitatively describing the main features of a collection of data. In this part, we calculate summary statistics like (mean, median, and standard deviation) for the numerical features in our dataset.

The results are shown below:

Table 3: descriptive statistics foe numerical features

Numerical Feature	Mean	Median	Standard Deviation
Postal code	98178.2	98125	2445.4
Model year	2021.02	2022	2.98
Electric range	50.6	0	96.9
Base MSRP	897.7	0	7653.5
Legislative district	28.9	32	14.9
DOL vehicle id	2.29e8	2.405e8	7.1e7
2020 census tract	5.29e10	5.3e10	1.5e9

It's noticed from above that the median is very close to the mean except for electric range and base MSRP, where the median gives zero since most of their values are zero, while the mean gives non-zero values since mean is very **sensitive to outliers**.

Spatial Distribution

We can observe from the figure below that we have a high density in certain areas For example, a dense cluster is present in and around major cities, which indicates that urban areas tend to have more EV adoption.

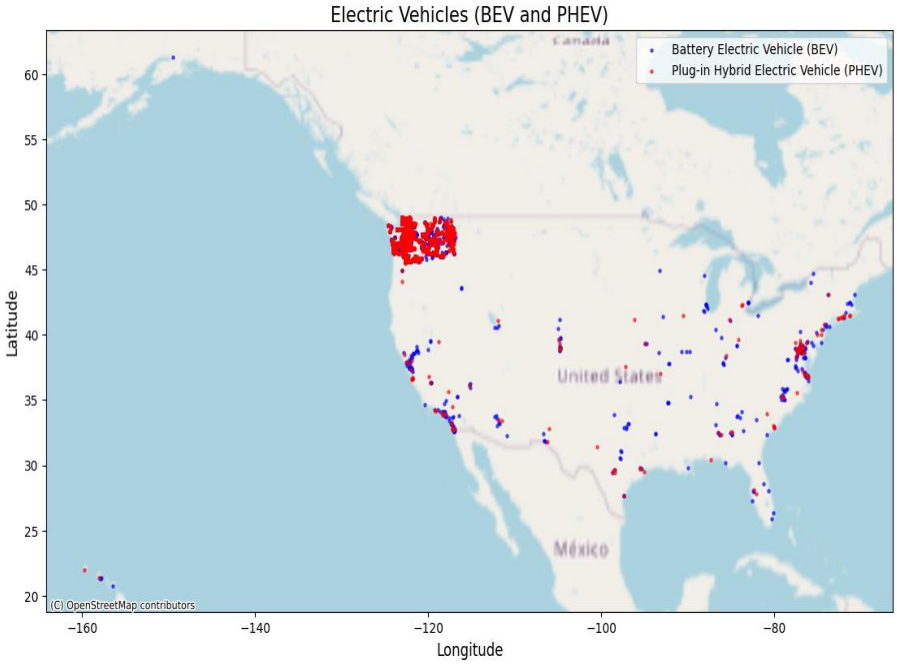

Figure 4: Spatial distribution

Model Popularity

In this part, we visualize the make and the model for each car as a bar chart to analyze the popularity of different EVs and identify the trends.

Make popularity:

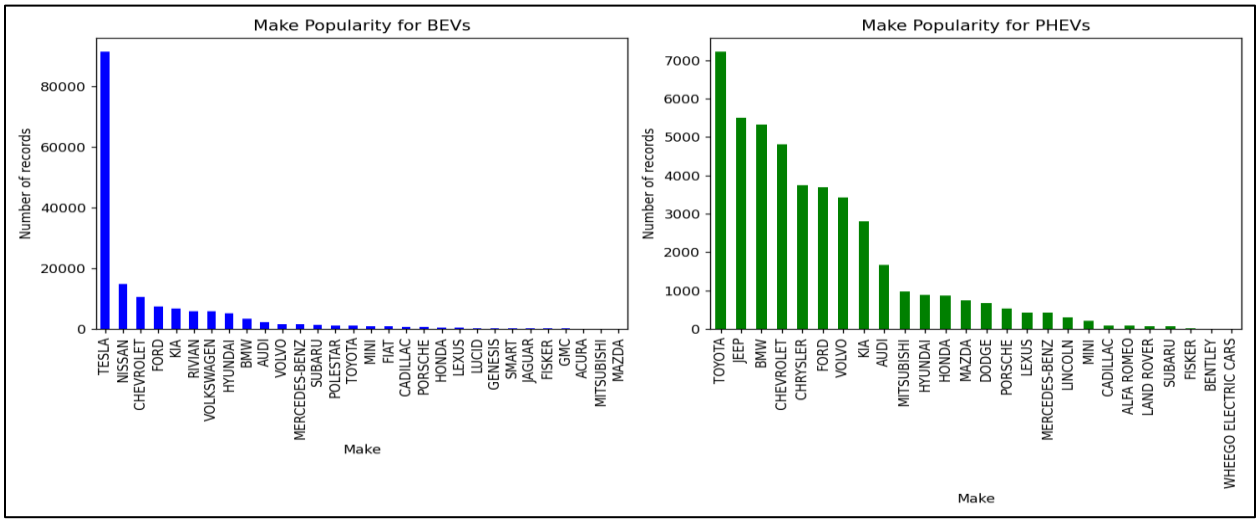

Figure 5: make popularity analysis

From the results above, we notice that the TESLA is the dominant company for the BEVs. While the population of PHEVs is more widely distributed among different companies like TOYOTA, BMW, JEEP, and CHEVEROLET…etc.

Model Popularity:

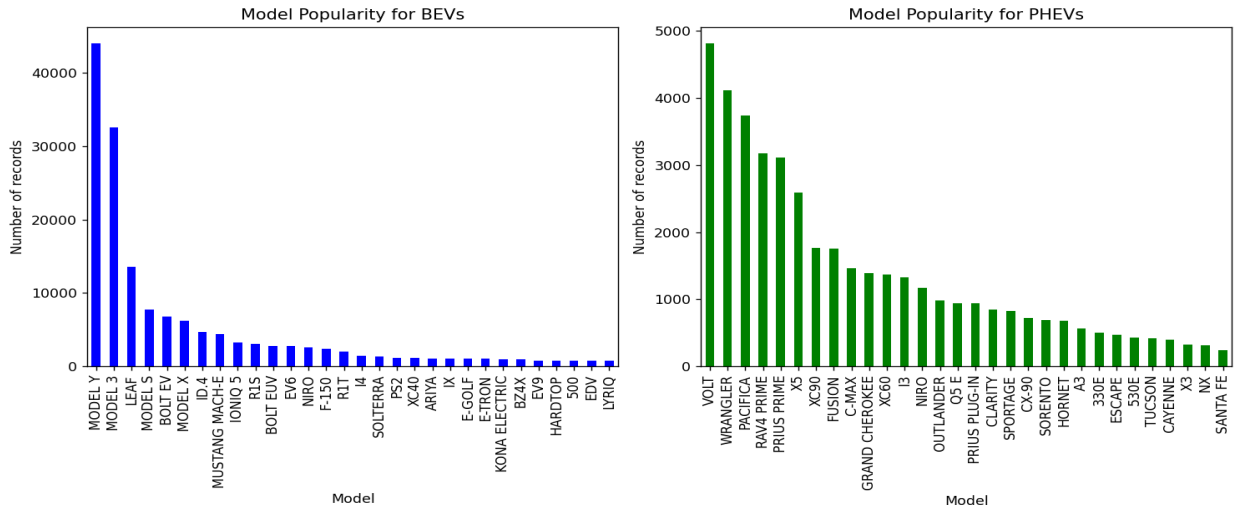

Figure 6: model popularity analysis

The results above show that MODEL Y and MODEL S are the trend models for the BEVs, while there are many popular models for PHEVs like VOLT, WRANGLER, and X5.

Relationships between numeric features

In this part, we investigate the relationships between numeric features. This is achieved by calculating the correlation between each pair of numeric features using correlation matrix as shown:

	Postal Code	Model Year	Electric Range	Base MSRP	Legislative District	DOL Vehicle ID	2020 Census Tract
Postal Code	1.000000	-0.001291	-0.000800	-0.003408	-0.412348	0.005862	0.508744
Model Year	-0.001291	1.000000	-0.513534	-0.230651	-0.016824	0.215703	0.004710
Electric Range	-0.000800	-0.513534	1.000000	0.114155	0.019025	-0.140689	-0.000323
Base MSRP	-0.003408	-0.230651	0.114155	1.000000	0.010477	-0.039501	-0.000283
Legislative District	-0.412348	-0.016824	0.019025	0.010477	1.000000	-0.010728	-0.100714
DOL Vehicle ID	0.005862	0.215703	-0.140689	-0.039501	-0.010728	1.000000	0.003347
2020 Census Tract	0.508744	0.004710	-0.000323	-0.000283	-0.100714	0.003347	1.000000

Figure 7: correlation matrix for numeric features

The correlation matrix above shows that the correlation ranges between -0.5 and 0.5 for most features which indicates a moderate but not very strong relationship between these features.

Visualization

Data Exploration Visualizations

In this part, we create different types of visualizations to explore the features and relationships between features.

- The following histograms describe some features like model year and electric range:

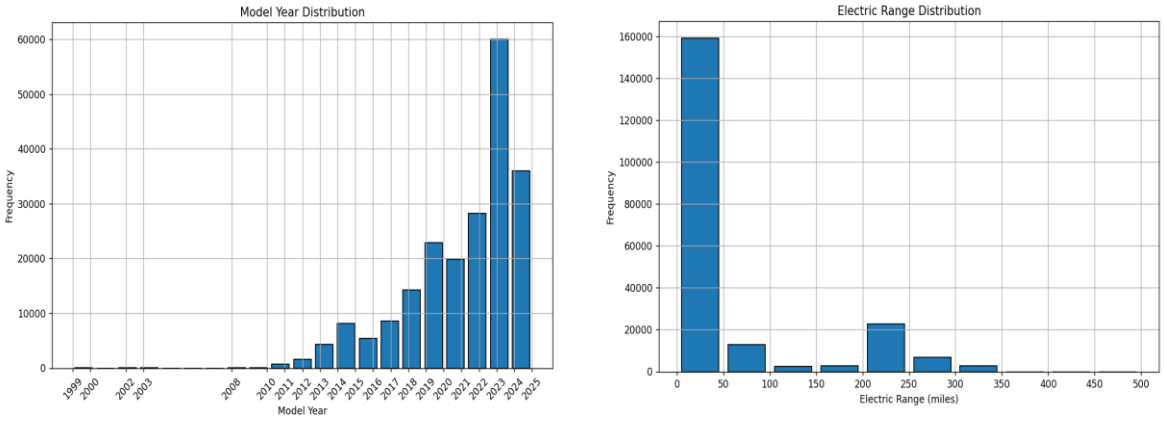

Figure 8: histograms for model year and electric range

From histograms shown above, we notice that the electric range (0-50 km) is the most popular range, while 2023 is the most popular model year.

- The following scatter plot and boxplot describes the relationship between the model year and the electric range:

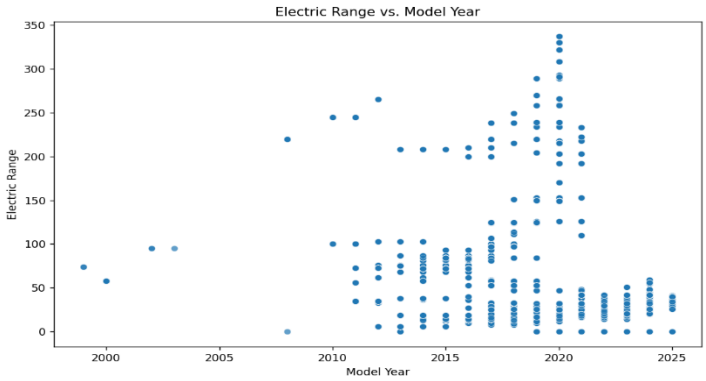

Figure 10: Model year and electric range scatter plot.

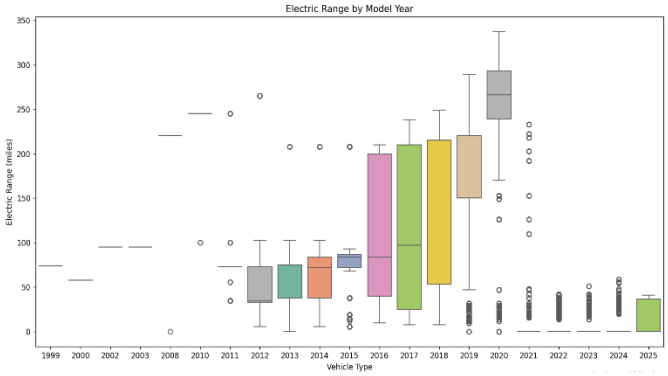

Figure 9: model year and electric range boxplot

The plots above shows that the electric range increases gradually as the model year increase. And the greatest electric range is obtained around the year 2020.

Comparative Visualization

In this part, we compare the distribution of EVs across different cities and counties using bar char visualization. County:

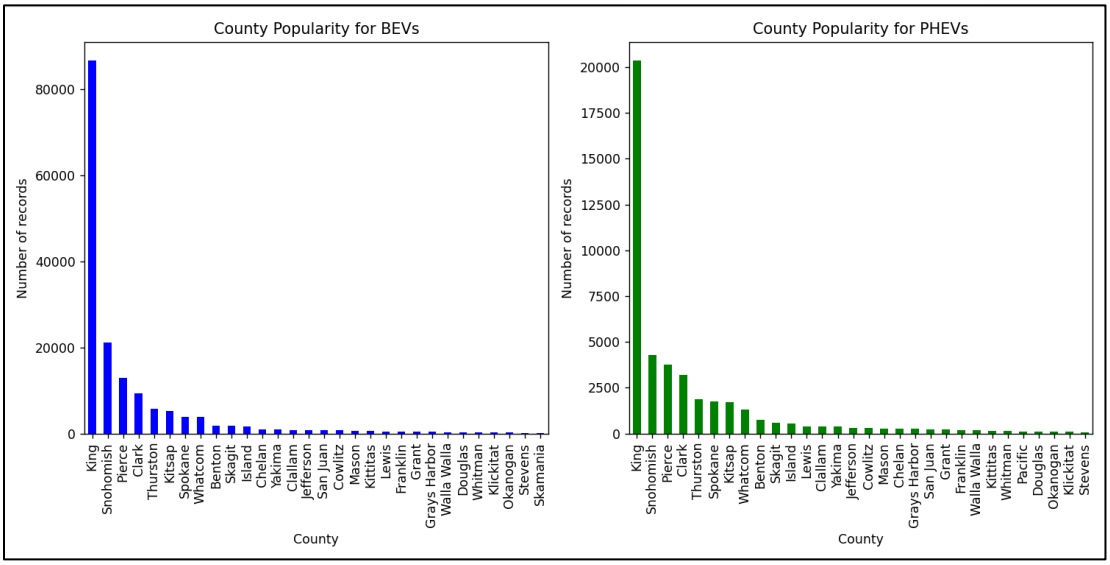

Figure 11: cars distribution over counties.

The bar charts above show that “king” is the most popular county for both types of EVs.

City:

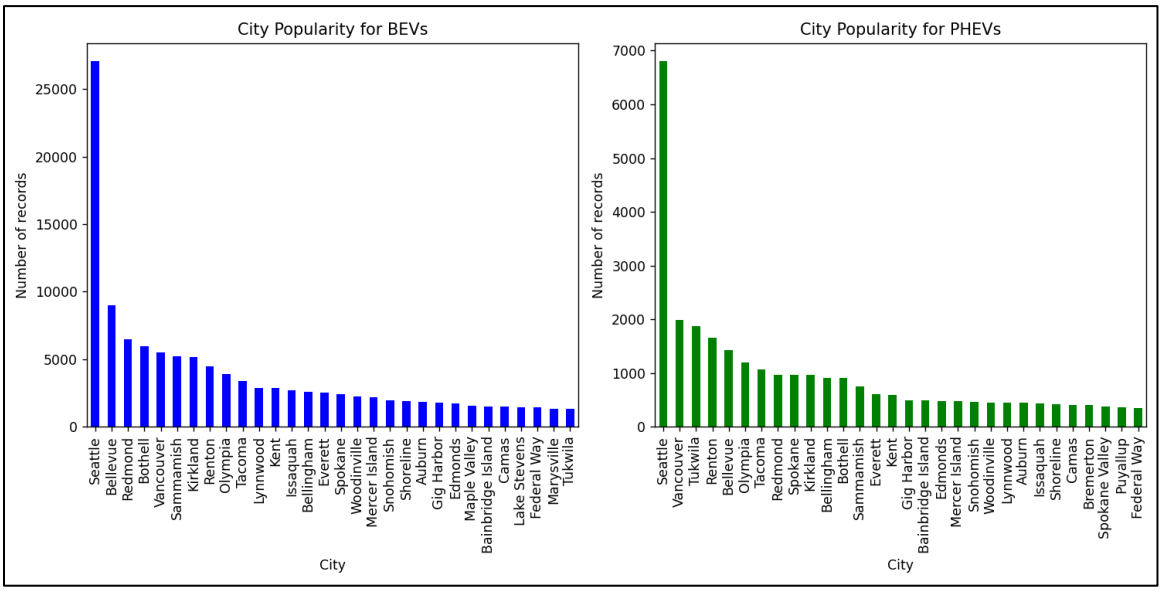

Figure 12: cars distribution over cities

It’s shown that “Seattle” is the dominant city for both car types.