CSE 676-B: Deep Learning, Spring 2024

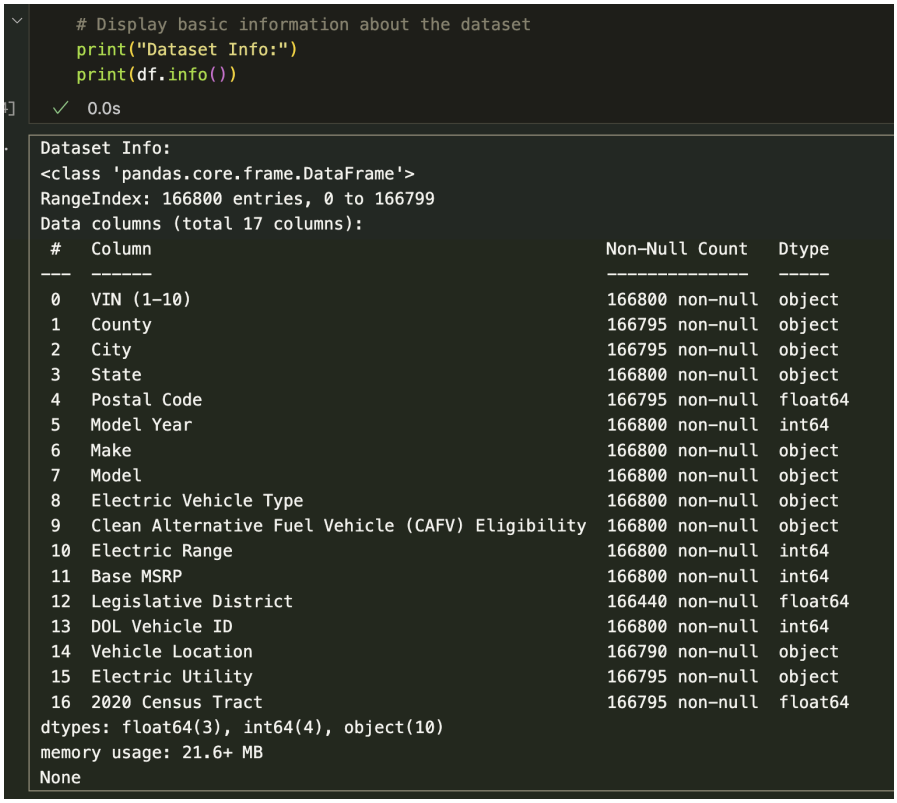
Assignment 0 From Data to ML and NN Models

1. Provide brief details about the nature of your dataset. What is it about? What type of data are we encountering? Provide the main statistics about the entries of the dataset (mean, std, number of missing values, etc.)

This dataset contains information about the Battery Electric Vehicles (BEVs), Plug-in Hybrid Electric Vehicles (PHEVs), Make, Model, Year, Range, and Clean Alternative Fuel Vehicle (CAFV) Eligibility that are currently registered through the Washington State Department of Licensing (DOL).

The dataset comprises a mix of categorical and numerical data types.

Info of dataset:



```
    # Display basic information about the dataset
    print("Dataset Info:")
    print(df.info())
```

✓ 0.0s

```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 166800 entries, 0 to 166799
Data columns (total 17 columns):
 #   Column                                             Non-Null Count   Dtype
---  ------                                             --------------   -----
 0   VIN (1-10)                                         166800 non-null  object
 1   County                                             166795 non-null  object
 2   City                                               166795 non-null  object
 3   State                                              166800 non-null  object
 4   Postal Code                                        166795 non-null  float64
 5   Model Year                                         166800 non-null  int64
 6   Make                                               166800 non-null  object
 7   Model                                              166800 non-null  object
 8   Electric Vehicle Type                              166800 non-null  object
 9   Clean Alternative Fuel Vehicle (CAFV) Eligibility  166800 non-null  object
 10  Electric Range                                     166800 non-null  int64
 11  Base MSRP                                          166800 non-null  int64
 12  Legislative District                               166440 non-null  float64
 13  DOL Vehicle ID                                     166800 non-null  int64
 14  Vehicle Location                                   166790 non-null  object
 15  Electric Utility                                   166795 non-null  object
 16  2020 Census Tract                                  166795 non-null  float64
dtypes: float64(3), int64(4), object(10)
memory usage: 21.6+ MB
None
```

Summary statistics of the dataset:

```
# Display summary statistics of the dataset
print("\nSummary Statistics:")
print(df.describe())
```

[5] ✓ 0.0s

```
Summary Statistics:
         Postal Code     Model Year  Electric Range      Base MSRP  \
count  166795.000000  166800.000000   166800.000000  166800.000000
mean    98173.713750    2020.341793       61.508993    1152.723171
std      2442.584415       3.001465       93.271747    8661.081091
min      1730.000000    1997.000000        0.000000       0.000000
25%     98052.000000    2018.000000        0.000000       0.000000
50%     98122.000000    2021.000000        0.000000       0.000000
75%     98371.000000    2023.000000       84.000000       0.000000
max     99577.000000    2024.000000      337.000000  845000.000000

       Legislative District  DOL Vehicle ID  2020 Census Tract
count         166440.000000    1.668000e+05       1.667950e+05
mean              29.178941    2.172420e+08       5.297709e+10
std               14.853534    7.727458e+07       1.569754e+09
min                1.000000    4.385000e+03       1.001020e+09
25%               18.000000    1.790741e+08       5.303301e+10
50%               33.000000    2.244045e+08       5.303303e+10
75%               42.000000    2.513421e+08       5.305307e+10
max               49.000000    4.792548e+08       5.603300e+10
```

Number of missing values:

Total rows - rows after deleting NA values = 166800 - 166435 = **365**

2. What kind of preprocessing techniques have you applied to this dataset?

Dropped missing values rows and duplicates.

```
# Drop rows with missing values
df.dropna(inplace=True)

# Remove duplicates
df.drop_duplicates(inplace=True)
```

✓ 0.1s

Dropped the features which are not required or not relevant to the dataset. Also renamed large feature names to small readable ones.

```
columns_to_drop = ['VIN (1-10)', 'State', 'Postal Code', 'Base MSRP', 'Legislative District',
                   'DOL Vehicle ID', 'Vehicle Location', 'Electric Utility', '2020 Census Tract']
df.drop(columns=columns_to_drop, inplace=True, axis=1)

new_names = {'Model Year': 'Model_Year', 'Electric Vehicle Type': 'EV_Type',
             'Clean Alternative Fuel Vehicle (CAFV) Eligibility': 'CAFV_Eligibility', 'Electric Range': 'Rang
df.rename(columns=new_names, inplace=True)
```

✓ 0.0s Python

There were only two numerical features named Range and Make Year, which contained real-world values with no outliers.

Performed One-Hot encoding using Pandas Dummies.

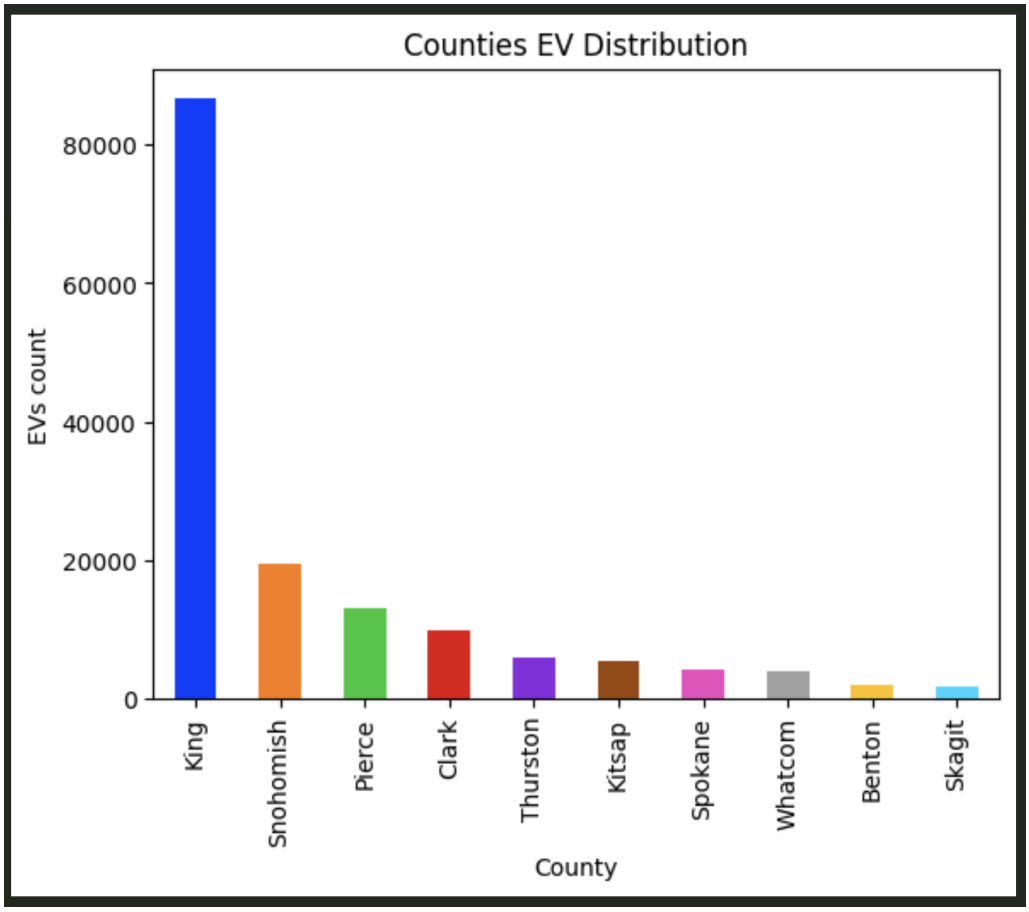
```
# One-Hot Encoding
df_encoded = pd.get_dummies(
    df, columns=['CAFV_Eligibility', 'EV_Type'], prefix=['CAFV', 'EV'])

columns_to_convert = ['CAFV_Clean Alternative Fuel Vehicle Eligible', 'CAFV_Eligibility unknown as battery ra
                      'CAFV_Not eligible due to low battery range', 'EV_Battery Electric Vehicle (BEV)', 'EV_

# Convert boolean values to integers (0 and 1)
df_encoded[columns_to_convert] = df_encoded[columns_to_convert].astype(int)
```

✓ 0.0s Python

+ Code + Markdown

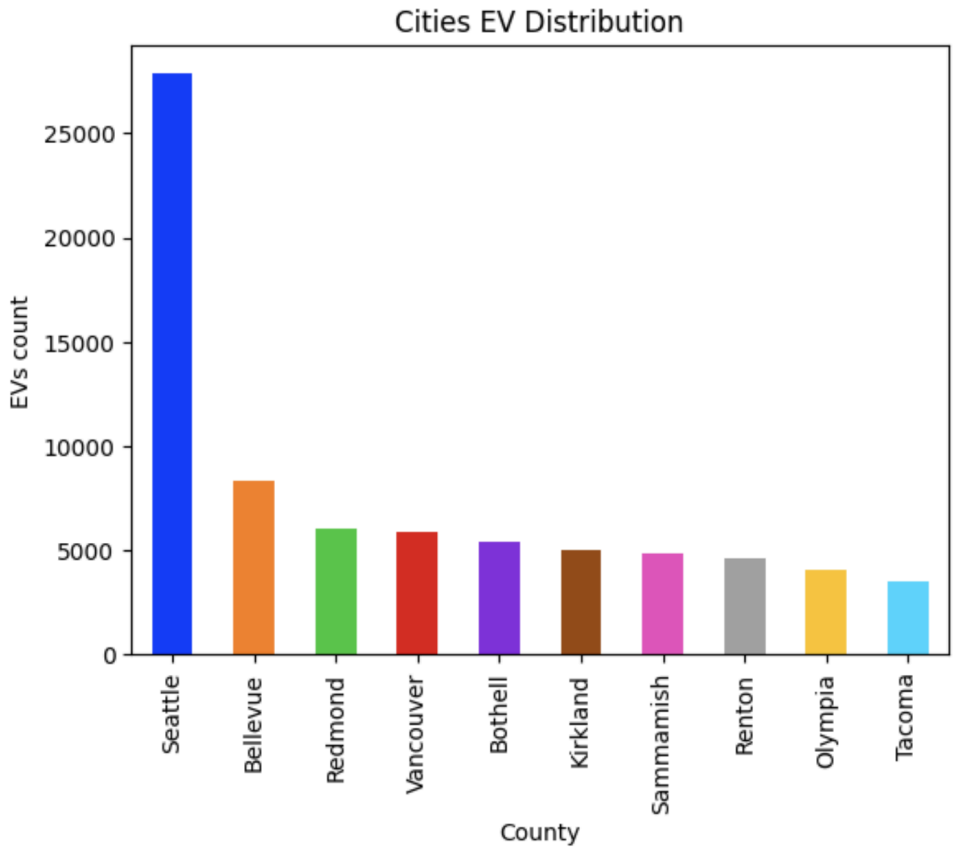
3. Provide at least 5 visualization graphs with a brief description for each graph, e.g. discuss if there are any interesting patterns or correlations.

Top ten counties with the highest number of EVs.



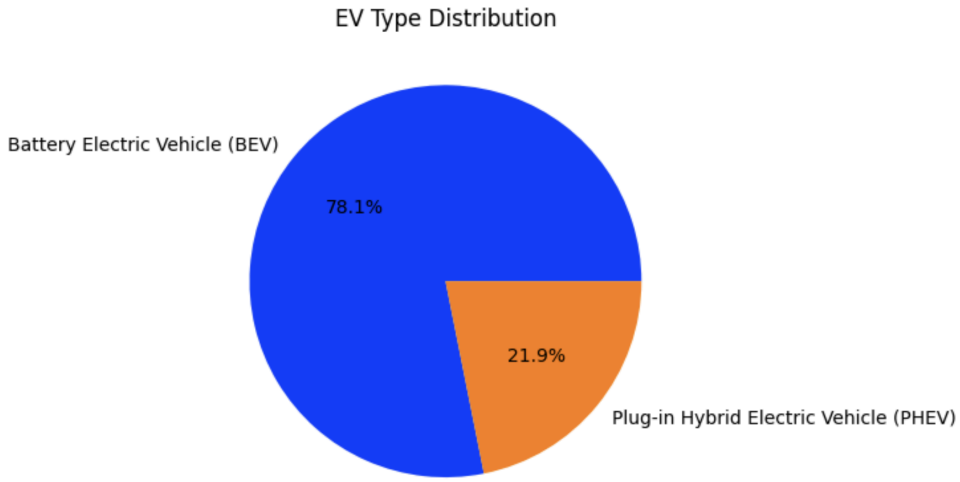
→ King County has the highest number of Electric vehicles in Washington State.

Top ten Cities with the highest number of EVs.



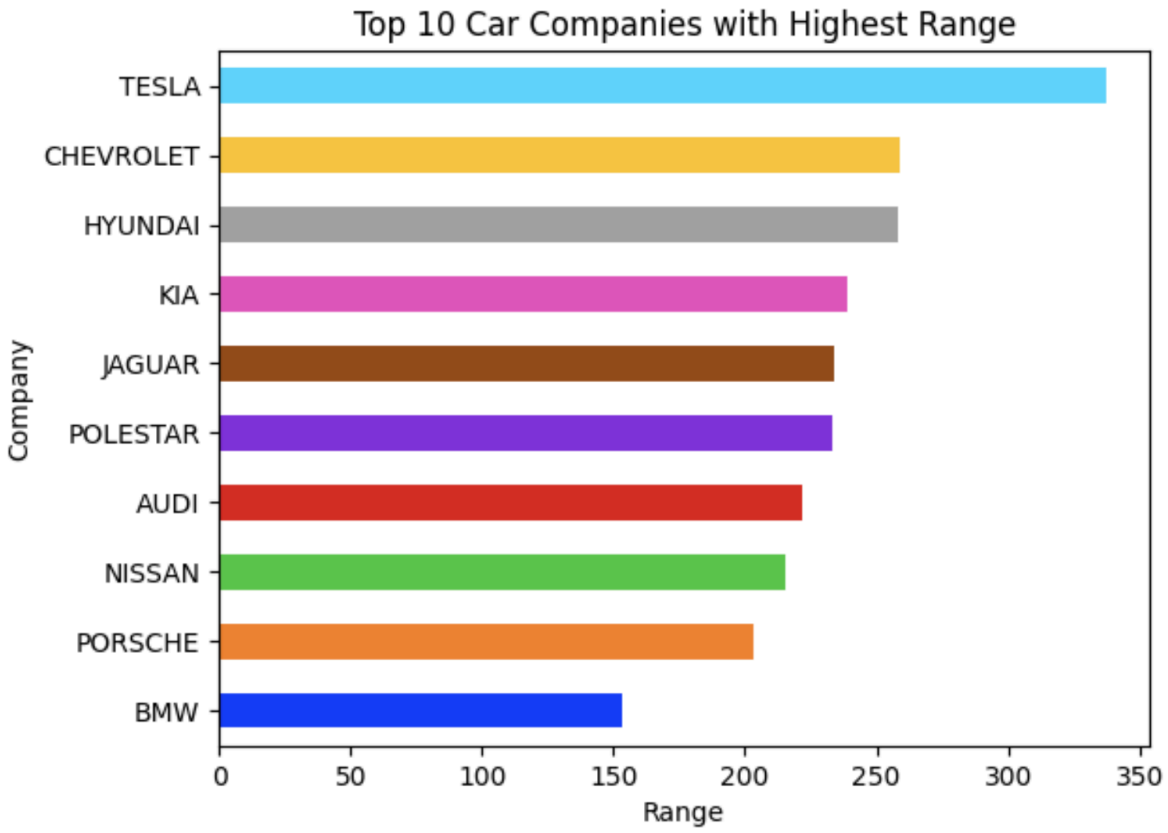
→ Seattle City has the highest number of Electric vehicles in Washington State.

EV Type Distribution.



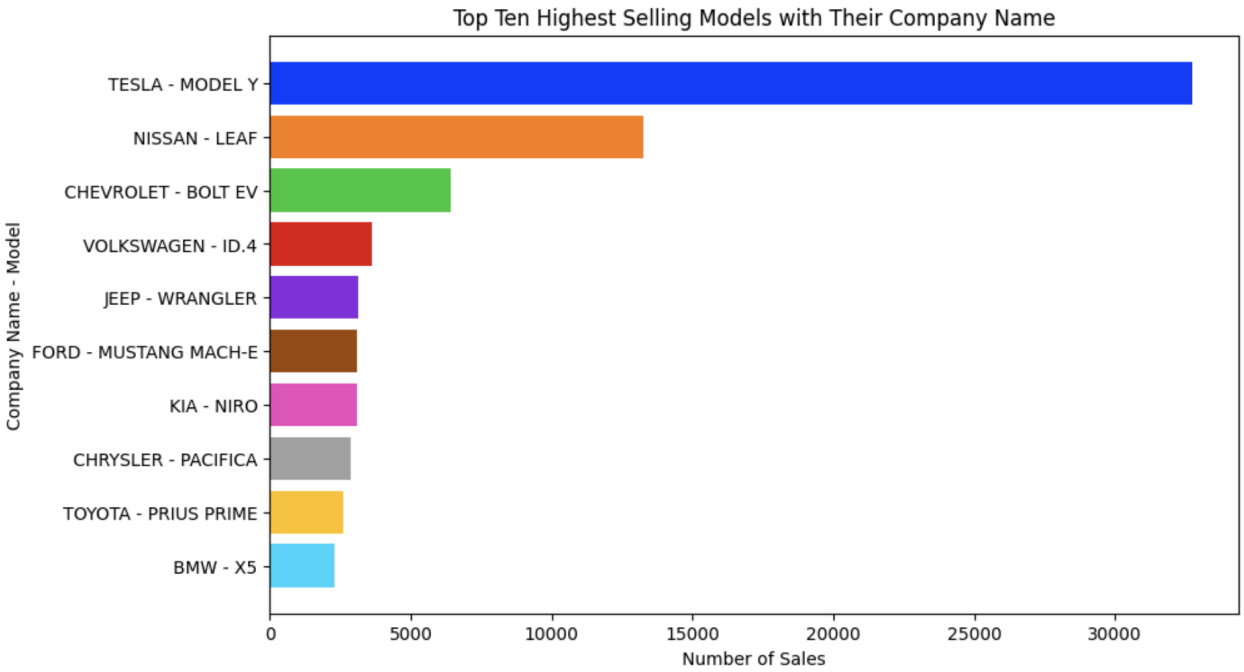
→ Battery-operated EVs are much higher in number compared to Plug-In EVs.

Top ten car companies with the highest range.



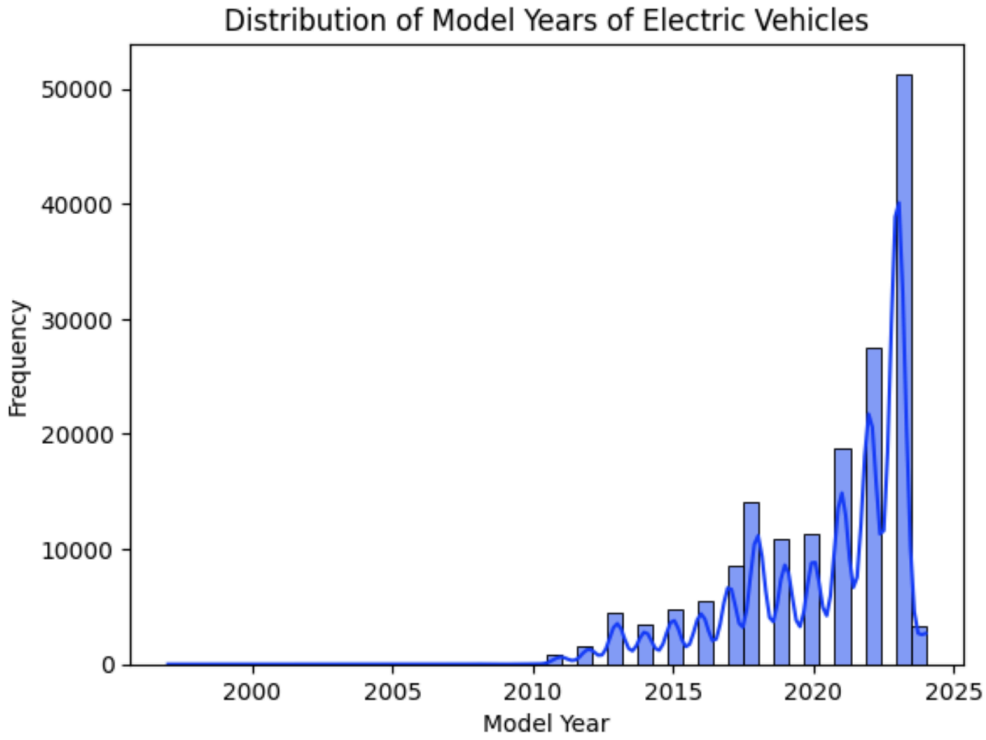
→ Tesla has the highest range of all-electric vehicles.

Top ten highest-selling models with their company name.



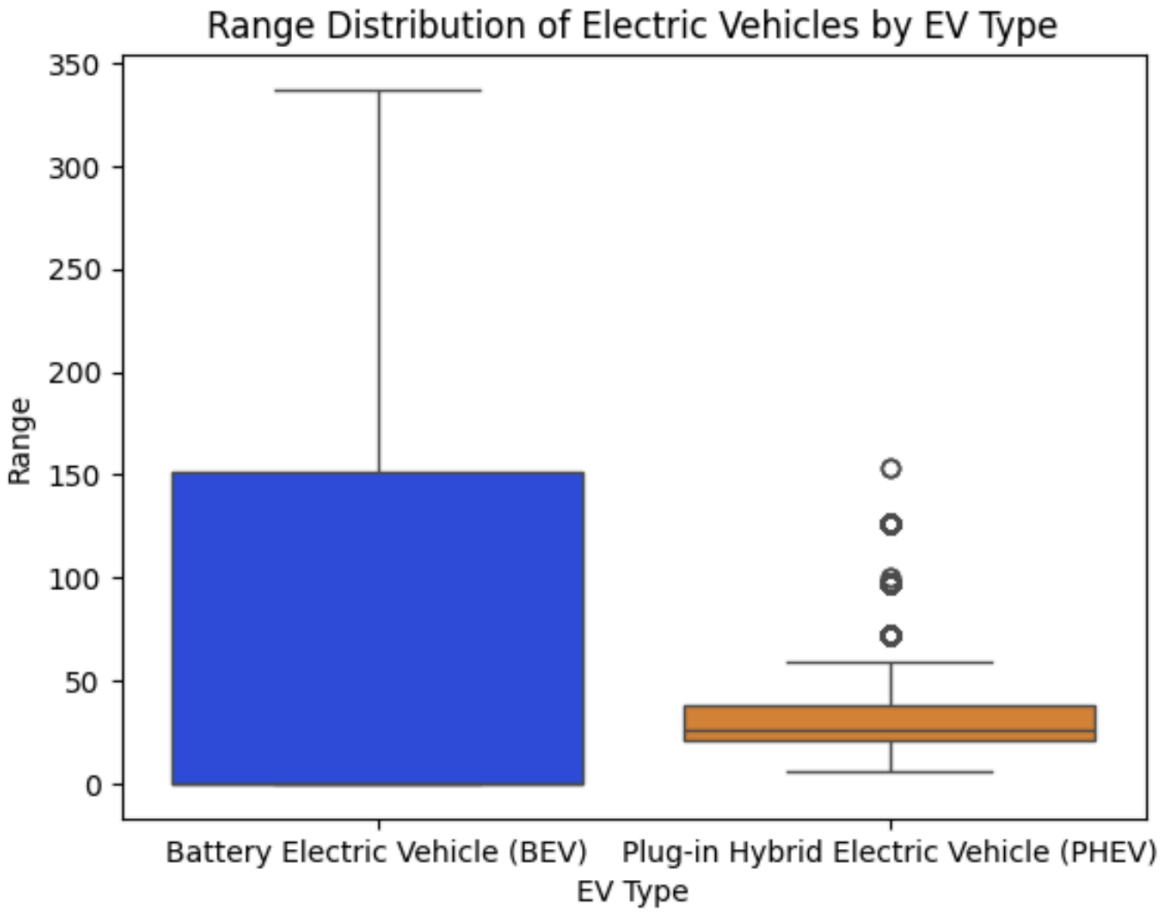
→ Tesla's Model-Y is the most selling EV of all time.

Distribution of Model Years of EVs.



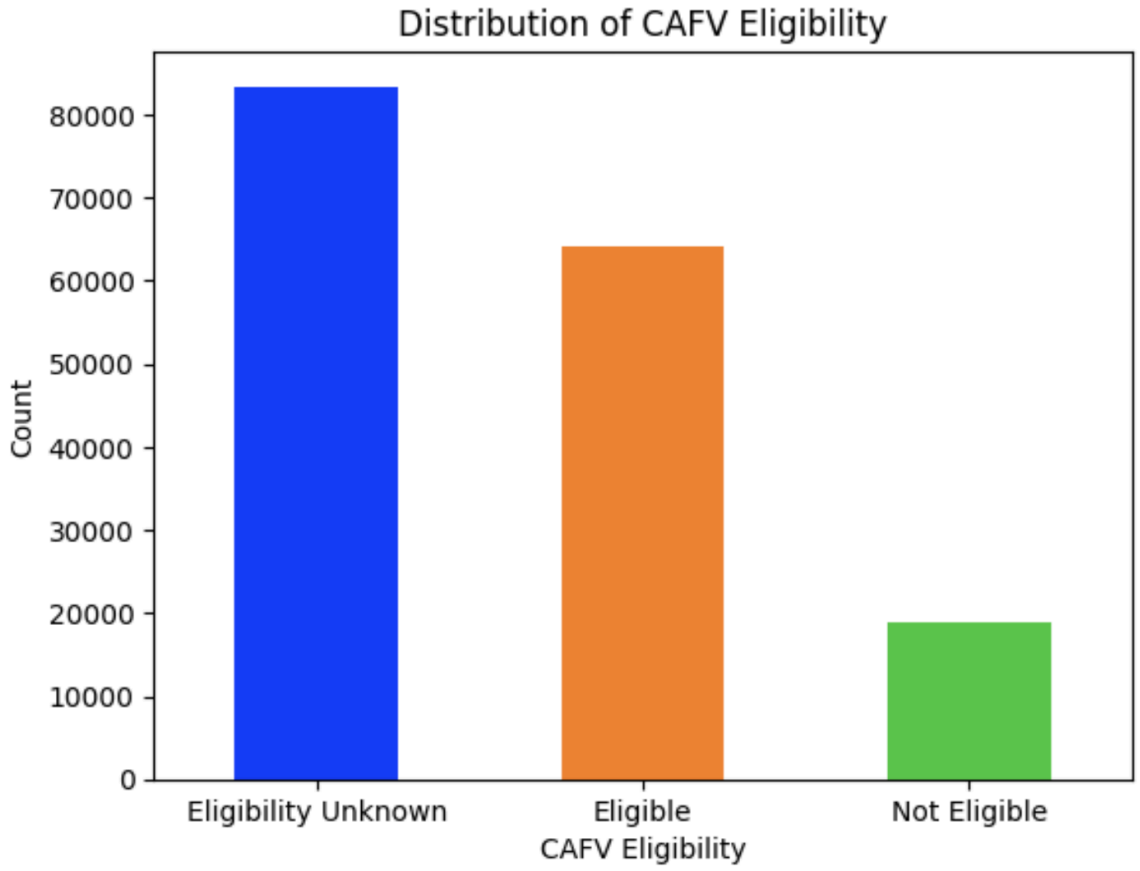
→ EV sales saw a steep growth from the year 2020-2025.

Range distribution of Electric Vehicles by EV types.



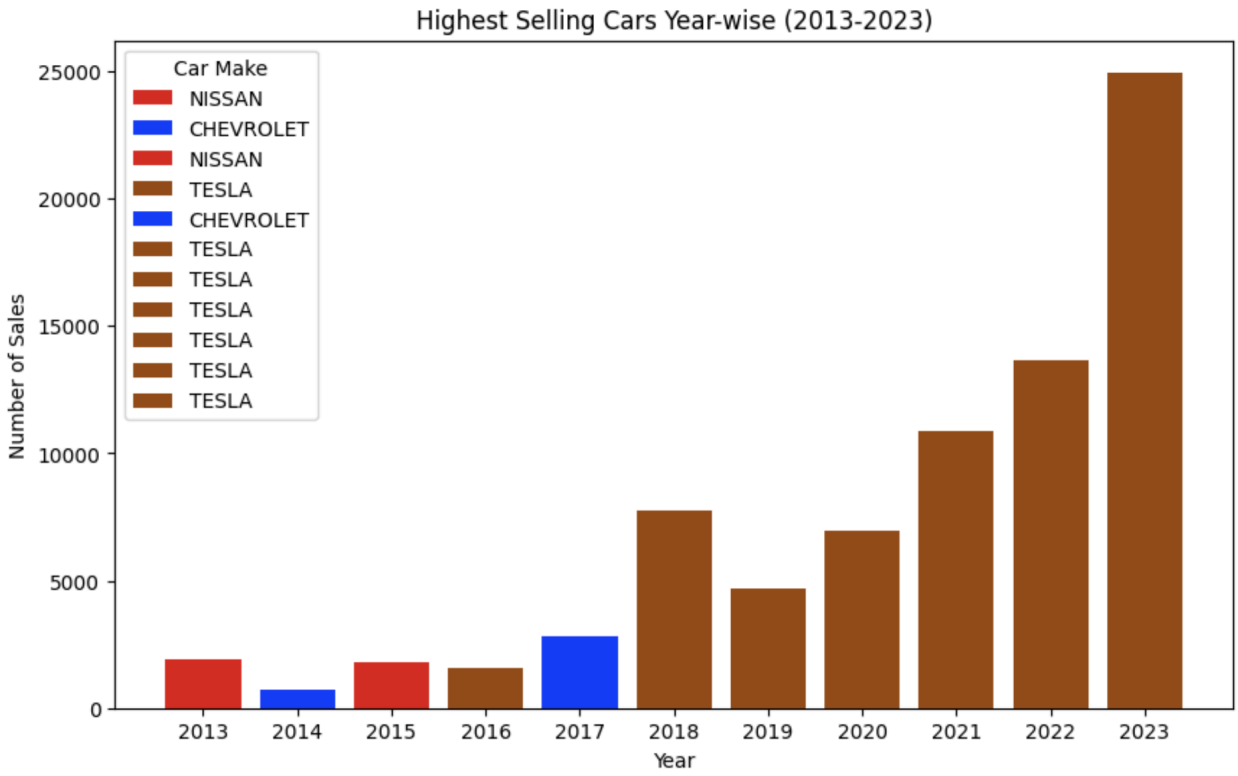
→ From above we can infer that BEV has a much higher range than PHEV.

Distribution of CAFV Eligibility.



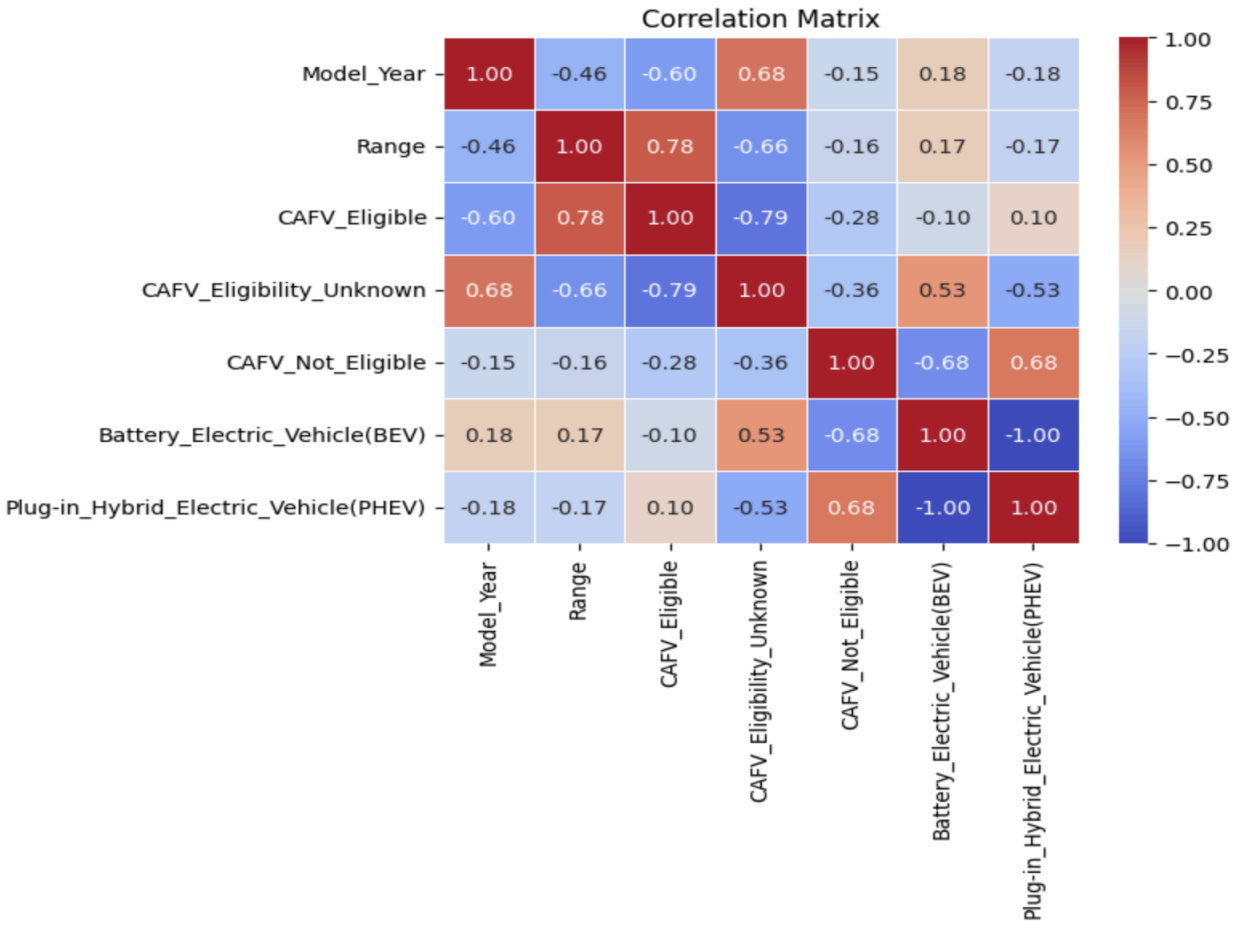
→ Most of the EVs in Washington state have their CAFV eligibility unknown.

Highest selling car year-wise in last decade.



→ From the year 2018-2023, Tesla has been the highest-selling car in Washington state.

Correlation Matrix.



→ Range and CAFV_Eligible have the highest correlation among all the features.

4. Provide brief details and mathematical representation of the ML methods you have used. What are the key features? What are the advantages/disadvantages?

Logistic Regression:

Logistic Regression is a binary classification algorithm used to predict the probability of a binary result based on one or more features. In short classification. It uses the logistic function to find the probability of the dependent variable as a function of the independent variables.

Key Features:

It’s a simple, efficient classification model and also works well with linearly separable data.

Its advantage is it’s easy to implement and interpret however is sensitive to outliers.

Random Forest:

Random Forest is an ensemble learning method that constructs multiple decision trees during training. It builds multiple decision trees and merges them to get accurate and stable predictions.

Key Features:
It handles non-linear relationships better and since it’s an ensemble learning, it combines multiple weak learners to create a strong learner.

Its advantage is it handles high dimensional data well however, it’s computationally expensive.

Support Vector Machines(SVM):

It’s an algorithm used for both classification and regression tasks. It works by finding the hyperplane that best separates the classes in the feature space. SVM aims to find the optimal hyperplane that maximizes the margin between the classes in the feature space.

Key Features:
It handles both linear and non-linear relationships better and is effective in high-dimensional data.

Its advantage is it handles high-dimensional data well however, it’s sensitive to noise.

5. Provide your loss value and accuracy for all 3 methods.

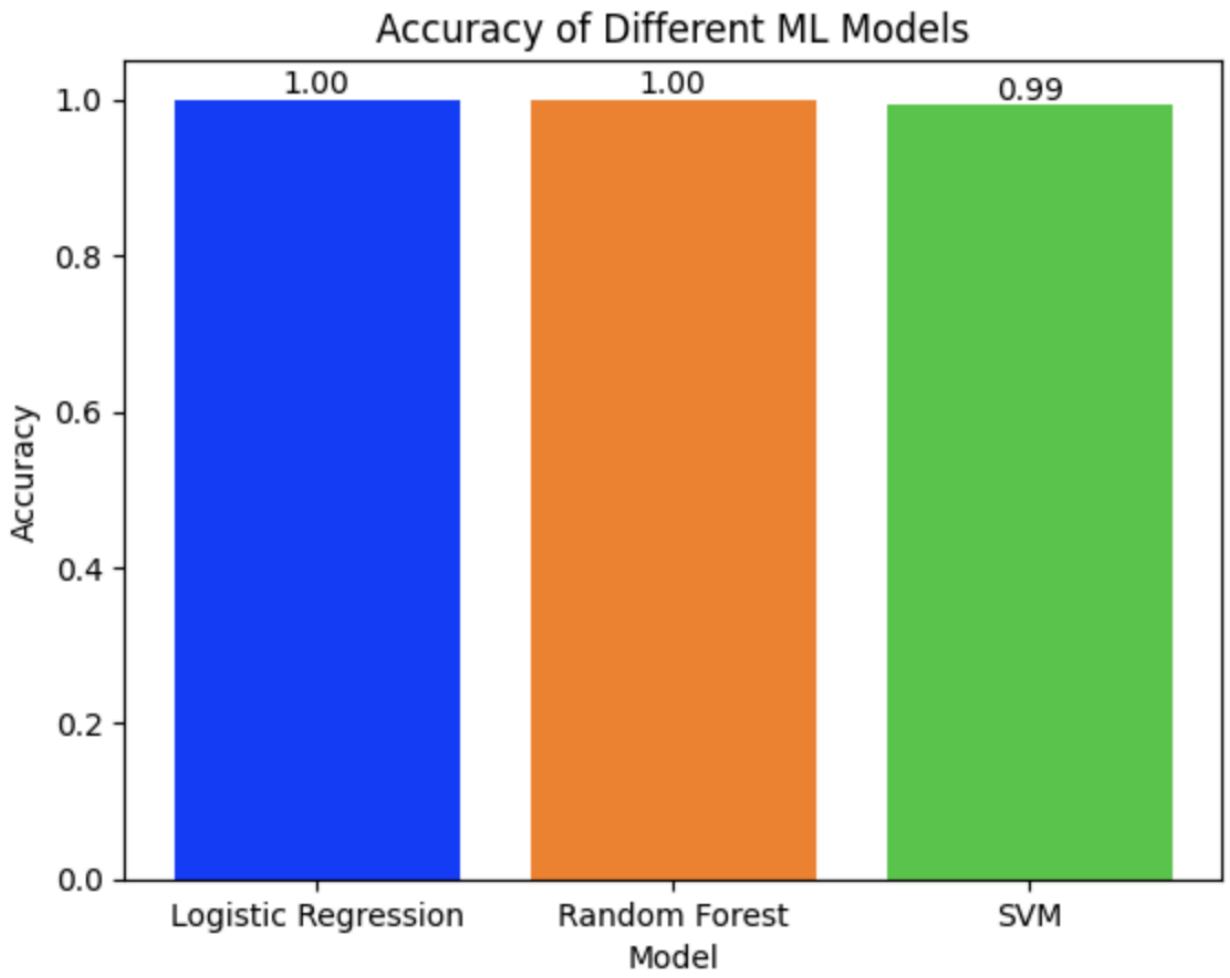
Validation Set

```
--- Validation Set Loss ---
Logistic Regression Loss on Validation Set: 2.2204460492503136e-16
Support Vector Machine(SVM) Loss on Validation Set: 0.6242489786109109
--- Random Forest Classification Report on Validation Set---
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     10231
           1       1.00      1.00      1.00      6413

    accuracy                           1.00     16644
   macro avg       1.00      1.00      1.00     16644
weighted avg       1.00      1.00      1.00     16644

--- Validation Set ---
Logistic Regression Accuracy on Validation Set: 1.0
Random Forest Accuracy on Validation Set: 1.0
Support Vector Machine(SVM) Accuracy on Validation Set: 0.9904470079307859
```

Testing Set

```
--- Testing Set Loss ---
Logistic Regression Loss on Testing Set: 2.2204460492503136e-16
Support Vector Machine(SVM) Loss on Testing Set: 0.6269903262632939
--- Random Forest Classification Report on Testing Set---
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     10307
           1       1.00      1.00      1.00      6336

    accuracy                           1.00     16643
   macro avg       1.00      1.00      1.00     16643
weighted avg       1.00      1.00      1.00     16643


--- Testing Set ---
Logistic Regression Accuracy on Testing Set: 1.0
Random Forest Accuracy on Testing Set: 1.0
Support Vector Machine(SVM) Accuracy on Testing Set: 0.9923090788920267
```

6. Show the plot comparing the predictions vs the actual test data for all methods used. Analyze the results. You can consider accuracy/time/loss as some of the metrics to compare the methods.



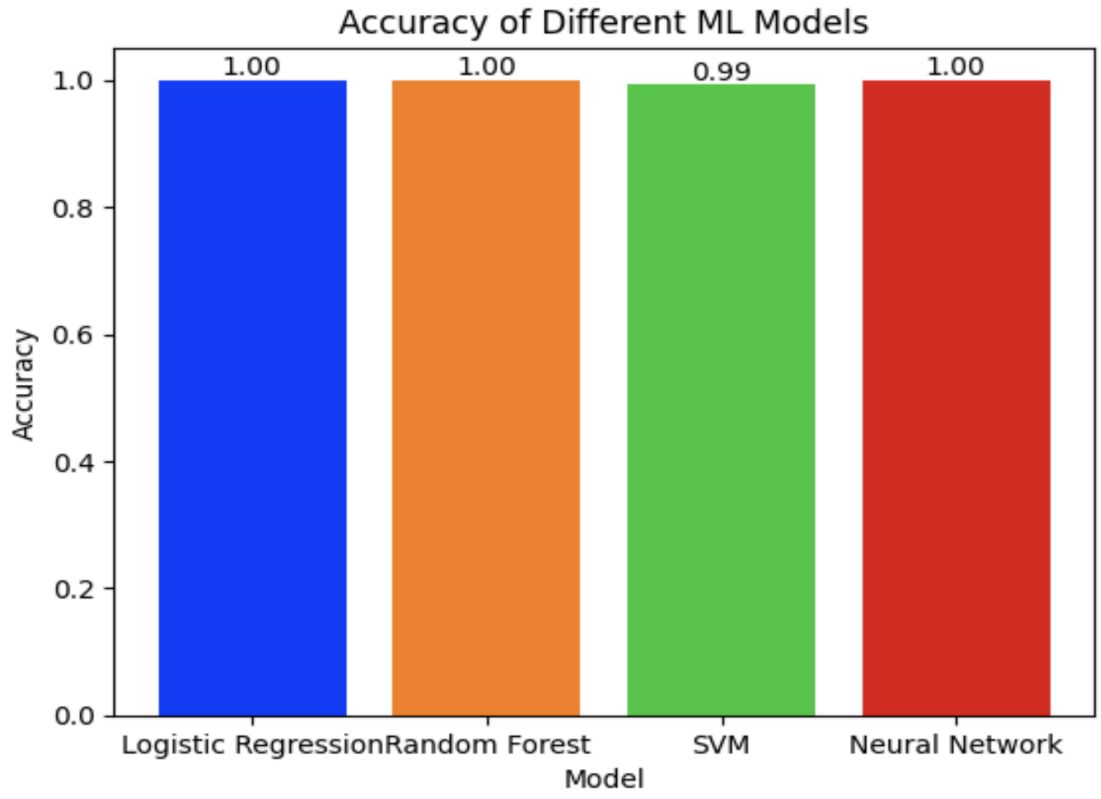
→ All three models Logistic Regression, Random Forest, and SVM gave similar accuracies.

7. Provide the neural network structure you have built to solve the problem defined in Part I. Show the plot. Analyze the results.

Neural Network Structure.

```python
# Define the neural network architecture
class NN(nn.Module):
    def __init__(self, input_size, hidden_size, output_size):
        super(NN, self).__init__()
        self.fc1 = nn.Linear(input_size, hidden_size)
        self.relu = nn.ReLU()
        self.fc2 = nn.Linear(hidden_size, output_size)

    def forward(self, x):
        x = self.fc1(x)
        x = self.relu(x)
        x = self.fc2(x)
        return x
```

Accuracies of different models.



→ All models Logistic Regression, Random Forest, SVM, and Neural Network gave similar accuracies.

References:

1. https://pandas.pydata.org/docs/
2. https://numpy.org/doc/
3. https://matplotlib.org/stable/index.html
4. https://scikit-learn.org/stable/
5. https://seaborn.pydata.org/
6. https://pytorch.org/tutorials/beginner/deep_learning_60min_blitz.html
7. https://optuna.org/
8. https://numpy.org/doc/stable/
9. Part I - Step 3 is based on CSE 574 Machine Learning Quiz 5 submission by Nikhil Gupta