Electrical and Computer Engineering

Department Machine Learning and Data Science - ENCS5341

Assignment #1

Submission deadline: 30/10/2024

Leyan & Rana

The objective of this assignment

is to work with a real-world dataset, focusing on data preprocessing, conducting exploratory data analysis (EDA), and effectively communicating your insights. Dataset Overview: Source and Description of the Dataset: The dataset used for this assignment is titled "Electric Vehicle Population Data" and can be found on Data.gov: 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. Requirements: Provide answers to the following questions as possible as you can. Provide a brief description, including the number of examples, number and type of features, and context. Data Cleaning and Feature Engineering: 1. Document Missing Values: Check for missing values and document their frequency and distribution across features. 2. Missing Value Strategies: If missing values are present, apply multiple strategies (e.g., mean/median imputation, dropping rows) and compare their impact on the analysis. 3. Feature Encoding: Encode categorical features (e.g., Make, Model) using techniques like one-hot encoding. 4. Normalization: Normalize numerical features if necessary for chosen analysis methods. Exploratory Data Analysis: 5. Descriptive Statistics: Calculate summary statistics (mean, median, standard deviation) for numerical features. 6. Spatial Distribution: Visualize the spatial distribution of EVs across locations (e.g., maps). 7. Model Popularity: Analyze the popularity of different EV models (categorical data) and identify any trends. 8. Investigate the relationship between every pair of numeric features. Are there any correlations? Explain the results. Visualization: 9. Data Exploration Visualizations: Create various visualizations (e.g., histograms, scatter plots, boxplots) to explore the relationships between features. 10. Comparative Visualization: Compare the distribution of EVs across different locations (cities, counties) using bar charts or stacked bar charts. Additional Analysis: 11. Temporal Analysis (Optional): If the dataset includes data across multiple time points, analyze the temporal trends in EV adoption rates and model popularity.

Dataset:

This dataset, sourced from the State of Washington, contains details on battery electric vehicles (BEVs) and plug-in hybrid electric vehicles (PHEVs) that are presently registered with the Washington State Department of Licensing. It includes 17 distinct columns of information, covering each vehicle's VIN, the county and city of registration, make, model, electric type, and electric range. The vehicles' model years span from 2013 to the present, with the data being regularly updated by the Washington government.

Data Cleaning and Feature Engineering

1. **Document Missing Values**

In this part, we found all missing data from the whole data set ordered by the attribute name, number of missing fields and their percentage. It turned out from the results that the most missing data belongs to Legislative District field with missing count = 455 and percentage equals to 0.21. to find these result we used isnull() function in python and we found the sum.

```
Missing Values Documentation:
                        Missing Count  Missing Percentage
County                              4            0.001903
City                                4            0.001903
Postal Code                         4            0.001903
Electric Range                      5            0.002379
Base MSRP                           5            0.002379
Legislative District              445            0.211738
Vehicle Location                   10            0.004758
Electric Utility                    4            0.001903
2020 Census Tract                   4            0.001903
```

```
# Count missing values
missing_count = data.isnull().sum()
```

We added additional choice which is to provide field name and show what rows numbers the field has missing in

```
Choose an option: 2
Enter the attribute name:
Input attribute name: City
Number of missing values in 'City': 4
        VIN (1-10) County City State  Postal Code  ...  Legislative District DOL Vehicle ID Vehicle Location Electric Utility 2020 Census Tract
117629  5YJ3E1EB2M    NaN  NaN    NS          NaN  ...                   NaN      179569743              NaN              NaN               NaN
148153  5YJXCAE24H    NaN  NaN    BC          NaN  ...                   NaN      159850029              NaN              NaN               NaN
180881  WBAJA9C50K    NaN  NaN    AE          NaN  ...                   NaN      244582593              NaN              NaN               NaN
205349  1G1RB6S53J    NaN  NaN    BC          NaN  ...                   NaN      477613216              NaN              NaN               NaN
```

To validate this, I went back to excel records to check and it gave me true results.

2. **Missing Value Strategies**

For missing data, we used two methods to handle them. First is using mean imputation for numeric values and most frequented one for non-numeric values

Before:

117631	5YJ3E1EB2M			NS		2021	TESLA	MODEL 3	Battery Ele	Eligibility u	0	0		1.8E+08			

After:

117631	5YJ3E1EB2	King	Seattle	NS	98178.21	2021	TESLA	MODEL 3	Battery Ele	Eligibility u	0	0	28.92995	1.8E+08	POINT (-12	PUGET SO	5.3E+10

Next method we used to drop rows with missing data

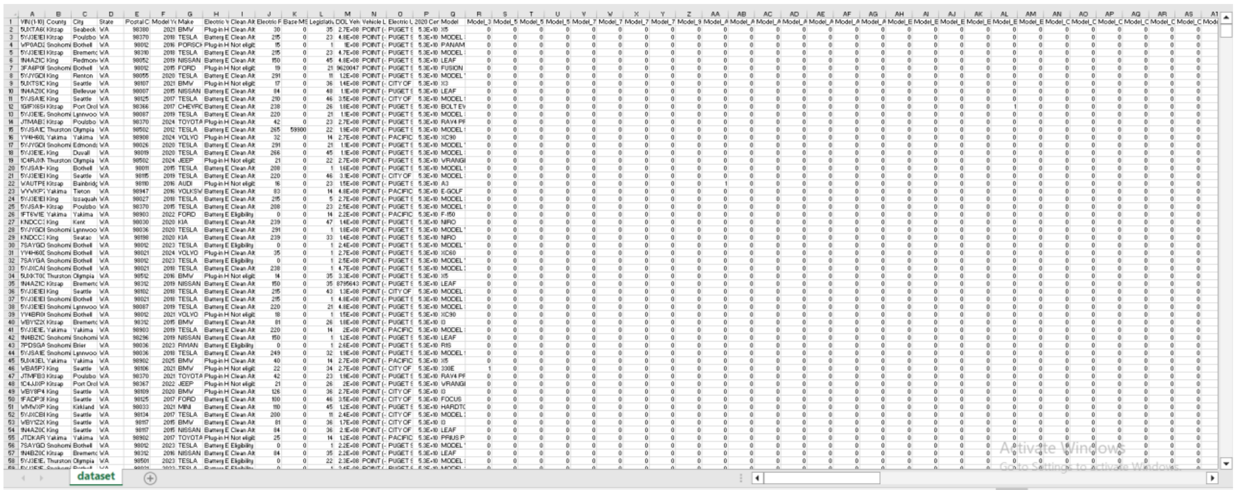
Recording to missing percentage in this dataset which is 0.22%, it is kind of not that much lost, it was handled by imputation and dropping rows.

```
Total cells: 3572805
Missing cells: 485
Total samples: 210165
Samples with missing data: 456 (0.22%)
Missing values handled by imputation.
```

```
Before deletion:
Total samples: 210165
Missing cells: 485
Samples with missing data: 456 (0.22%)

After deletion:
Total samples: 209709
Total cells: 3565053
Rows with missing values have been dropped.
```

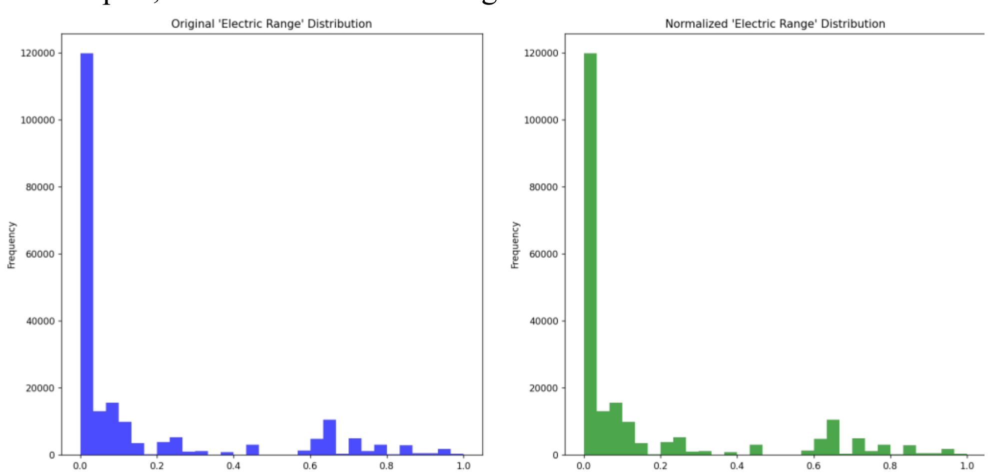
3. Feature Encoding

For this part, we used one hot encoding for Model field and added it to the original dataset
This is how the dataset look like after applying one hot encoding to the dataset for feature 'Model'



4. Normalization

For this part, I normalized electric range field with min-max scaler and this is the results



After normalization before normalization

Electric Range
0.089020772
0.637982196
0.044510386
0.637982196
0.445103858
0.056379822
0.863501484
0.050445104

Electric Range
30
215
15
215
150
19
291
17
84
210
238
220

Exploratory Data Analysis:

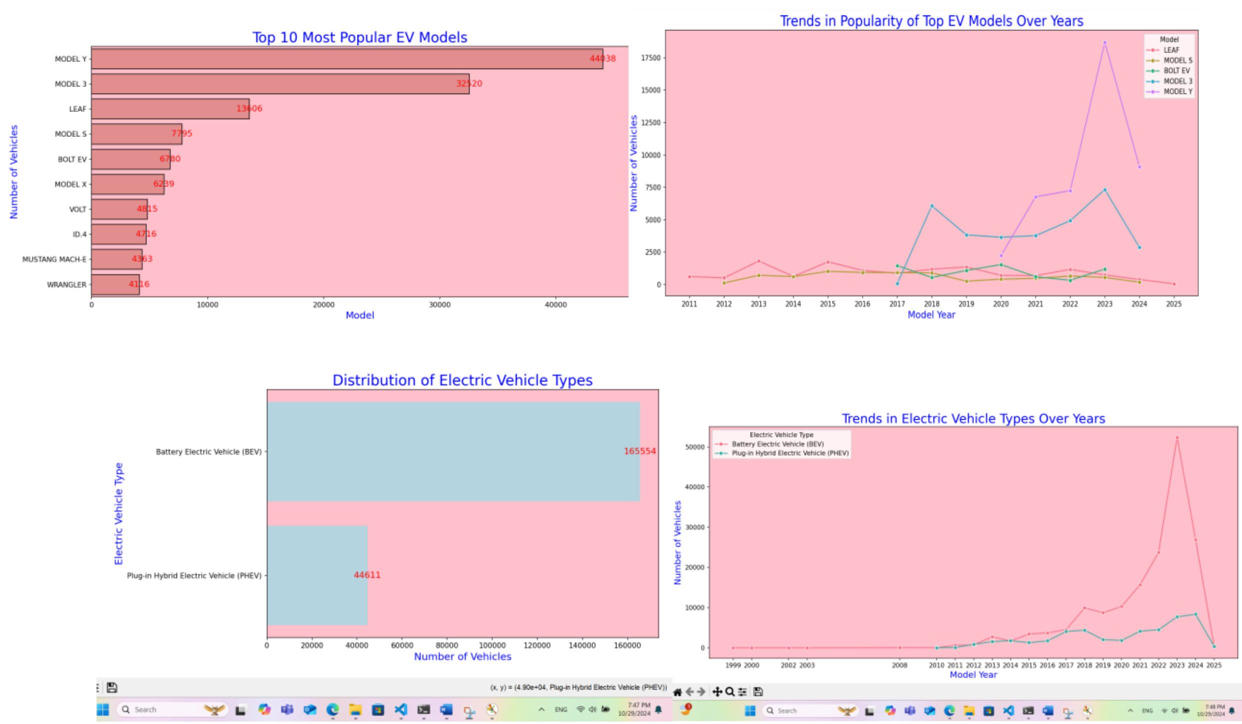
5. **Descriptive Statistics**
   In this part, we found statistics (mean, median, standard deviation) for numerical features.

```
Choose an option: 7
Descriptive Statistics for 'Legislative District':
Mean: 28.929954224680532
Median: 32.0
Standard Deviation: 14.892600518270855


Descriptive Statistics for 'Base MSRP':
Mean: 897.6768890369243
Median: 0.0
Standard Deviation: 7653.4975599676545


Descriptive Statistics for 'Electric Range':
Mean: 0.15015501824807212
Median: 0.0
Standard Deviation: 0.2580776709432204


Descriptive Statistics for 'Postal Code':
Mean: 98178.20940612198
Median: 98125.0
Standard Deviation: 2445.4061302251735
```

6. **Spatial Distribution**
   In this part, we used spatial distribution of EVs across locations using map, first we used unique() function to find a list of all countries in the dataset, then we used FIPS codes and used choropleth map to visualize it

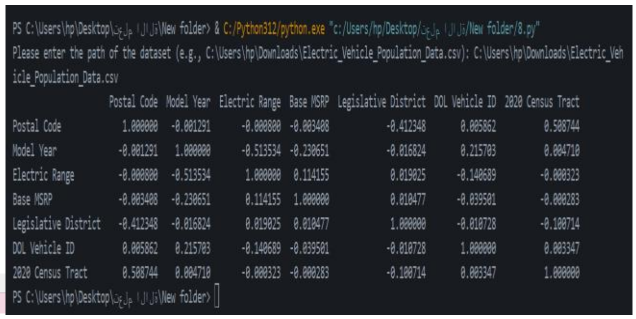
7.Model Popularity

The code loads and cleans an EV dataset, then creates visualizations to explore model popularity, trends over time, county-specific popularity, and type distribution, highlighting top models and trends in EV adoption. Key outputs include bar and line charts displaying top models, geographic distributions, and yearly trends by model and EV type. The output plots offer a comprehensive look at model-specific trends, EV type distribution, and geographic popularity, helping to identify patterns in EV adoption and model preferences.



8.correlation analysis

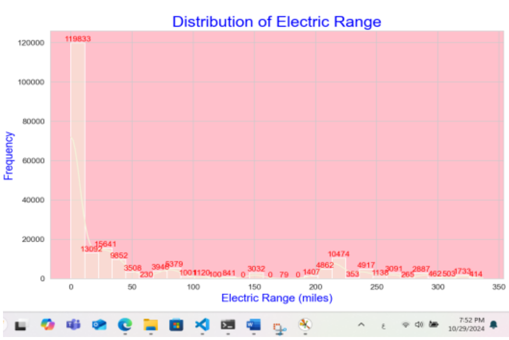
This code loads and cleans an EV dataset, then calculates and visualizes a correlation matrix for numeric features. The heatmap shows the strength and direction of relationships, with some features displaying strong positive or negative correlations, indicating possible dependencies where higher values in one feature could predict changes in another. The annotated heatmap and printed matrix support easy interpretation and further analysis.

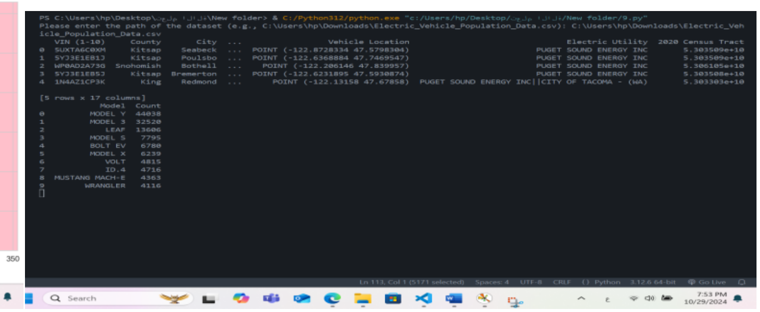


```
PS C:\Users\hp\Desktop\[illegible]\New folder> & C:/Python312/python.exe "c:/Users/hp/Desktop/[illegible]/New folder/8.py"
Please enter the path of the dataset (e.g., C:\Users\hp\Downloads\Electric_Vehicle_Population_Data.csv): C:\Users\hp\Downloads\Electric_Vehicle_Population_Data.csv
                      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
PS C:\Users\hp\Desktop\[illegible]\New folder>
```

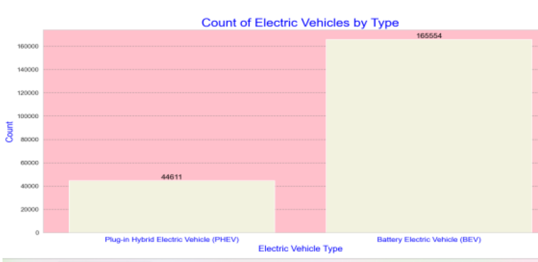
9.Data Exploration Visualizations

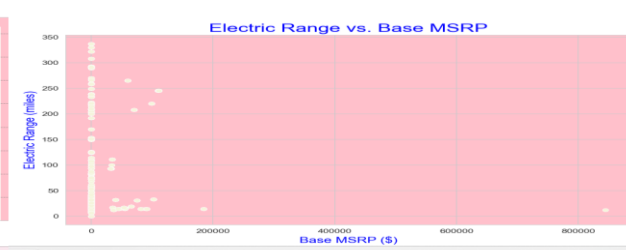
The code loads, cleans, and analyzes EV data, calculating model popularity and vehicle types. It uses various plots (distribution of electric range, bar charts for vehicle types, scatter plots, box plots, and correlation heatmaps) to illustrate relationships among key variables.Visuals are customized with distinct colors (pink background, beige and blue elements, and red bar labels) to enhance readability.

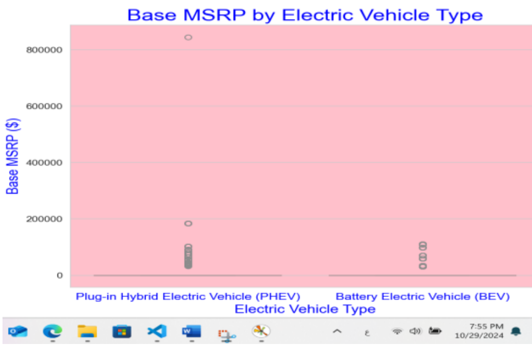


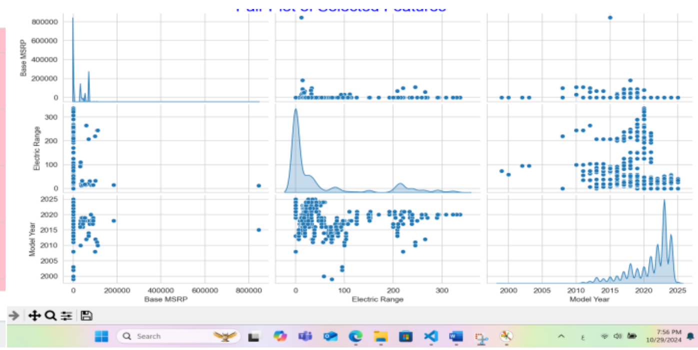


```
PS C:\Users\hp\Desktop\[illegible]\New folder> & C:/Python312/python.exe "c:/Users/hp/Desktop/[illegible]/New folder/9.py"
Please enter the path of the dataset (e.g., C:\Users\hp\Downloads\Electric_Vehicle_Population_Data.csv): C:\Users\hp\Downloads\Electric_Vehicle_Population_Data.csv
   VIN (1-10)    County       City  ...                 Vehicle Location                              Electric Utility  2020 Census Tract
0  5UXTA6C0XM    Kitsap    Seabeck  ...  POINT (-122.8728334 47.5798304)                        PUGET SOUND ENERGY INC       5.303509e+10
1  5YJ3E1EB1J    Kitsap    Poulsbo  ...  POINT (-122.6368884 47.7469547)                        PUGET SOUND ENERGY INC       5.303509e+10
2  WP0AD2A73G  Snohomish   Bothell  ...   POINT (-122.206146 47.839957)                         PUGET SOUND ENERGY INC       5.306105e+10
3  5YJ3E1EB5J    Kitsap  Bremerton  ...  POINT (-122.6231895 47.5930874)                        PUGET SOUND ENERGY INC       5.303508e+10
4  1N4AZ1CP3K      King    Redmond  ...     POINT (-122.13158 47.67858)  PUGET SOUND ENERGY INC||CITY OF TACOMA - (WA)       5.303303e+10

[5 rows x 17 columns]
           Model  Count
0        MODEL Y  44038
1        MODEL 3  32520
2           LEAF  13606
3        MODEL S   7795
4        BOLT EV   6780
5        MODEL X   6239
6           VOLT   4815
7           ID.4   4716
8  MUSTANG MACH-E   4363
9       WRANGLER   4116
```



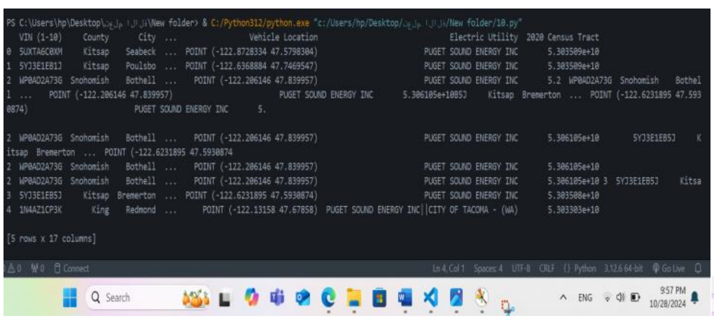


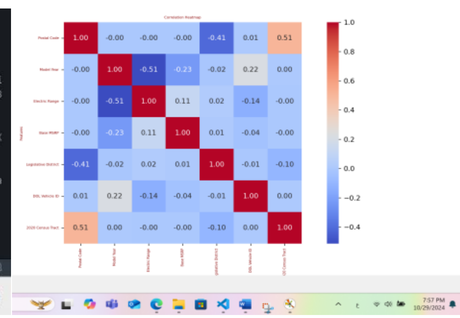




10.Comparative Visualization

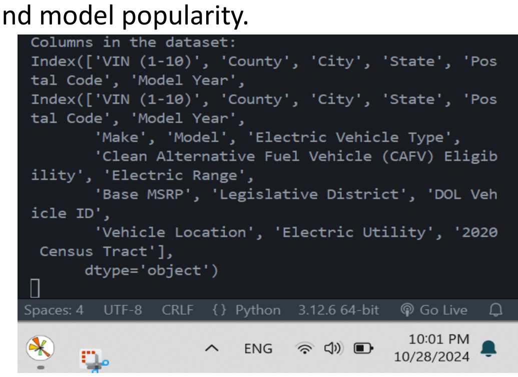
This analyzes and visualizes electric vehicle (EV) data by first loading it from a user-specified path. It calculates correlations among numerical features, highlighting them in a heatmap with a pink background, and identifies the top 15 counties and cities by EV count, shown in bar charts with distinct colors for readability. The code also visualizes the distribution of different EV types across counties and cities using stacked bar charts, which makes it easier to compare the relative popularity of each EV type across locations.

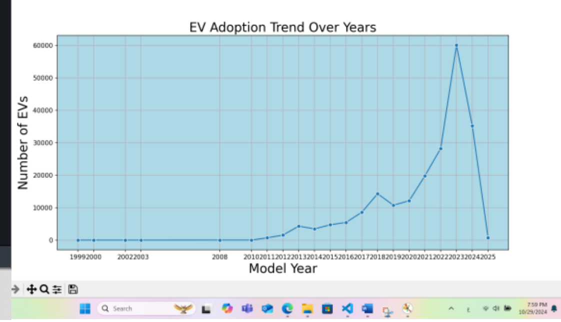


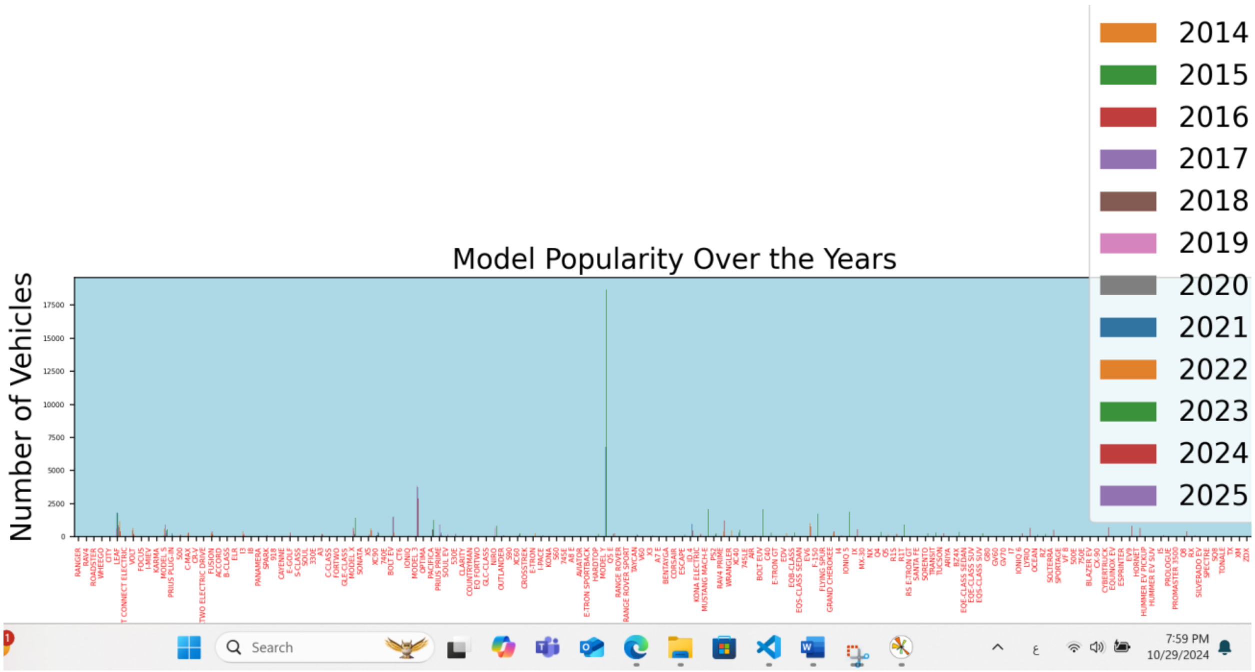


11. Temporal Analysis

If the dataset includes data across multiple time points, analyze the temporal trends in EV adoption rates and model popularity.







Model Popularity Over the Years
Number of Vehicles
17500
15000
12500
10000
7500
5000
2500
0
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
Search
7:59 PM
10/29/2024