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.

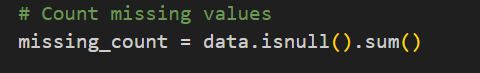
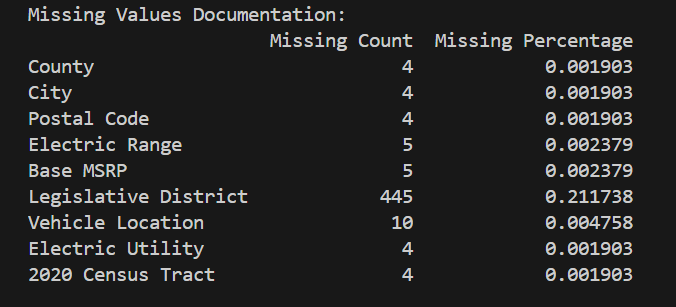
# **Dataset:**

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.

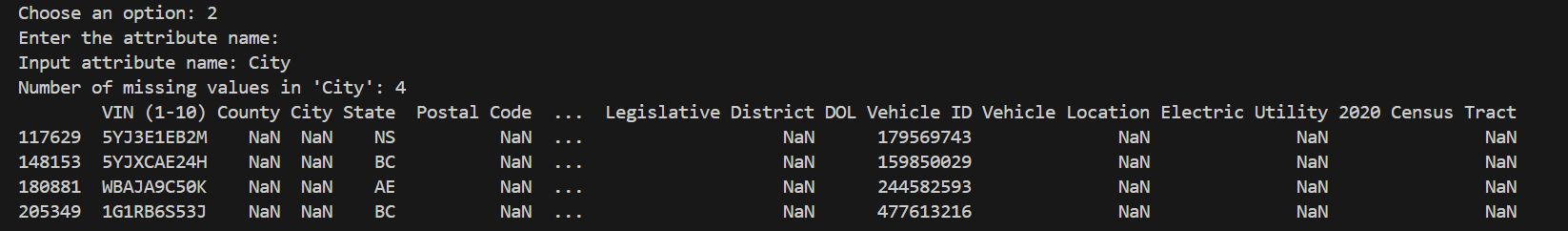
# **Data Cleaning and Feature Engineering**

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.



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



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

1. **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:

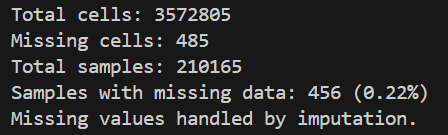
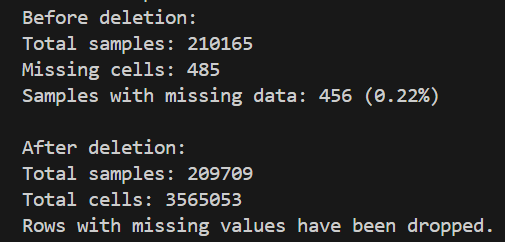


After:



Next method we used to drop rows with missing data

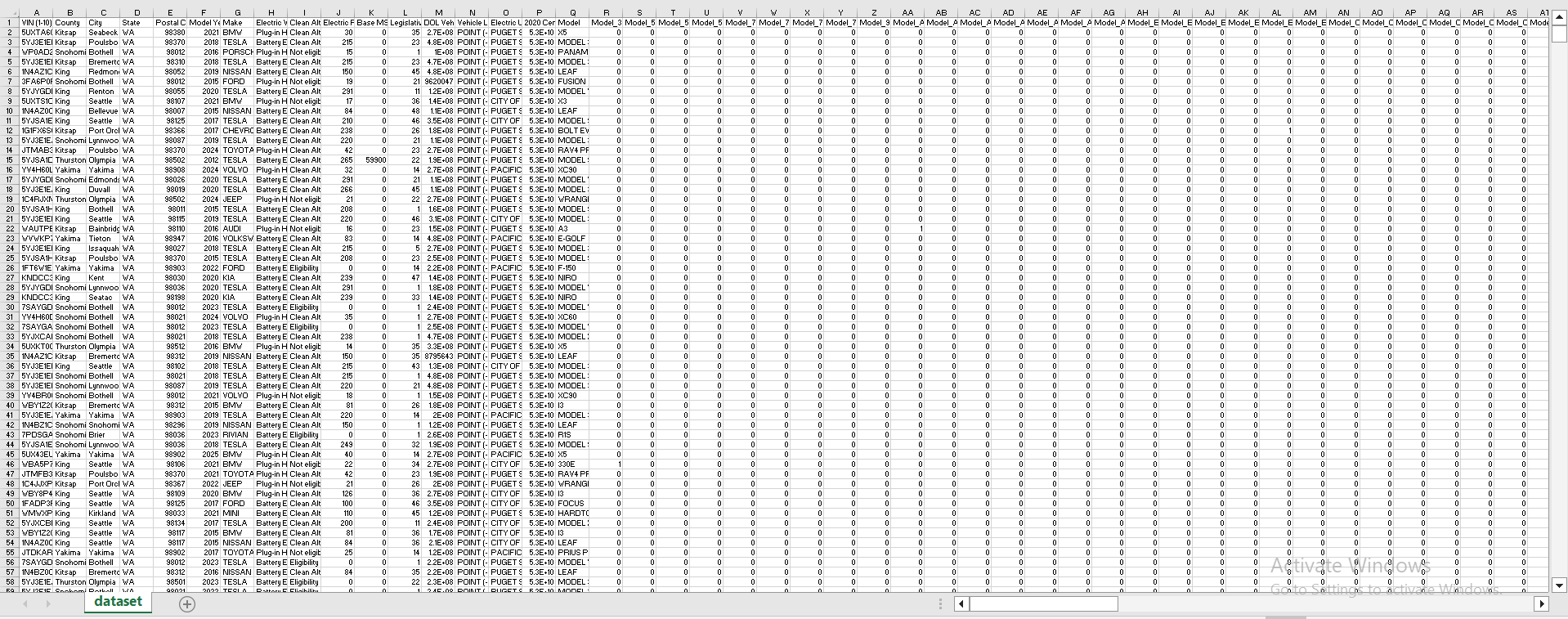
Recording to missing percentage in this dataset which is 0.22%, it is kind of not that much lost, it was handled by imputation and dropping rows.

1. **Feature Encoding**

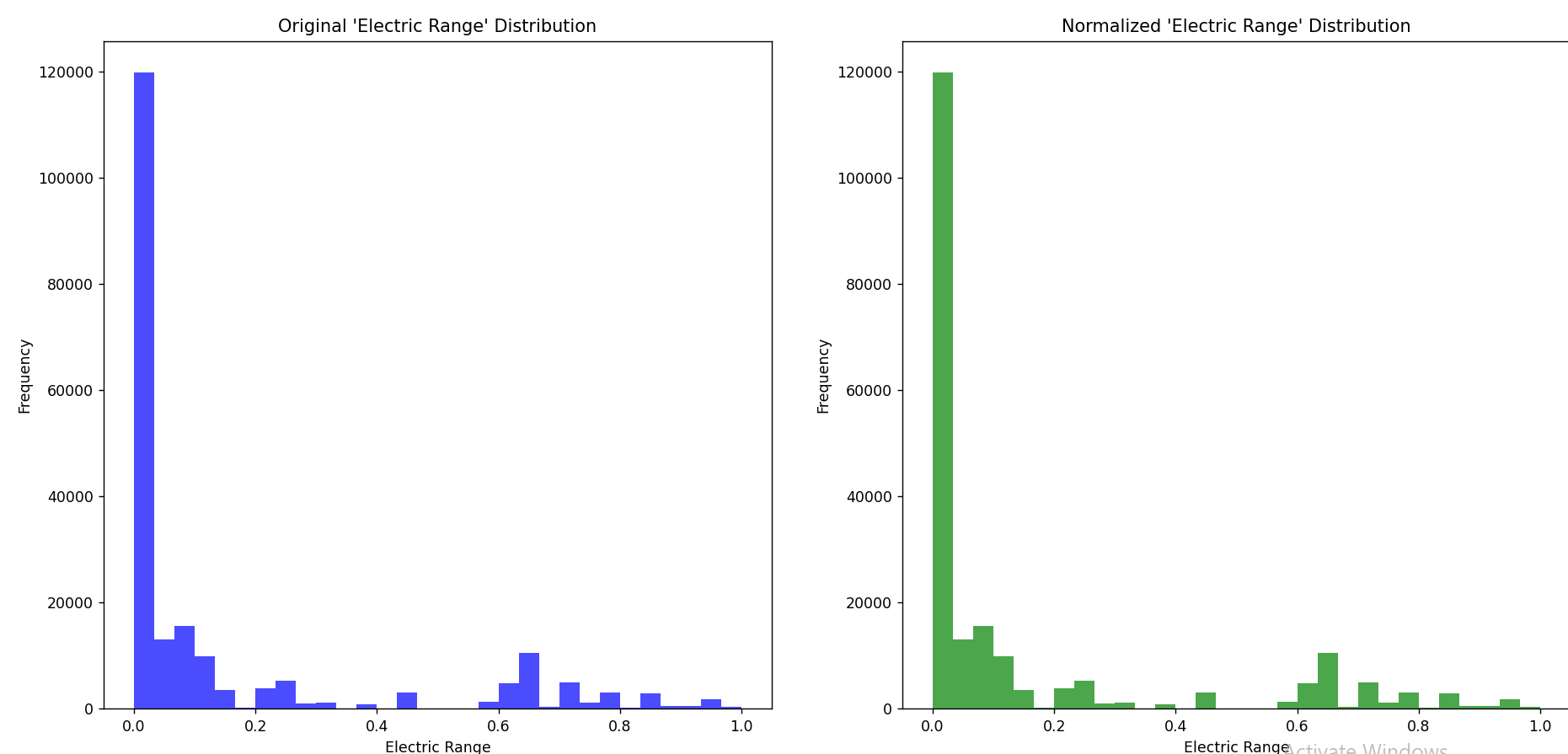
For this part, we used one hot encoding for Model field and added it to the original dataset

This is how the dataset look like after applying one hot encoding to the dataset for feature ‘Model’

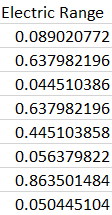
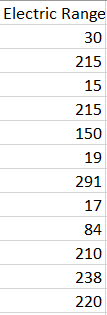


1. **Normalization**

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



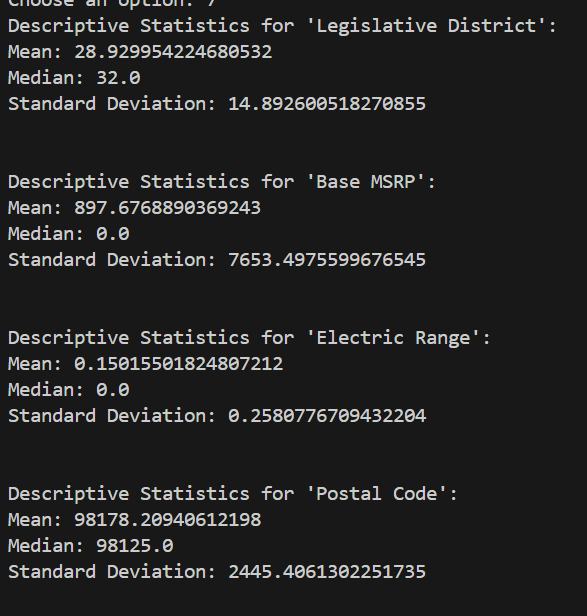
After normalization before normalization

# **Exploratory Data Analysis:**

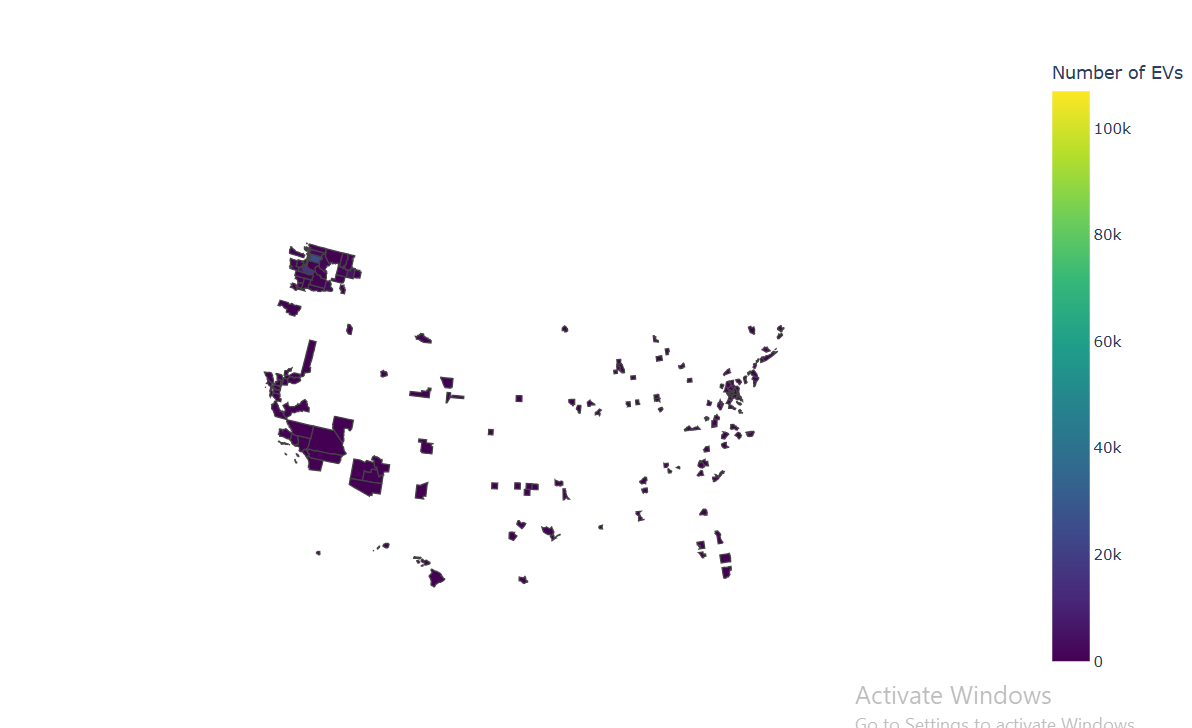
1. **Descriptive Statistics**

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



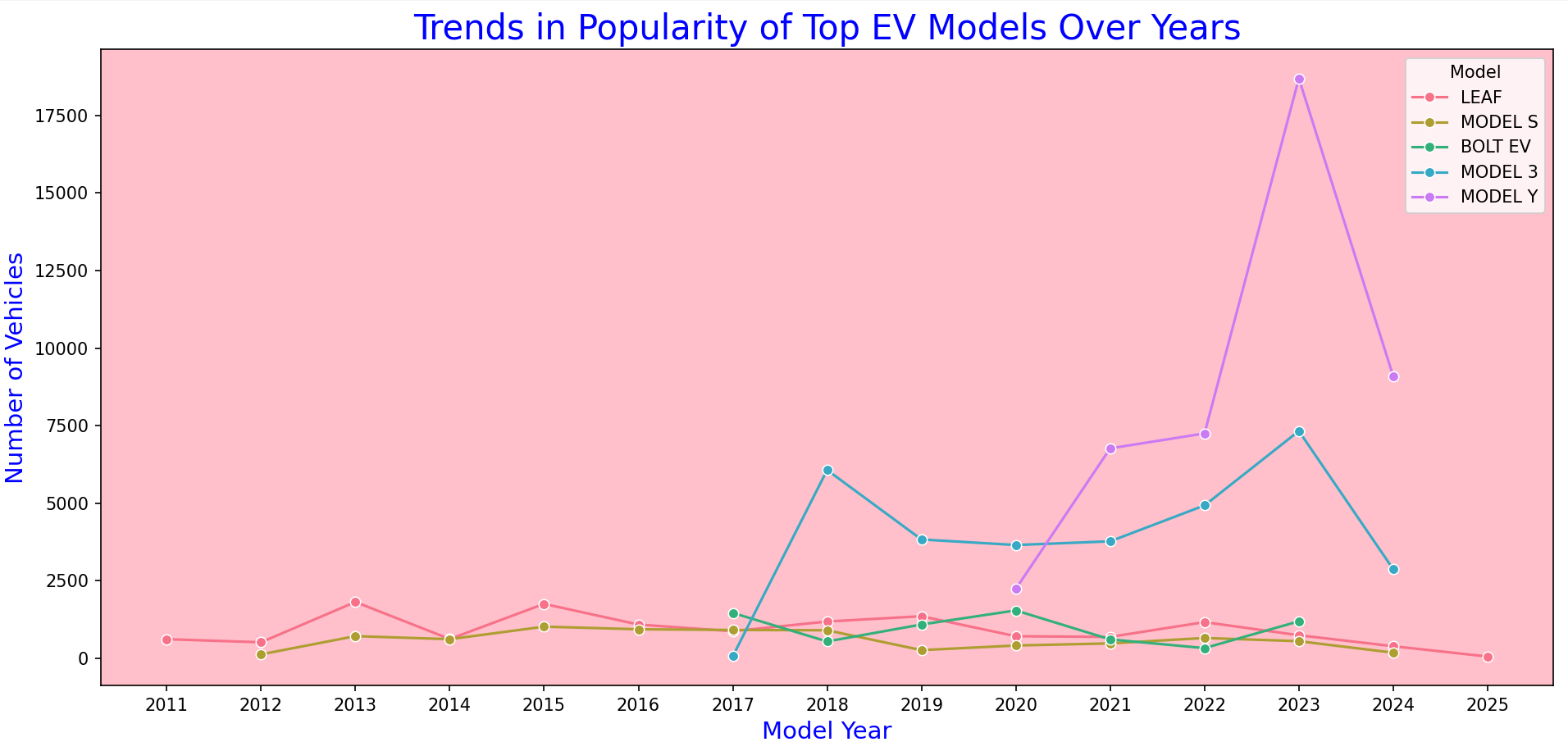
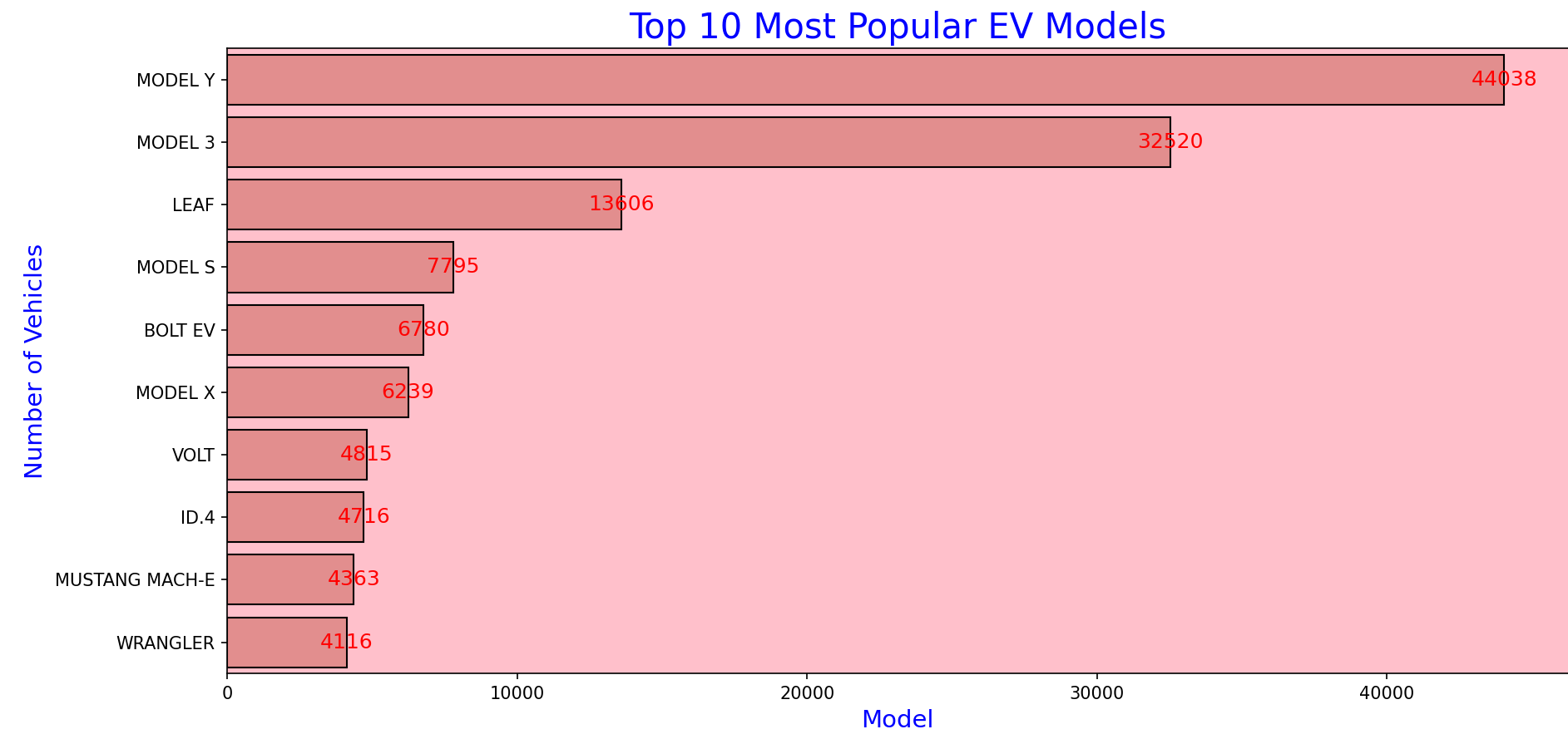
1. **Spatial Distribution**

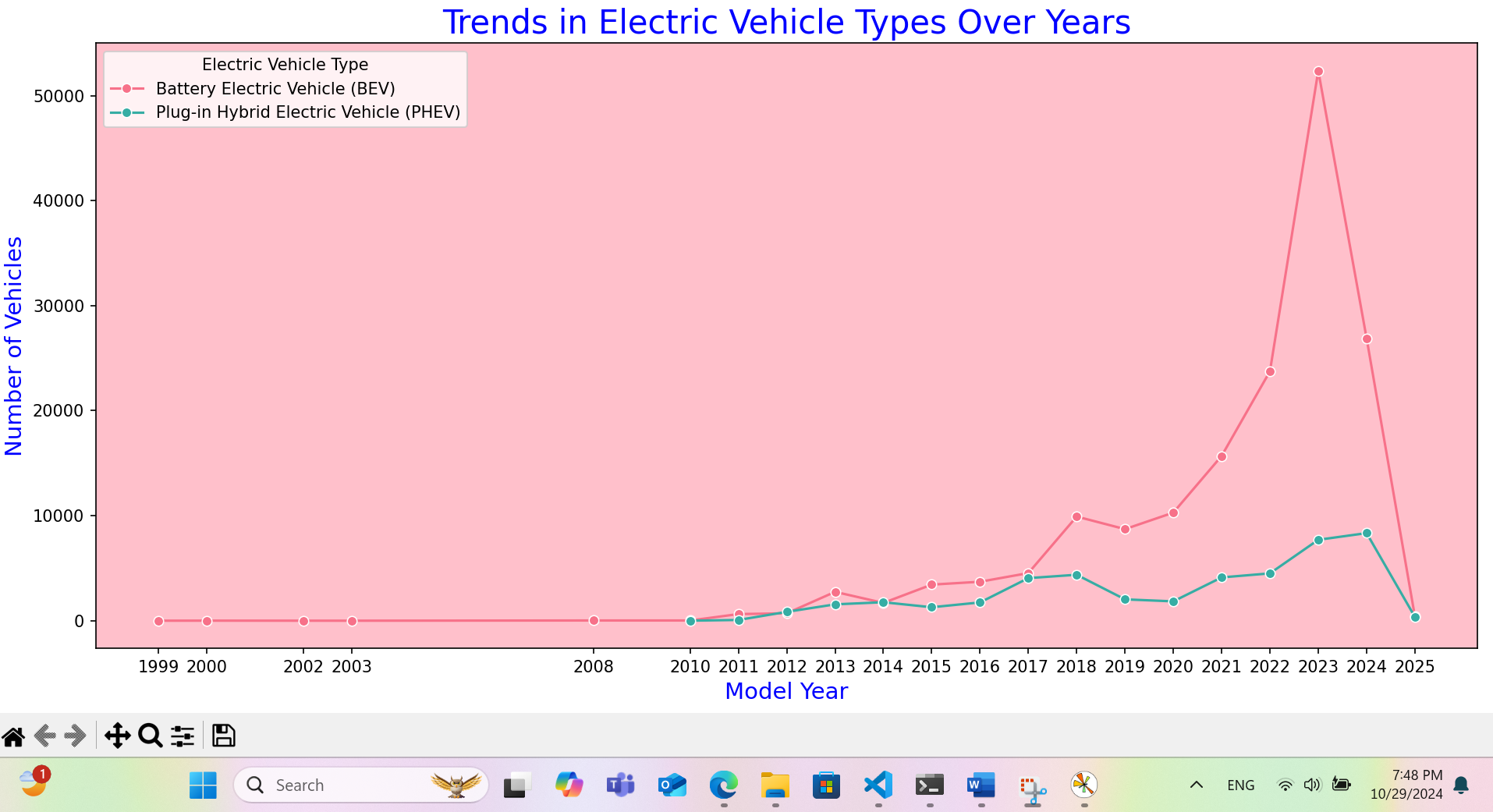
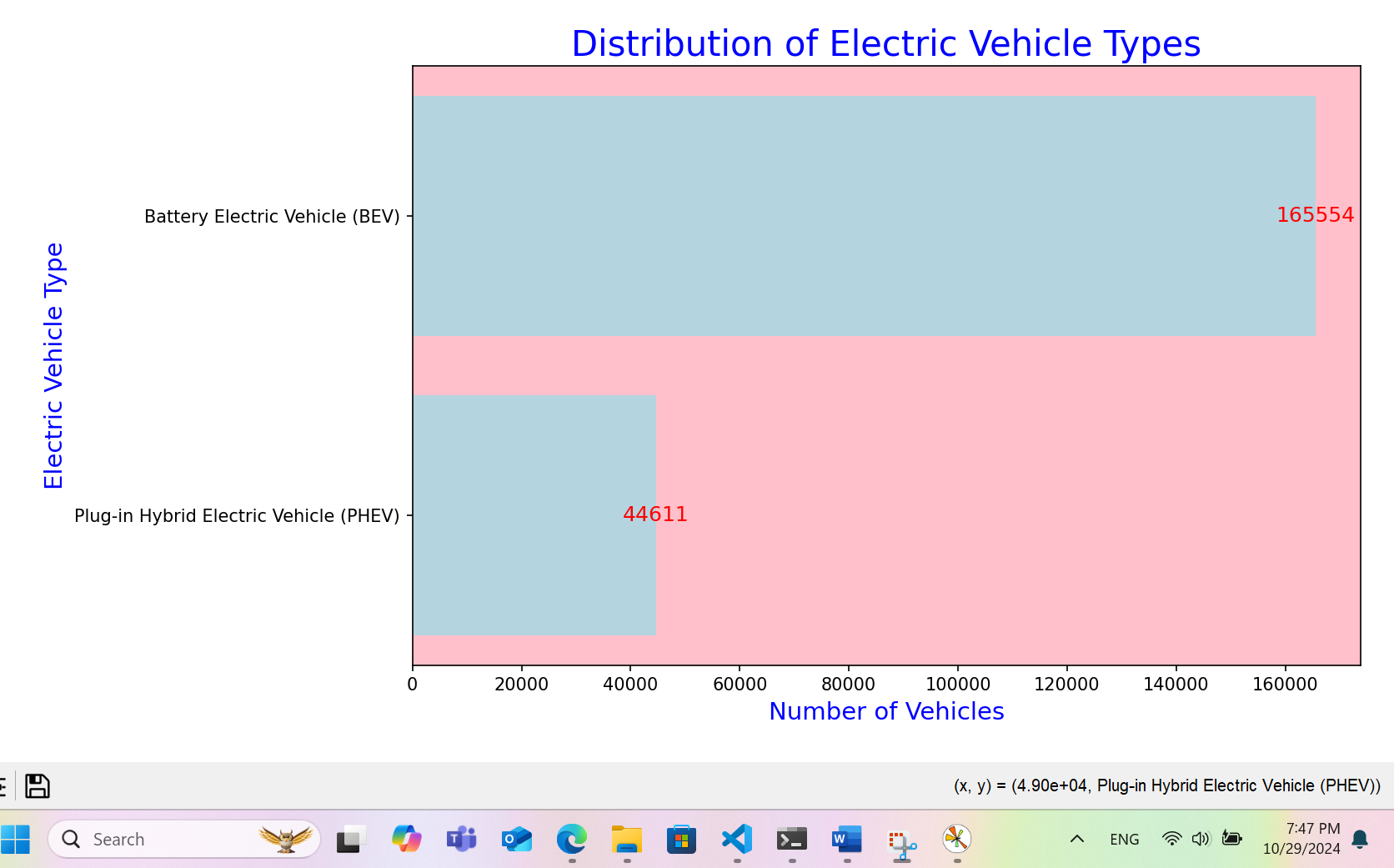
In this part, we used spatial distribution of EVs across locations using map, first we used unique() function to find a list of all countries in the dataset, then we used FIPS codes and used choropleth map to visualize it



**7.Model Popularity**

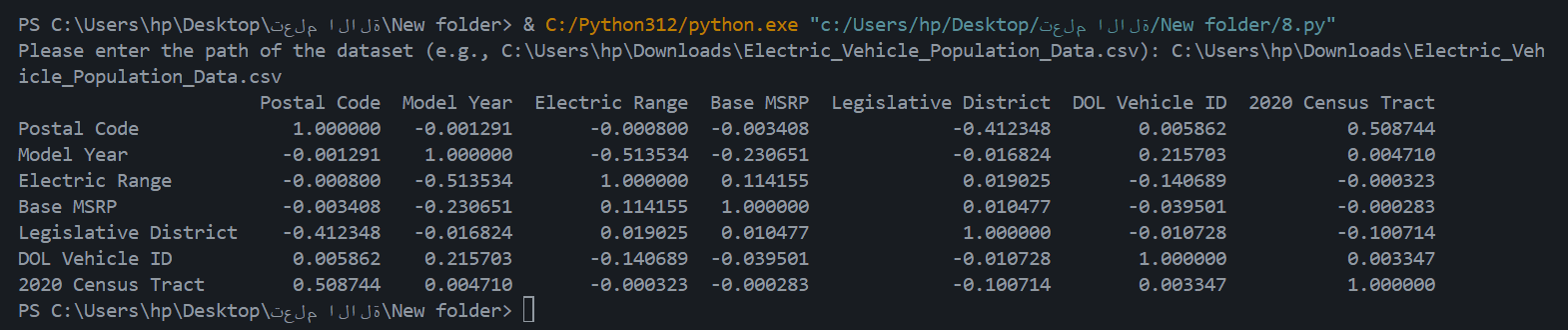
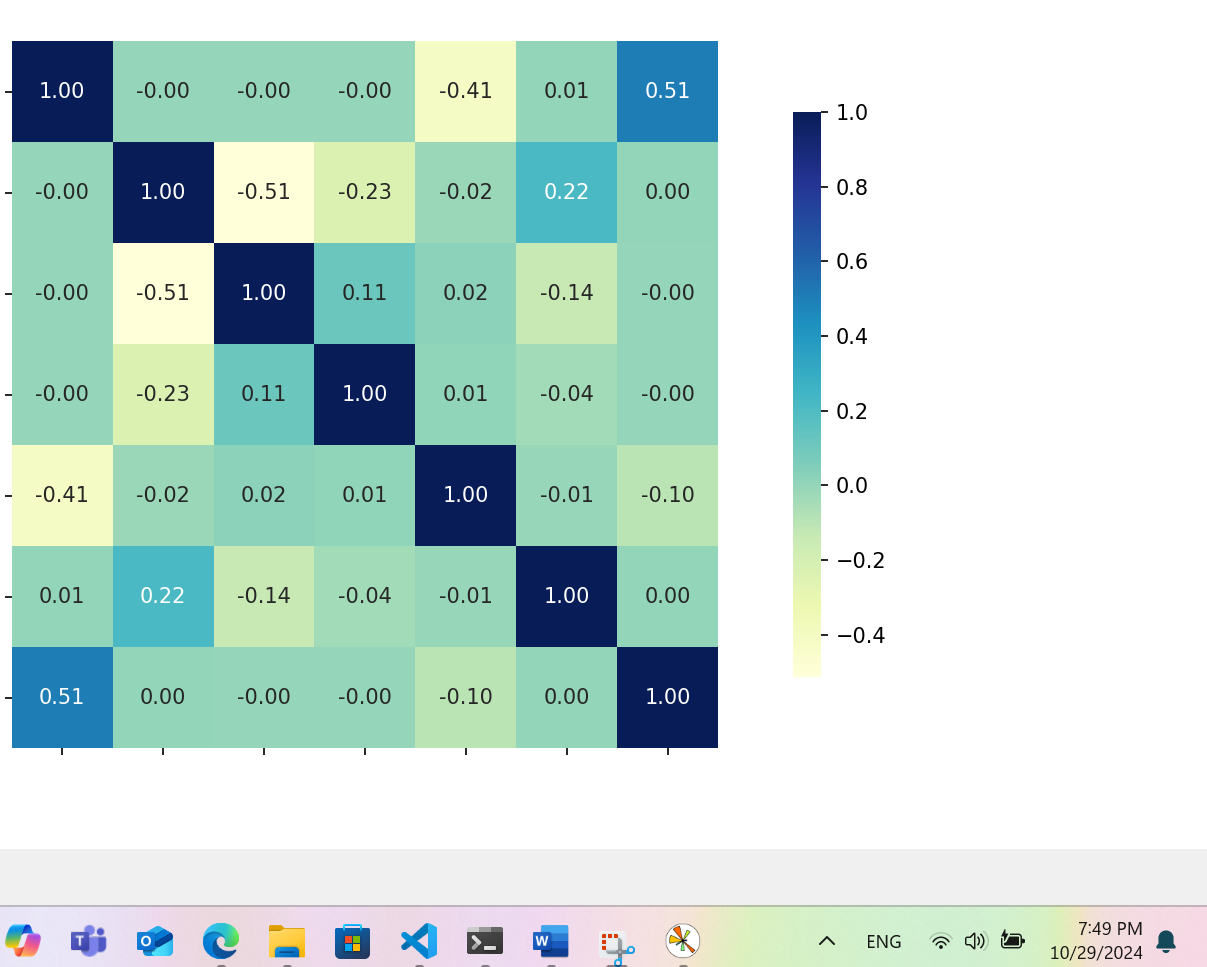
The code loads and cleans an EV dataset, then creates visualizations to explore model popularity, trends over time, county-specific popularity, and type distribution, highlighting top models and trends in EV adoption. Key outputs include bar and line charts displaying top models, geographic distributions, 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.





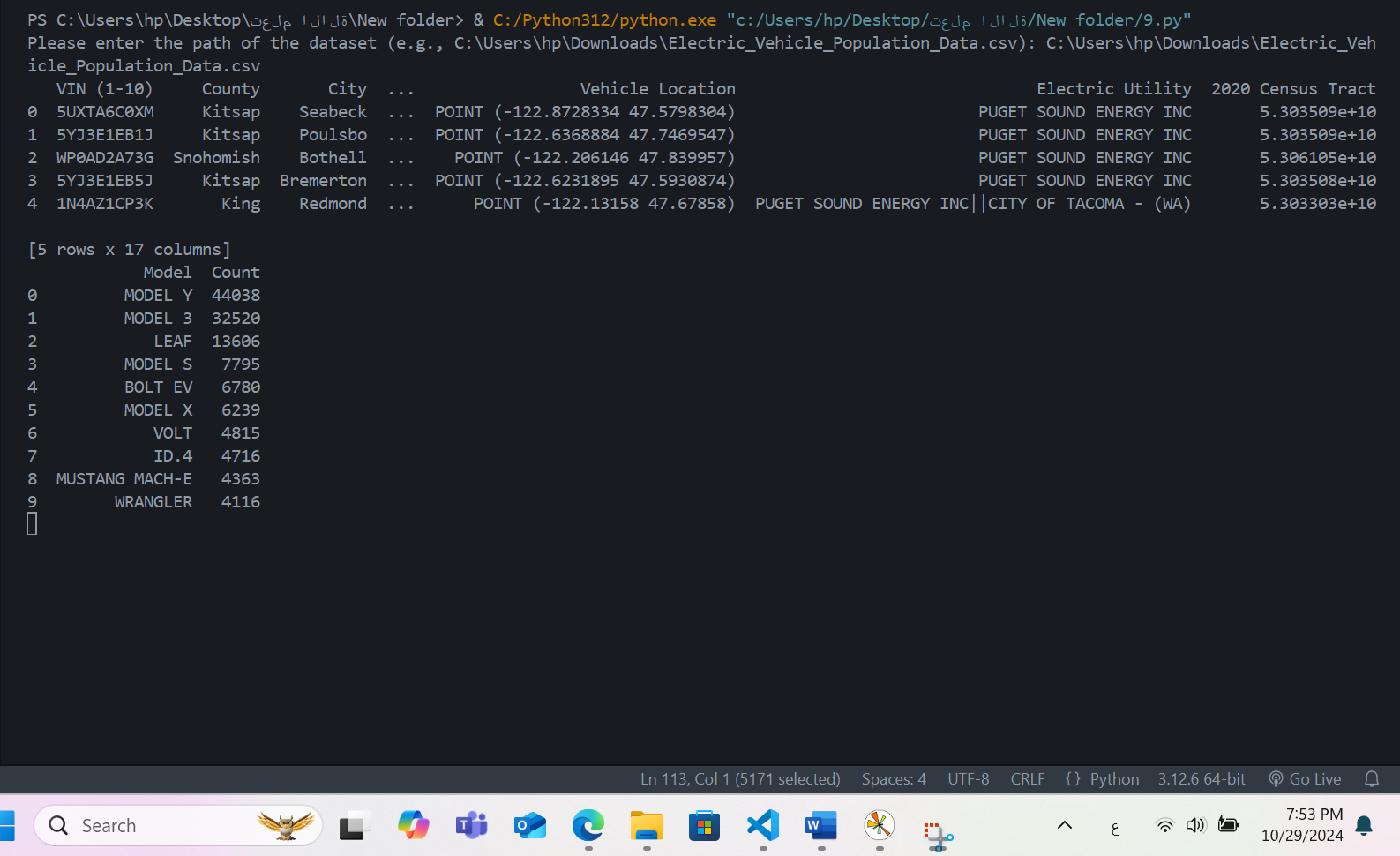
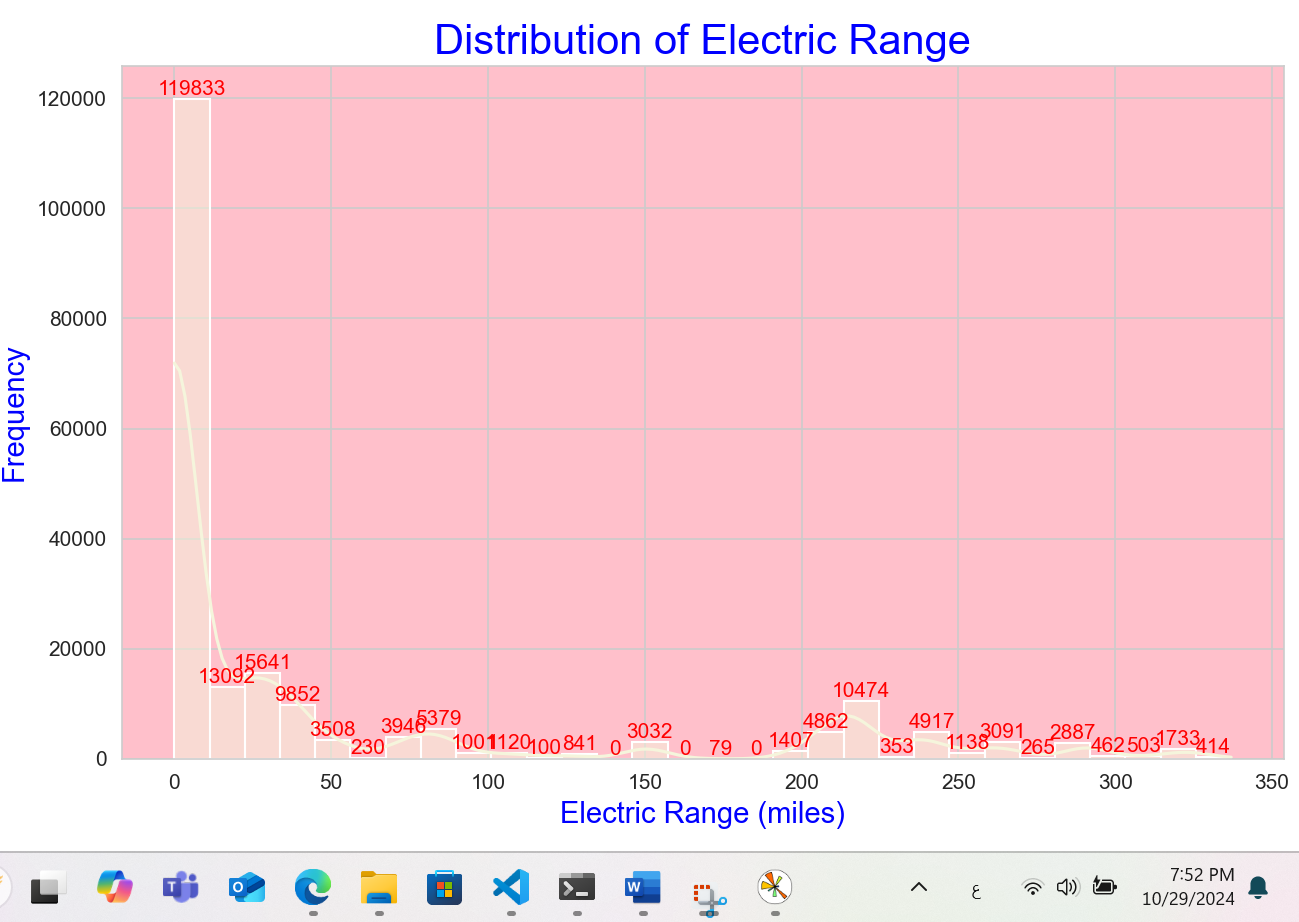
**8.correlation analysis**

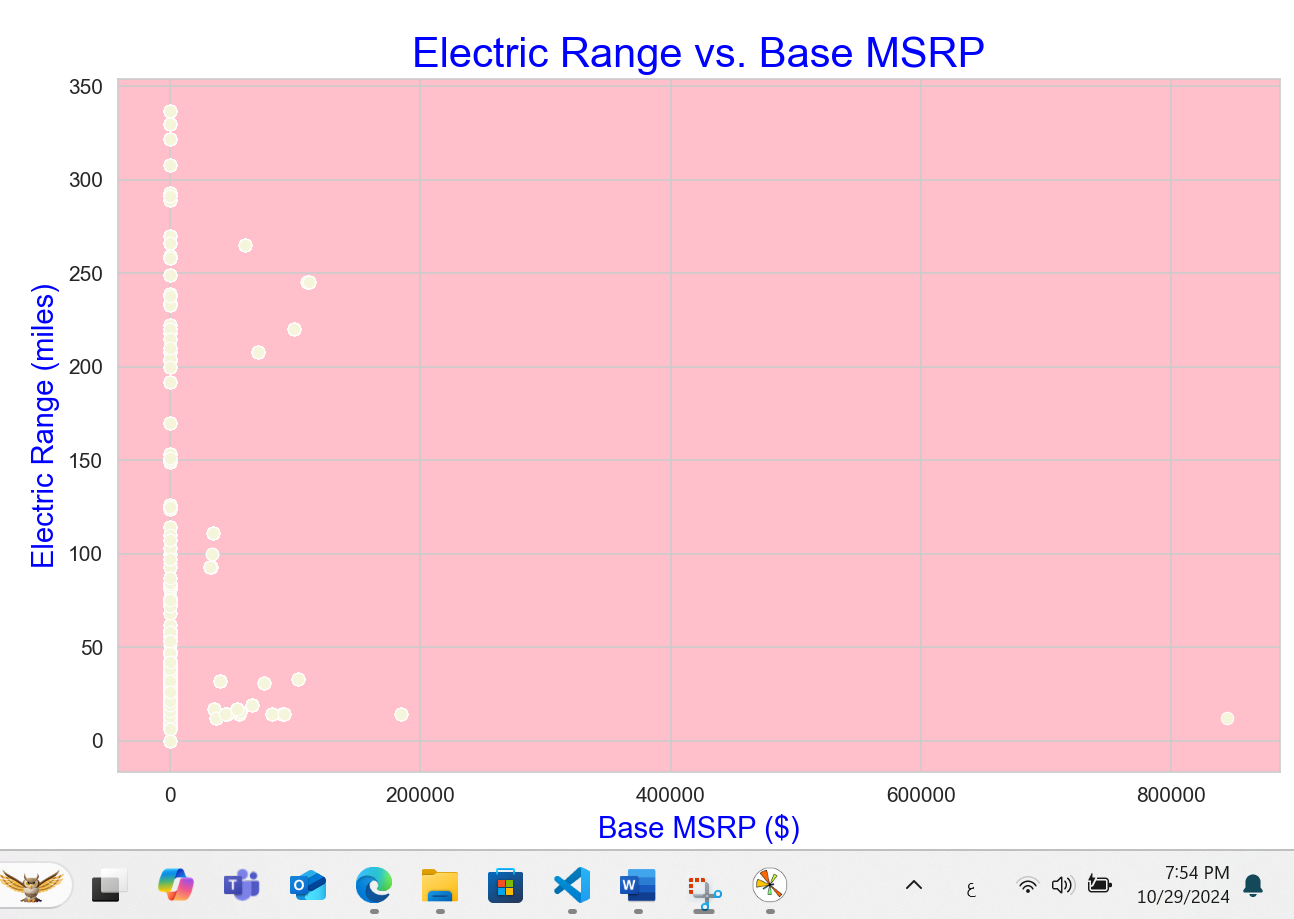
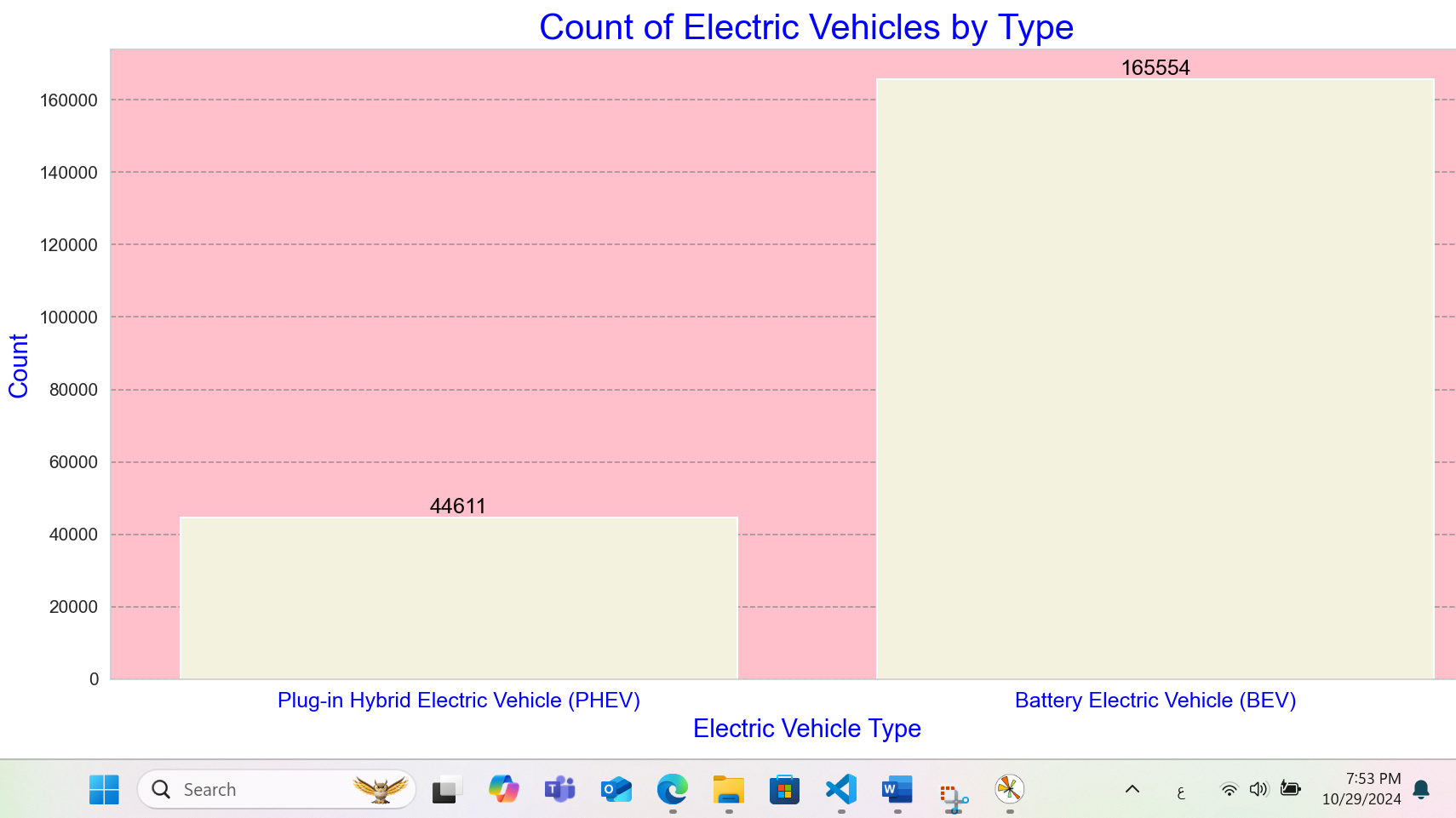
This code loads and cleans an EV dataset, then calculates and visualizes a correlation matrix for numeric features. The heatmap shows the strength and direction of relationships, with some features displaying strong positive or negative correlations, indicating possible dependencies where higher values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

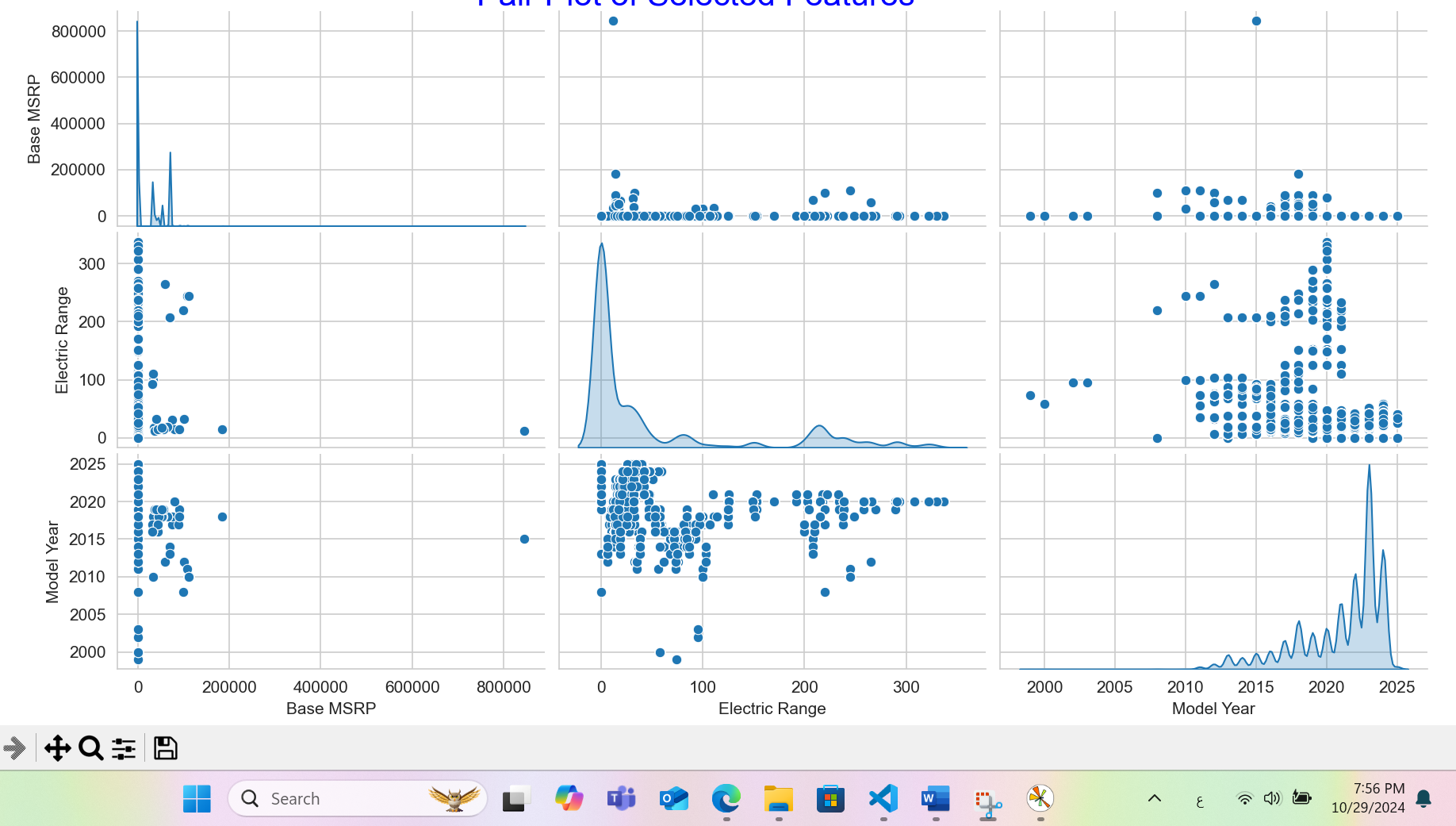
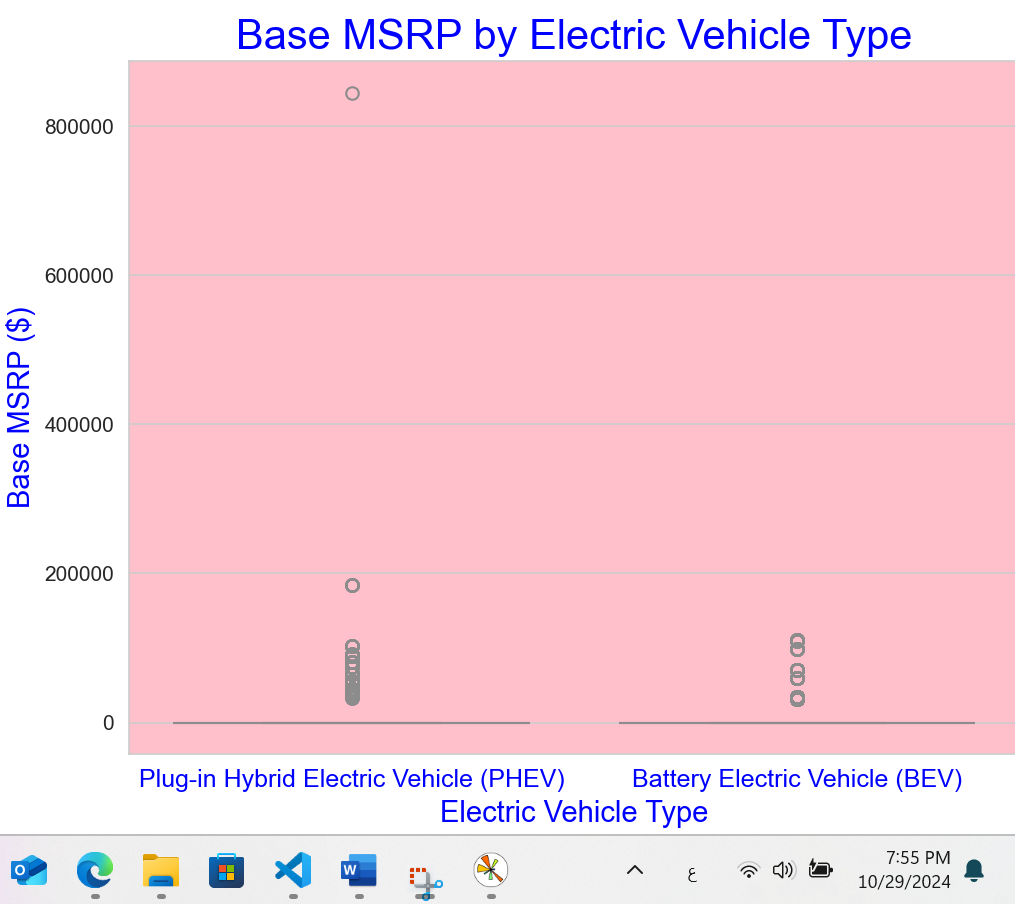


**9.Data Exploration Visualizations**

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.

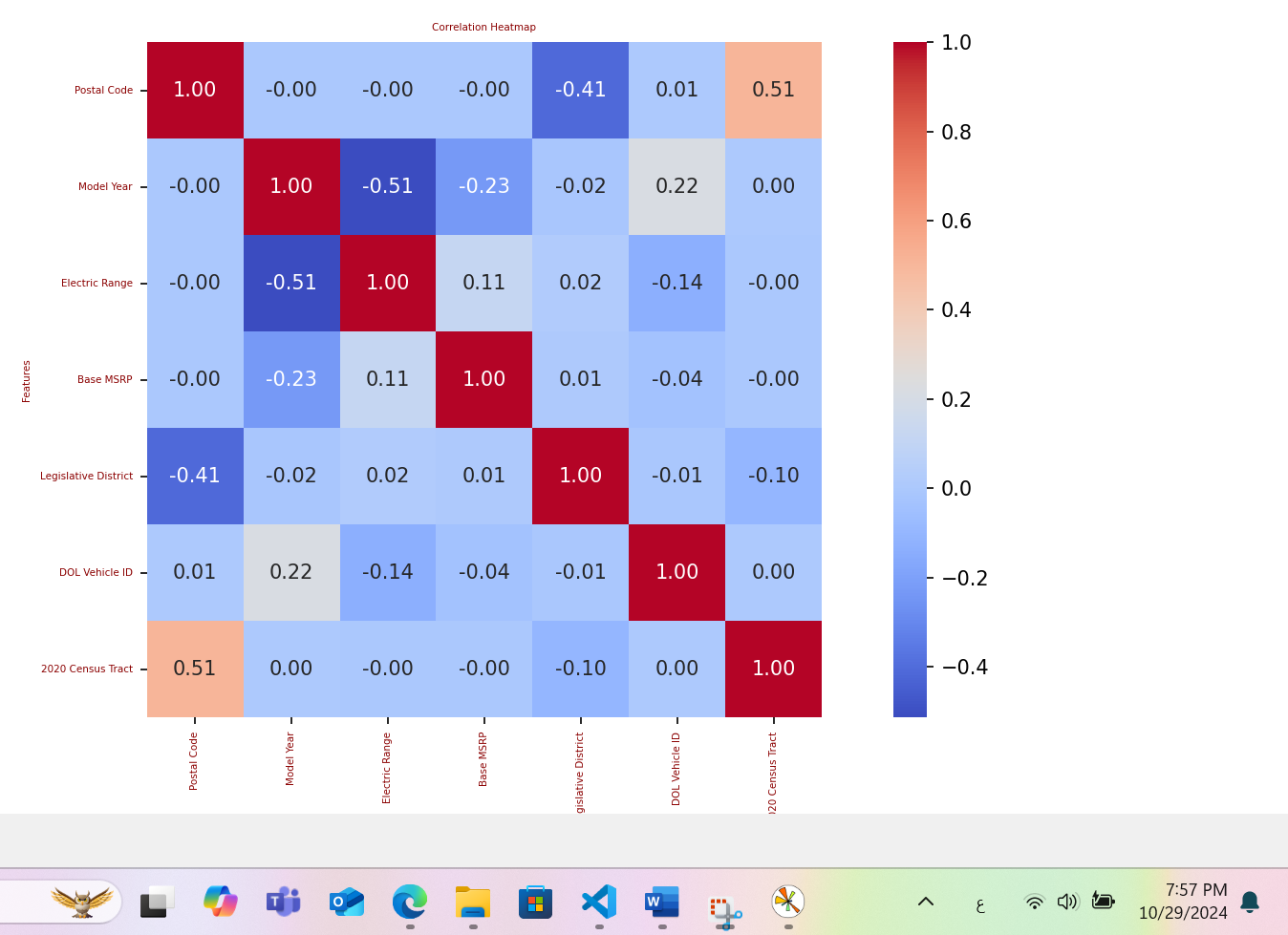
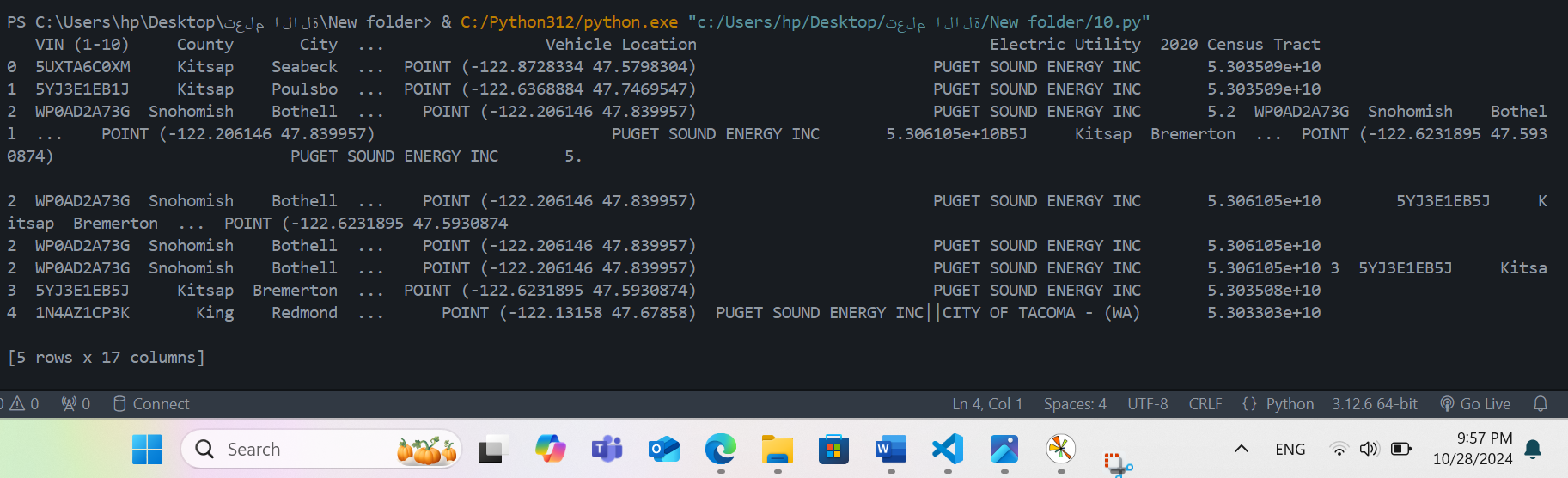






**10.Comparative Visualization**

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.

