



**Faculty of Engineering & Technology Electrical &
Computer Engineering Department**

ENCS5341

Machine Learning

Assignment 1

Prepared by:

Alhasan Manasra 1211705

Mohammad Khdour 1212517

Instructor: Dr. Ismail Khater

Section: 2 & 3

Date: October 30, 2024

Abstract

This assignment analyzes a real-world data set titled "Electric Vehicle Population Data". The data set contains various features related to electric vehicles (EVs), including model type, electric range, and registration details across different cities and counties. The primary objectives of the assignment are to preprocess the data, perform exploratory data analysis (EDA), and visualize key insights. We identify and handle missing values, apply feature encoding techniques such as one-hot encoding, and normalize numerical features for further analysis. Descriptive statistics and correlation analyses were conducted to explore relationships between the features. The spatial distribution of EVs was visualized across different locations, and the popularity of various EV models was analyzed. Additionally, multiple visualizations such as histograms, scatter plots, boxplots, and bar charts were generated to provide deeper insights into the data.

Dataset and Attributes Description

According to the dataset in EV, there are 210166 samples and 17 features distributed in the following form:

- **VIN (1-10):** The first ten characters of the Vehicle Identification Number (VIN), a unique identifier for each vehicle.
- **County:** The county in which the vehicle is registered.
- **City:** The city of the vehicle's registration.
- **State:** The U.S. state where the EV is registered is useful for analyzing state-level adoption patterns and potential policy impacts.
- **Postal Code:** The postal code of the vehicle's registration.
- **Model Year:** The manufacturing year of the EV, which can be used to study trends in EV adoption over time and assess the popularity of different model years.
- **Make:** The manufacturer or brand of the vehicle, indicating the brand distribution in the EV market (e.g., Tesla, Nissan, Chevrolet).
- **Model:** The specific model of the vehicle (e.g., Tesla Model 3, Nissan Leaf) helps to determine model popularity and user preferences.
- **Electric Vehicle Type:** The type of EV, which may include categories like Battery Electric Vehicle (BEV) and Plug-in Hybrid Electric Vehicle (PHEV), is useful for comparing adoption across different EV types.
- **Clean Alternative Fuel Vehicle (CAFV) Eligibility:** Indicates whether the vehicle qualifies as a Clean Alternative Fuel Vehicle, which may make it eligible for certain incentives or benefits, aiding in policy and eligibility analysis.
- **Electric Range:** The maximum range the vehicle can travel on electric power alone, in miles. It's important to understand the capabilities of EVs on the market and user preferences.
- **Base MSRP:** The manufacturer's suggested retail price at the base model level is useful for analyzing EV costs and market trends by price point.
- **Legislative District:** The legislative district where the vehicle is registered, which can assist in understanding EV distribution in specific political regions and help inform local policy decisions.
- **DOL Vehicle ID:** A unique identifier assigned by the Department of Licensing (DOL), which tracks vehicle registration within the state's records.
- **Vehicle Location:** The specific location data (potentially latitude and longitude) of the registered vehicle, enabling precise geographic analysis if spatial coordinates are included.

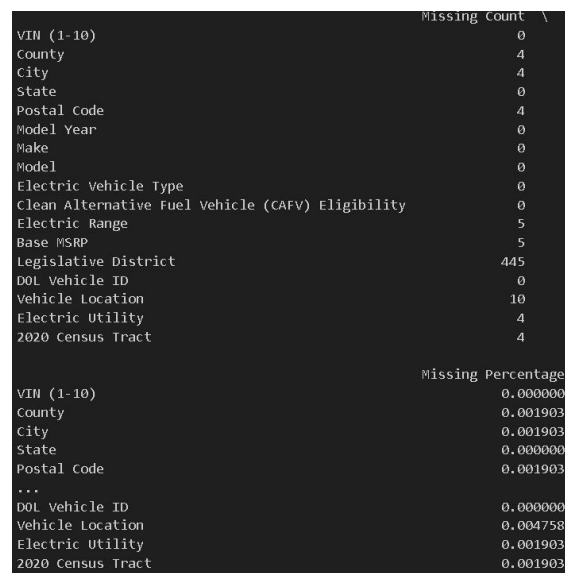
- **Electric Utility:** The electric utility company servicing the area where the vehicle is registered, is relevant for assessing EV infrastructure support and energy resource allocation.
- **2020 Census Tract:** The Census Tract as defined in the 2020 U.S. Census, provides detailed socioeconomic and demographic data for the region, useful for socio-economic analysis of EV adoption.

First, we should read the CSV file to complete our work on this data.

Data Cleaning and Feature Engineering

- **Document Missing Values:**

In this task, we analyzed the dataset to identify any missing values across the different features. Missing values are common in real-world datasets and can affect the quality of the analysis if not properly handled. We systematically searched for missing values in the dataset by examining each column for missing or null entries.



	Missing Count
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 (CAEV) 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

	Missing Percentage
VIN (1-10)	0.000000
County	0.001903
City	0.001903
State	0.000000
Postal Code	0.001903
...	
DOL Vehicle ID	0.000000
Vehicle Location	0.004758
Electric Utility	0.001903
2020 Census Tract	0.001903

Figure 1 - Task 1

- **Missing Value Strategies:**

We applied multiple strategies to handle missing values and compared their impact on the analysis. Dealing with missing values is crucial to ensuring the integrity of the dataset, and we employed methods like Median, Mean, Mode, and Drop.

After applying these strategies, we compared their impact by observing the differences in key summary statistics, model performance, and the distribution of the data. This comparison helped us choose the most appropriate method for handling missing values in the final dataset, ensuring minimal bias and maintaining data integrity for subsequent analyses.

We noticed that all missing values appeared after filling it with strategies.

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

Figure 2 - Task 2

● Feature Encoding:

We focused on transforming categorical features into a numerical format suitable for analysis and machine learning algorithms. Specifically, we encoded the categorical features such as Make and Model, which represent the manufacturer and model of the electric vehicles, using one-hot encoding.

One-hot encoding is a technique that converts categorical variables into a series of binary columns, where each column represents one unique category. The presence of a category in a given row is indicated by a 1, while all other categories are marked with 0. This approach ensures that the categorical data can be properly used by machine learning models without introducing any ordinal relationships.

Make_ROLLS-ROYCE	Make_SMART	Make_SUBARU	Make_TESLA	Make_THINK	Make_TOYOTA	Make_VINFAST	Make_VOLKSWAGEN	Make_VOLVO	Make_WHEEGO ELECTRIC CARS
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 3 - task 3

● Normalization:

We normalized the numerical features in the dataset to ensure they are on a similar scale, which is crucial for many analysis methods and machine learning algorithms. Normalization is especially useful when the numerical features have different ranges or units, as this can cause features with larger ranges to dominate the results, leading to biased model outcomes. In this case, we use min-max normalization, to normalize the numerical feature

Electric Range	Base MSRP	Legislative District
0.089021	0.0	0.708333
0.637982	0.0	0.458333
0.044510	0.0	0.000000
0.637982	0.0	0.458333
0.445104	0.0	0.916667

Figure 4 - task 4

Exploratory Data Analysis

● Descriptive Statistics:

In this task, we calculated summary statistics for the key numerical features in the dataset: Electric Range, Legislative District, and Base MSRP. These statistics provide insights into the central tendency and variability of the data, which helps in understanding the overall distribution of each feature.

Electric Range mean = 0.1501560210983959	Legislative District mean = 0.5818799526751172	Base MSRP mean = 0.0010618824736152055
Electric Range median = 0.0	Legislative District median = 0.6458333333333334	Base MSRP median = 0.0
Electric Range std = 0.2580809626199416	Legislative District std = 0.31026792338992387	Base MSRP std = 0.009056232377523865

Figure 5 - Task 5

● Spatial Distribution:

We visualized the spatial distribution of electric vehicles (EVs) across different locations to gain insights into how EVs are geographically distributed across cities and states. By mapping the location data, we aimed to identify regions with high concentrations of EV registrations and potential geographic trends in EV adoption.

Spatial Distribution of Electric Vehicles

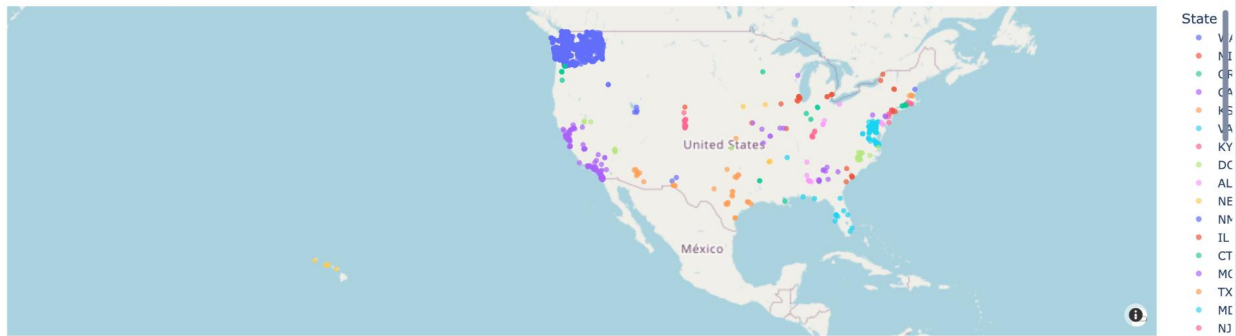


Figure 6 - task 6

For each point and position for cars and for each state, we give it a specific color to distinguish it from others.

● Model Popularity:

We analyzed the popularity of different electric vehicle (EV) models by examining the number of registered vehicles for each model in the dataset. The goal was to identify the most popular EV models and detect any trends in consumer preferences.

From the model counts, we selected the top 10 most popular EV models based on the number of registered vehicles. These top models represent the most widely adopted EVs in the dataset.

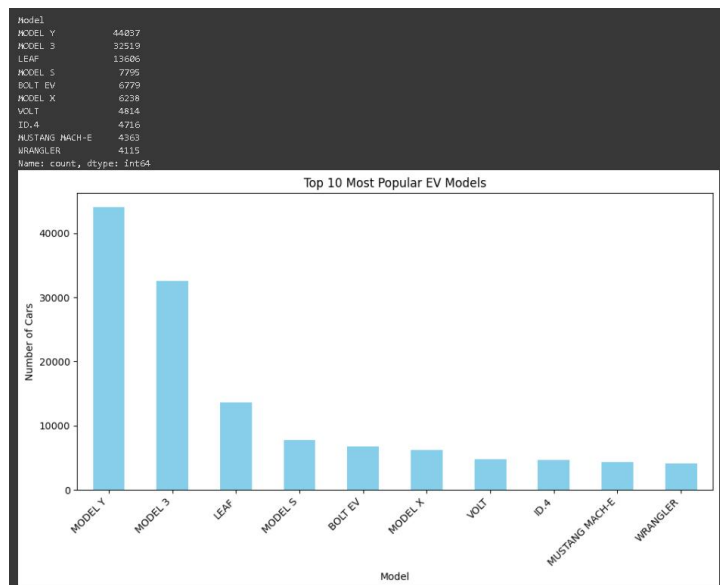


Figure 7 - task 7

The bar chart illustrates the Top 10 Most Popular EV Models based on vehicle registrations in the dataset. The Tesla Model Y and Tesla Model 3 dominate the chart with approximately 44,000 and 32,500

registrations, respectively, showcasing Tesla's significant market presence. The Nissan Leaf ranks third with 13,600 vehicles, followed by other Tesla models (Model S and Model X) and non-Tesla models like the Chevy Bolt EV, Volkswagen ID.4, and Ford Mustang Mach-E. While Tesla leads the market with four models in the top 10, the chart also highlights increasing competition from other automakers. The stark difference in registration numbers between the top two Tesla models and the rest reflects Tesla's strong influence in the EV market.

● Investigate the correlation

we investigated the relationships between all pairs of numerical features in the dataset to identify any significant correlations. Correlation analysis is important for understanding how different numerical variables are related to one another, which can inform feature selection and help us understand underlying patterns in the data.

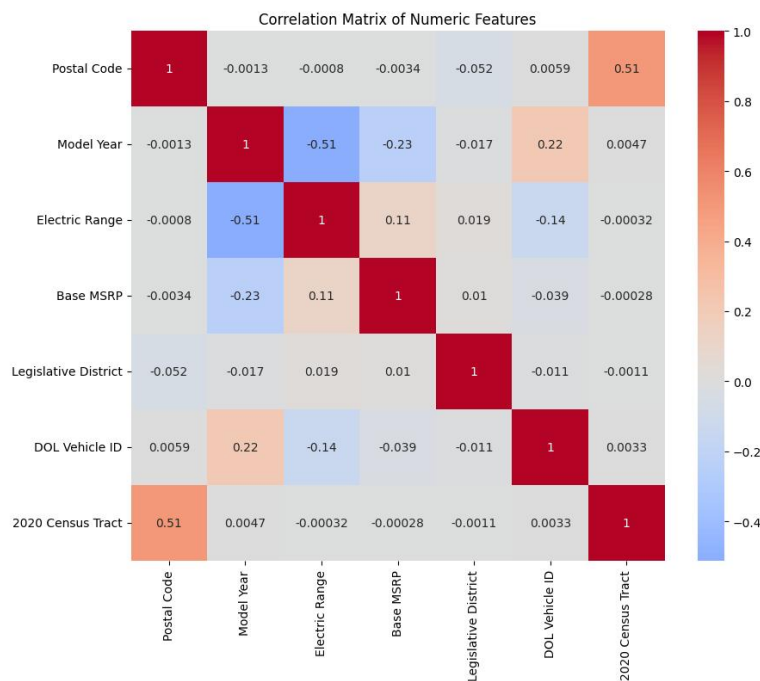


Figure 8 - task 8

The heatmap above illustrates the correlation matrix of the numerical features in the dataset. It shows the pairwise relationships between features such as Postal Code, Electric Range, Base MSRP, Legislative District, DOL Vehicle ID, and 2020 Census Tract. The correlation values range from -1 to 1, where red indicates a positive correlation (strongest relation), and blue indicates a negative correlation. A strong positive correlation (0.51) is observed between Postal Code and 2020 Census Tract, suggesting that these geographic identifiers are closely related. There is a Strong negative correlation (-0.51) between Model Year and Electric range, indicating that cars with a higher electric range tend to have a slightly higher MSRP. Other feature pairs exhibit weak or near-zero correlations, such as the slight negative correlation (-0.00028) between Base MSRP and 2020 Census Tract, implying that most features in the dataset are largely independent of one another.

Visualization

● Data Exploration Visualizations:

We created multiple visualizations to explore the relationships between different features in the dataset. These visualizations help us gain insights into the distribution and correlation between various variables, allowing for a better understanding of the data.

The figure below is hisplot Distribution of the Legislative District

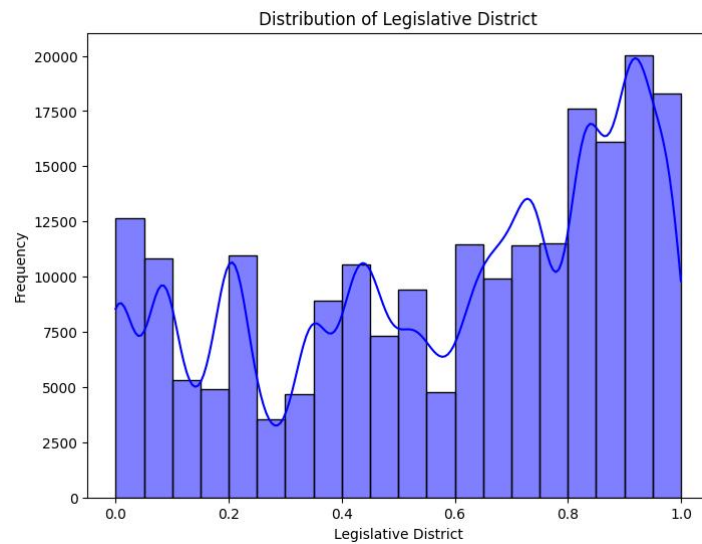


Figure 9 - task 9 part I

The figure below is Scatterplot for Electric Range vs. Model Year

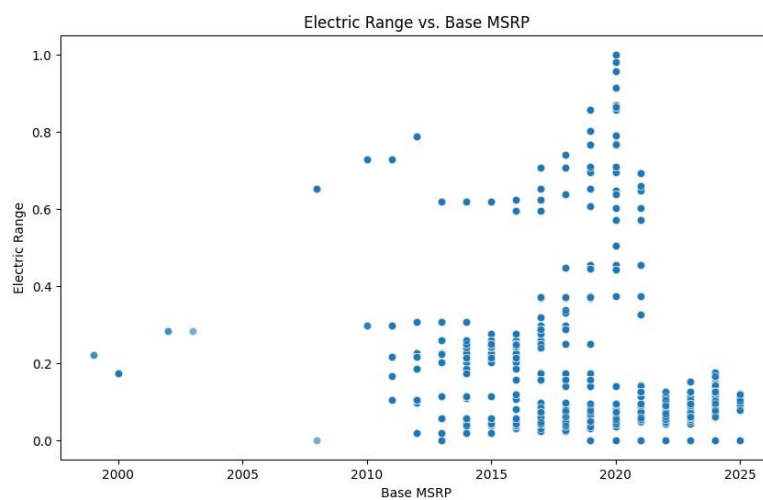


Figure 10 - task 9 part II

The figure above shows that there is no correlation between these two features.

● Comparative Visualization:

We created a bar chart to compare the distribution of electric vehicles (EVs) across the top 10 cities. The x-axis represents the cities, and the y-axis shows the number of registered EVs in each city. Using the head (10) function, we selected the top 10 cities with the highest number of EV registrations. The chart highlights that Seattle has the highest concentration of EVs, with more than 30,000 registrations, followed by Bellevue and Vancouver. The remaining cities have significantly fewer registered EVs, indicating that EV adoption is concentrated in a few urban areas. This visualization helps to easily identify the cities with the highest and lowest EV adoption rates among the top 10.

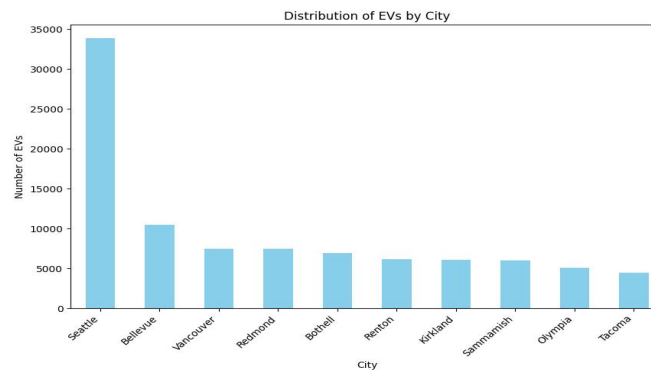


Figure 11: Task 10 bar chart

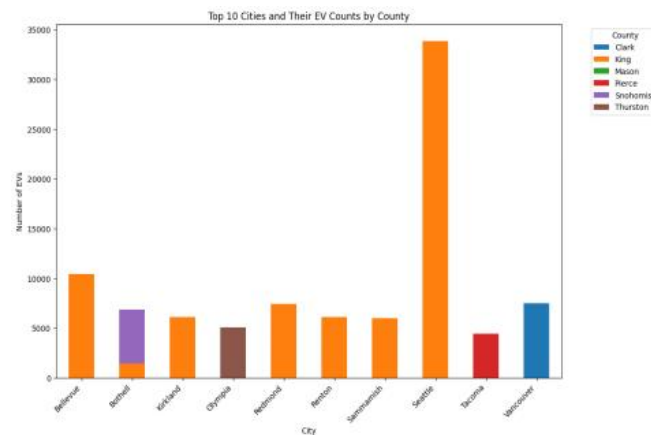


Figure 12: Task 10 stack bar chart

● Temporal Analysis:

We analyzed the temporal trends in electric vehicle (EV) adoption and the popularity of different EV models over time.

We calculated the overall trend of EV adoption by grouping the data by Model Year and plotting the number of registered EVs per year. The line plot shows a significant increase in EV registrations starting from around 2015, with a sharp rise around 2020, peaking in 2022. This trend indicates the growing adoption of EVs in recent years, with a noticeable surge as electric vehicles became more widely available and popular. The spike in 2022 suggests a substantial increase in EV registrations, likely driven by advancements in technology, incentives, and growing environmental concerns.

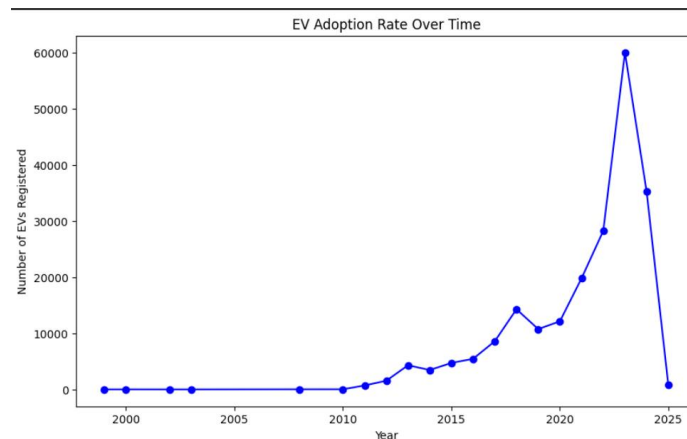


Figure 13 - task 11 part I

We also analyzed the popularity trends of the top 3 most popular EV models—Model Y, Model 3, and LEAF—over time. We filtered the dataset to include only these models and plotted their registrations by year.

