

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

**Machine Learning and Data Science ‐ ENCS5341**

**Assignment 1**

**Data Analysis Project**

**Prepared by:** QOSSAY RIDA 1211553

MAI BEITNOBA 1210260

**Instructor:** Ismail Khater

**Date**: 30/10/2024

# Introduction:

In this report, an analysis of the "Electric Vehicle Population Data" is conducted to apply data science techniques, with a focus on data cleaning, feature engineering, and exploratory data analysis (EDA). The dataset is prepared by addressing missing values and encoding categorical data to improve quality and usability. Patterns and trends in electric vehicle registrations across Washington State are explored through descriptive statistics and visualizations, allowing insights to be effectively communicated. This project is designed to demonstrate a structured approach to handling real-world data while uncovering meaningful trends in electric vehicle adoption.

# Data Cleaning and Feature Engineering:

## Document Missing Values

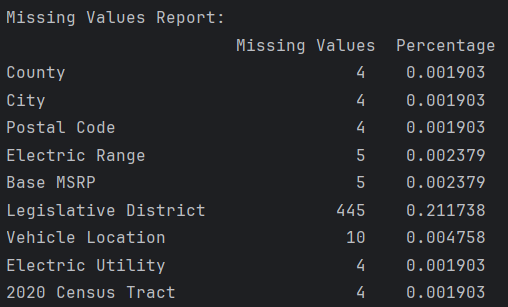
A missing values analysis was done to understand the frequency and distribution of null entries across all features in the dataset. The "Legislative District" column had the highest number of missing values, with 445 entries missing, representing 21.17% of the total dataset. Other features, including "County," "City," "Postal Code," "Electric Range," and "Electric Utility," had a very low percentage of missing data, all below 0.2%. This initial assessment provides a clear view of which columns will require attention during data cleaning.

Figure :Missing values Report

## Missing Value Strategies

To handle missing values, several approaches were tested, including mean imputation for numerical columns and most common value imputation for categorical columns, as well as median imputation with forward and backward fills for categorical data. These methods allowed the dataset size to be preserved while addressing missing values in different ways. Ultimately, the approach of dropping rows with missing values was selected, reducing the dataset from 210,165 to 209,709 entries. This ensured that no filled-in values were introduced, resulting in a cleaner dataset for analysis.

## Feature Encoding

One-hot encoding was applied to the “Electric Vehicle Type” column, which initially contained two categories: Plug-in Hybrid Electric Vehicle (PHEV) and Battery Electric Vehicle (BEV). With this transformation, each category was represented by a binary indicator (0 or 1) rather than text labels.

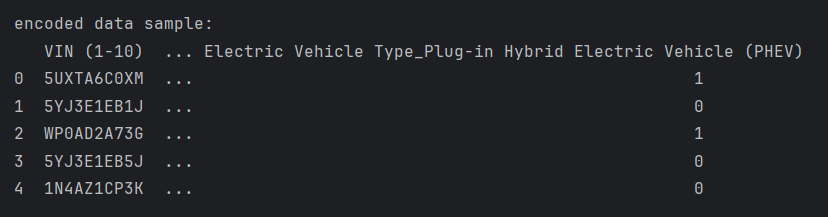


Figure 2:Sample of encoded Data

As shown in Figure 2, the new encoded data consists of zeros and ones in the respective column, making the categorical data compatible with analytical methods that require numerical input. Then, the encoded data was saved in a CSV file for easy access and further analysis.

## Normalization

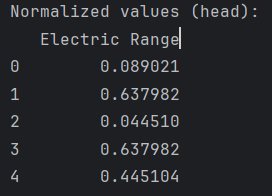
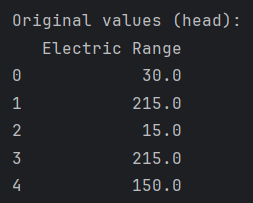
Normalization was applied to the “Electric Range” feature to adjust its values to a common scale, typically between 0 and 1. This scaling helps ensure that differences in scale won’t impact analysis results. Here, the Min-Max normalization method was used, which adjusts values based on the minimum and maximum in the column.

Figure :Electric Range before normalization

Figure :Electric Range after normalization

Originally, "Electric Range" values had a wide spread, from 15 to 215 miles in the sample shown. After normalization, these values were compressed between 0 and 1, maintaining their relative distances but making them easier to compare. For example, an initial value of 30 miles became 0.089, while 215 miles became 0.638.

# Exploratory Data Analysis:

## Descriptive Statistics

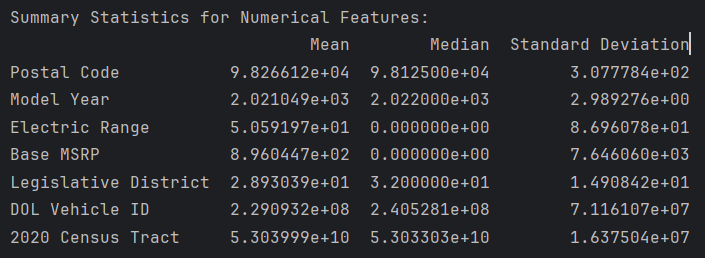
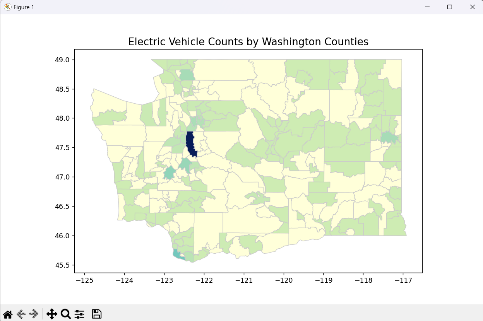


Figure 5 :Descriptive Statistics

Descriptive statistics were calculated for the numerical features to understand their averages and spread. For each feature, the mean, median, and standard deviation were computed, giving insight into the general distribution of values. For example, "Model Year" has a mean of 2021 and a low standard deviation, indicating that most vehicles are from recent years.

In contrast, "Electric Range" and "Base MSRP" show greater variability. "Electric Range" has a mean of around 50.6 miles, but a standard deviation of 86.96, suggesting a broad range of values. "Base MSRP" also has a high standard deviation (7,646) compared to its median of zero, possibly due to widely varied values.

## Spatial Distribution

To explore the spatial distribution of electric vehicles in Washington State, two mapping approaches were used:

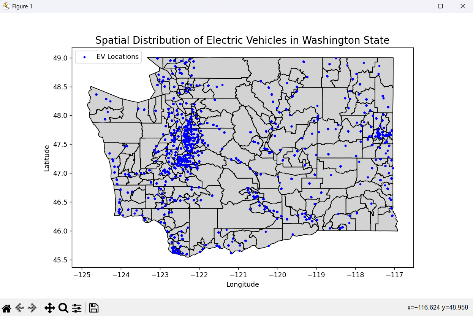
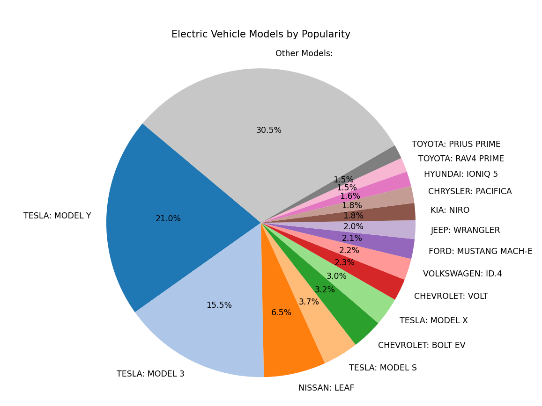
**Mapping EV Counts by City as shown in figure 6**:  
A map was created to display the number of EVs in each city across Washington’s counties. A county boundary shapefile was loaded and projected to the correct geographic system (EPSG:4326). EV counts for each city were calculated from the dataset and merged with the county map. The result was a color-coded gradient map, where yellow represented the lowest EV counts, moving through shades of green to navy blue for the highest concentrations.

Figure :Mapping EV Counts by City

**Mapping Individual EV Locations** **as shown in figure 7**:  
To show precise EV locations, longitude and latitude coordinates from the dataset were used to plot each vehicle on a base map of Washington counties. This scatter plot highlights specific clusters of EVs across the state.

Figure :Mapping Individual EV Locations

## Model Popularity

The analysis of electric vehicle (EV) model popularity reveals clear preferences among consumers, with Tesla models leading in popularity. Tesla’s Model Y, Model 3, and Model S dominate the top rankings. Specifically, the Tesla Model Y is the most popular, followed closely by the Model 3. The Nissan Leaf also appears as a popular choice.

Beyond the top models, other EVs are grouped into "Other Models," accounting for around 30.5% of the total, which indicates a wide variety of EVs with smaller market shares.

Figure :Model Popularity

## Correlation Analysis

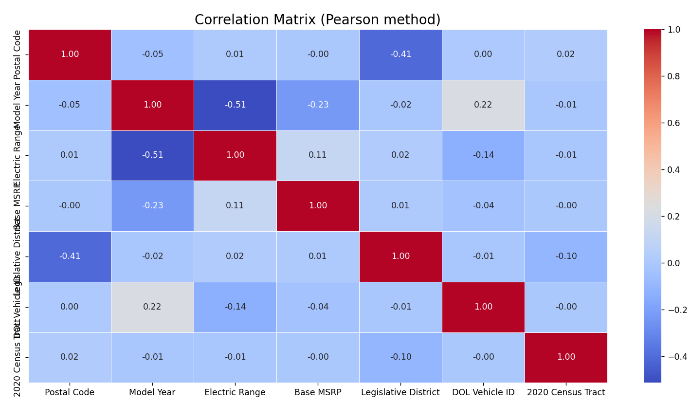
The correlation analysis investigates relationships between numeric features in the dataset by calculating the correlation coefficients for each pair of variable. The generated heatmap visually represents these correlations, with colors indicating the strength and direction of relationships: red for positive correlations and blue for negative. For instance, variables with values close to 1 are strongly positively correlated, meaning they tend to increase together, while those closer to -1 are inversely correlated, where one increases as the other decreases.

Figure :Correlation Matrix

This specific analysis found no correlations above the defined threshold of ±0.8, suggesting that there are no strongly linear relationships between any pairs of numeric features in this dataset. The result indicates that the numeric features vary independently of one another, without significant linear influence.

# Visualization

## Data Exploration Visualizations