BIRZEIT UNIVERSITY

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

Table of Contents:

Table of figures

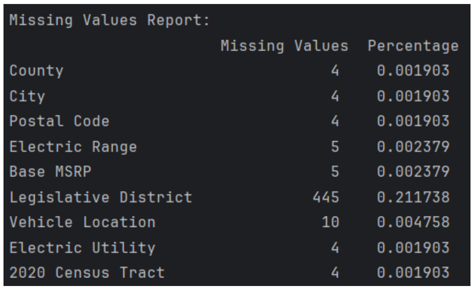
Introduction:

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.



```
Missing Values Report:
                      Missing Values  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
```

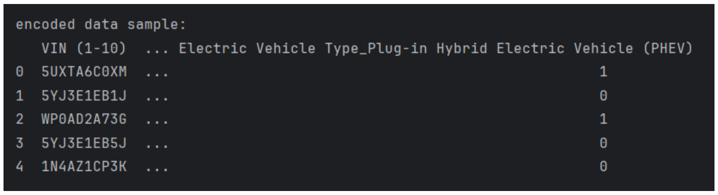
Figure 1:Missing values Report

➢ 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.



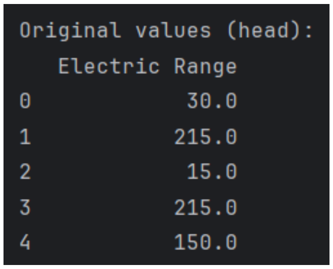
```
encoded data sample:
   VIN (1-10)  ... Electric Vehicle Type_Plug-in Hybrid Electric Vehicle (PHEV)
0  5UXTA6C0XM  ...                                                  1
1  5YJ3E1EB1J  ...                                                  0
2  WP0AD2A73G  ...                                                  1
3  5YJ3E1EB5J  ...                                                  0
4  1N4AZ1CP3K  ...                                                  0
```

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.

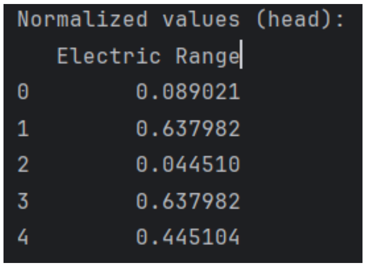
➢ Normalization

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.



```
Original values (head):
   Electric Range
0            30.0
1           215.0
2            15.0
3           215.0
4           150.0
```

Figure 3 :Electric Range before normalization



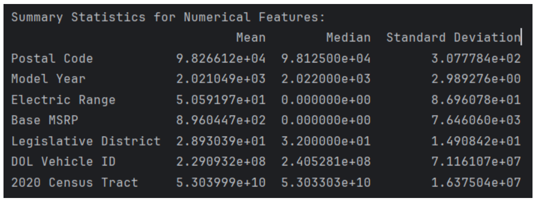
```
Normalized values (head):
   Electric Range
0        0.089021
1        0.637982
2        0.044510
3        0.637982
4        0.445104
```

Figure 4: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



```
Summary Statistics for Numerical Features:
                             Mean        Median  Standard Deviation
Postal Code          9.826612e+04  9.812500e+04        3.077784e+02
Model Year           2.021049e+03  2.022000e+03        2.989276e+00
Electric Range       5.059197e+01  0.000000e+00        8.696078e+01
Base MSRP            8.960447e+02  0.000000e+00        7.646060e+03
Legislative District 2.893039e+01  3.200000e+01        1.490842e+01
DOL Vehicle ID       2.290932e+08  2.405281e+08        7.116107e+07
2020 Census Tract    5.303999e+10  5.303303e+10        1.637504e+07
```

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.

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.

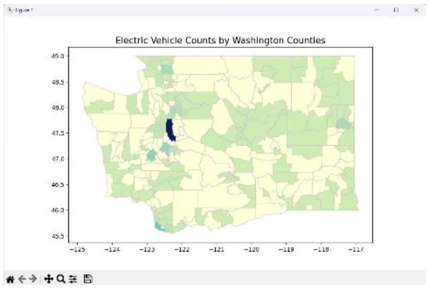


Figure 6:Mapping EV Counts by City

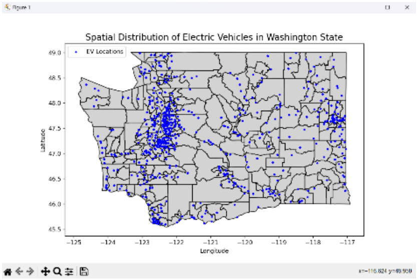


Figure 7:Mapping Individual EV Locations

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.

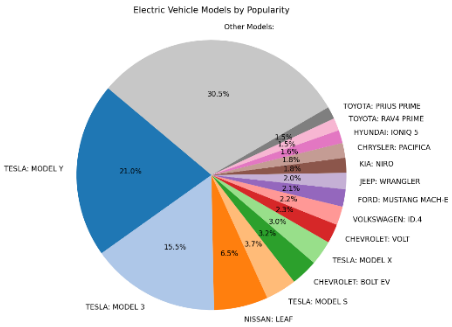


Figure 8:Model Popularity

➢ Correlation Analysis

The correlation analysis investigates relationships between numeric features in the dataset by calculating the correlation coefficients for each pair of variables. 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.

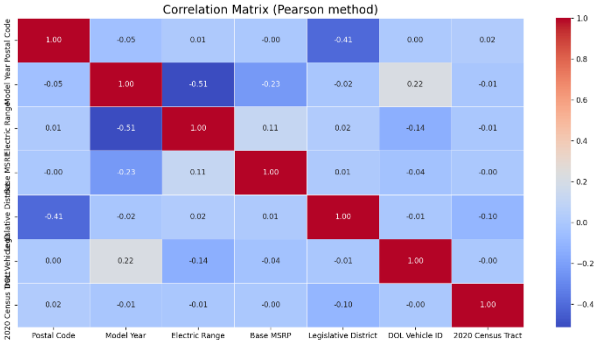


Figure 9:Correlation Matrix

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

In this section, data was visualized using multiple techniques, including histograms, scatter plots, and boxplots.

• Histograms

two histograms created, **Electric Range** and **Model Year,** offer essential insights into the distribution of these features.

The **Electric Range** histogram shows that most vehicles in the dataset have a very low range, with a significant drop in frequency as the range increases. This suggests that short-range electric vehicles are far more common.

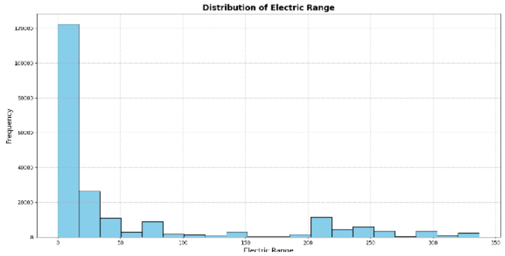


Figure 10:Distribution of Electric Range

The **Model Year Histogram** reveals that most electric vehicles in the dataset are from recent years, particularly from 2010 onward. This trend suggests an increase in electric vehicle production and availability in recent years, with a peak around 2020-2023.

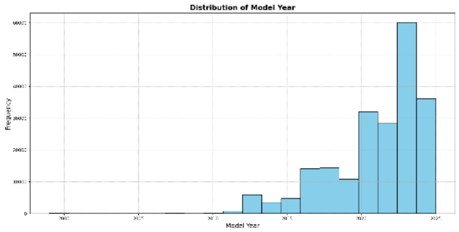


Figure 11:Distribution of Model year

- ## Scatter Plots

The following scatter plots explore key relationships between model year, electric range, and price, offering insights into trends within the dataset.

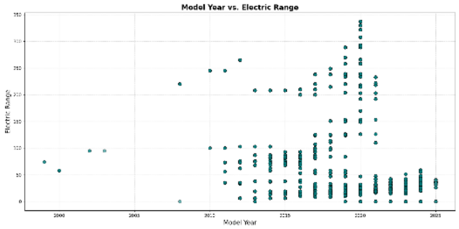


Figure 12:Model Year vs. Electric Range Scatter

This scatter plot shows that the electric range of vehicles has generally increased over the years. Early models mostly had ranges below 100 miles, but starting around 2015, a noticeable improvement is seen, with many models reaching over 200 miles.

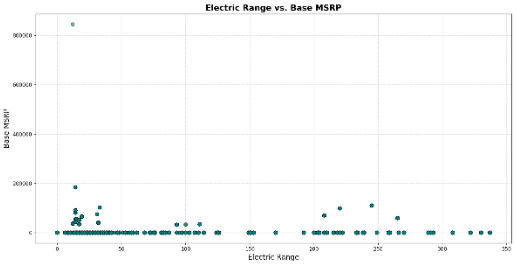


Figure 13: Electric Range vs. Base MSRP Scatter Plot

This scatter plot of Electric Range vs. Base MSRP shows that most electric vehicles (EVs) with a wide range of electric ranges are clustered in the lower MSRP range. However, there are a few outliers with extremely high MSRP values and lower electric ranges. Overall, it suggests that electric range does not necessarily corelate directly with the MSRP, as even lower-priced models can achieve respectable ranges.

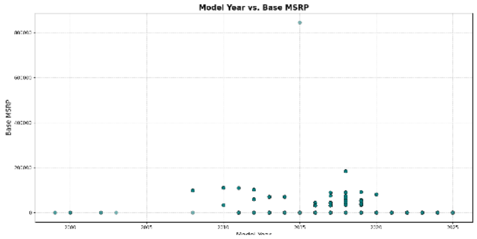


Figure 14:Model year vs. Base MSRP Scatter Plot

This scatter plot demonstrates the relationship between electric vehicle model year and base MSRP It reveals that while most vehicles are priced on the lower end, a few outliers have considerably higher prices, with one model surpassing 800,000 USD. The majority of vehicles are concentrated in the more recent model years, particularly from 2015 onward, reflecting a trend of more frequent electric vehicle releases in recent years.

- ## Boxplots

The boxplots below provide a comparison of electric vehicle ranges across different categories, highlighting variations influenced by factors like legislative district and vehicle make.

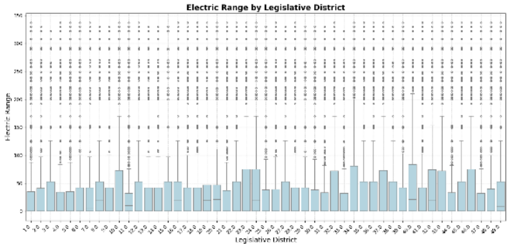


Figure 15:Electric Range by Legislative District

The boxplot in figure 15 shows the distribution of electric vehicle ranges across various legislative districts, highlighting differences in EV range within each district. The spread of values and any outliers help identify districts where EVs typically have higher or lower ranges.

This boxplot displays the distribution of electric vehicle ranges across different car makes. It highlights the variability in electric range within each brand, with some manufacturers showing a broad range of values, indicating diversity in their EV models. Outliers are also visible, suggesting that certain models within these brands achieve exceptionally high or low ranges compared to the typical offerings.

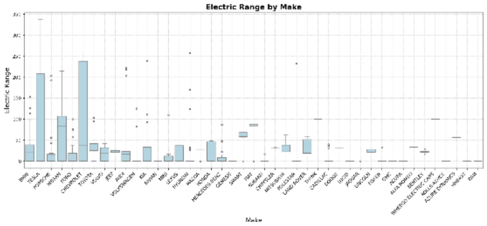


Figure 16:Electric Range by Make

➢ Comparative Visualization

This section visualizes the distribution of electric vehicles across top locations, providing insight into the concentration and variety of EV makes in these areas.

The bar chart shown in figure 17 illustrate the distribution of various electric vehicle (EV) makes across the top 10 cities by EV count. Seattle stands out with the highest number of EVs, exceeding 30,000, while other cities like Bellevue and Redmond also have significant but smaller counts. The chart illustrates the diversity of EV brands in each city, with Tesla being prominently represented across most cities. Other brands like Nissan, Chevrolet, and BMW also contribute to the EV population in these areas, but in smaller proportions.

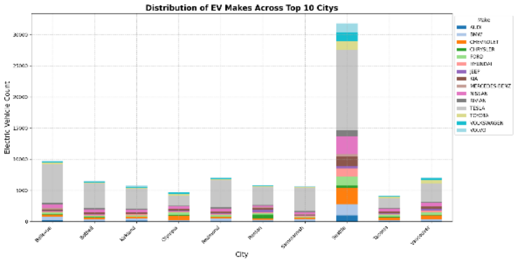


Figure 17:Distribution over EV Makes Across Top 10 Cities

Temporal Analysis

This section examines trends in EV adoption rates and model popularity over time. By analyzing changes across different time points.

The line graph shows the trends in electric vehicle adoption over time, with a notable increase starting around 2015. Adoption rates rise gradually until around 2020, after which there is a sharp spike, peaking close to 60,000 EVs in a single year before a slight decline. This trend indicates a growing interest and investment in electric vehicles in recent years.

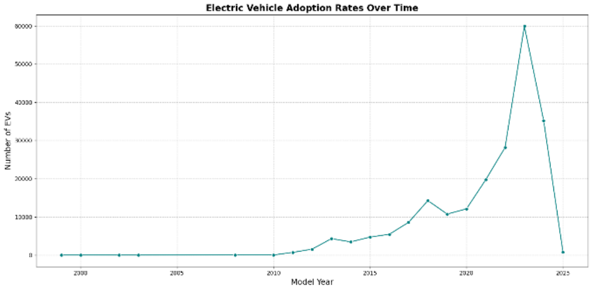


Figure 18: EV Adoption Rates Over Time

This bar chart shows the rise in popularity of the top 5 electric vehicle models over time. Tesla leads the growth, with a sharp increase starting around 2018 and peaking in 2023 at over 25,000 vehicles. Other brands like Ford, Nissan, Chevrolet, and Kia also show steady growth, with a slight dip in 2024.

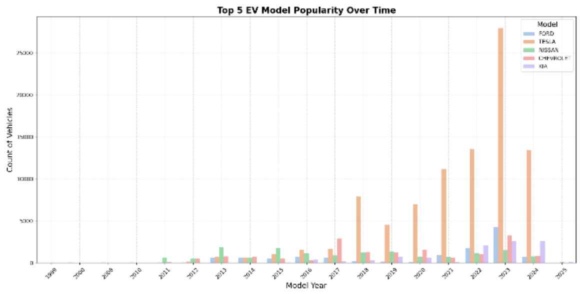


Figure 19:Top 5 EV Model Popularity Over Time

Conclusion

In conclusion, this project provided valuable experience in data manipulation and analysis, focusing on essential techniques like data cleaning, feature engineering, and exploratory data analysis. By working with the "Electric Vehicle Population Data," we practiced handling missing values, encoding features, and normalizing data, which are crucial skills in preparing any dataset for analysis. This assignment successfully reinforced foundational data science skills.