**Washington Electric Vehicle Data Analysis and Modelling**

**Name:** Sravani bodapati

**Date:** 29/09/2025

**Course Details:** Data Analytics

**Introduction**

This report presents an analysis of a dataset containing information on Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) registered in Washington State. The data is provided by the Washington State Department of Licensing (DOL). The objective of this analysis is to clean the data, explore key trends in electric vehicle (EV) adoption, visualize these trends, and build a linear regression model to predict a vehicle's electric range based on its features.

The dataset includes various columns such as vehicle details (VIN, Make, Model, Model Year), registration location (County, City, Postal Code), and technical specifications (Electric Vehicle Type, Electric Range, Base MSRP).

**Section 1: Data Cleaning:**

Proper data preparation is crucial for accurate analysis. The following steps were taken to clean and prepare the dataset.

**Handling Missing Values:**

* The first step was to identify the number of missing values in each column.
* For the 'Base MSRP' and 'Electric Range' columns, missing and zero values were addressed by replacing them with the mean of their respective columns to maintain data integrity for analysis.

**Managing Duplicate Records:**

* To ensure each vehicle is represented only once, duplicate records were identified and removed based on the unique 'DOL Vehicle ID' column.

**Anonymizing VINs:**

* Vehicle Identification Numbers (VINs) were anonymized to protect privacy while maintaining their uniqueness for analytical purposes. This was achieved by applying a hash function to the first 10 characters of the VIN.

**Cleaning GPS Coordinates:**

* The 'Vehicle Location' column, which contained GPS coordinates in a single string, was parsed into separate 'latitude' and 'longitude' columns for better readability and to facilitate geospatial analysis.

**Section 2: Data Exploration**

This section explores the cleaned dataset to uncover key trends and insights regarding EV adoption in Washington.

**Top EV Makes and Models:**The most common EV makes and models were identified by counting their occurrences. The top 5 makes are TESLA, NISSAN, CHEVROLET, FORD, and KIA, while the top 5 models are MODEL Y, MODEL 3, LEAF, MUSTANG MACH-E, and BOLT EV.

**EV Distribution by County:** Analysis of EV registrations by county reveals that \*King County has the most registered EVs\*, followed by Snohomish, Pierce, Clark, and Thurston. This suggests a higher concentration of EV adoption in certain populous areas.

**EV Adoption Over Time:** By analyzing the distribution of EVs by model year, a clear upward trend in adoption is visible, particularly in recent years.

**Average Electric Range:** The \*average electric range\* for all EVs in the dataset was calculated to provide a benchmark for performance.

**CAFV** **Eligibility**:A significant portion of the vehicles are eligible for Clean Alternative Fuel Vehicle (CAFV) incentives. The analysis calculated the exact percentage of eligible versus non-eligible vehicles.

**Range and Price Analysis**:

* The electric range varies significantly across different makes and models.
* Similarly, the average Base MSRP was calculated for each EV model to understand pricing differences.

**Section 3: Data Visualization**

Visualizations were created to effectively communicate the findings from the data exploration phase.

**Top 5 EV Makes:** A bar chart was created to display the top 5 most popular EV makes by count, providing a clear visual comparison.

**EV Adoption by Model Year:** A line graph illustrates the growth trend of EV registrations over different model years, highlighting the acceleration in adoption.

**Electric Range vs. Base MSRP:**A scatter plot was generated to examine the relationship between a vehicle's price and its electric range. This visualization helps identify pricing and performance trends.

**CAFV Eligibility**:A pie chart shows the proportion of EVs that are eligible for CAFV incentives versus those that are not, offering an at-a-glance view of the distribution.

Geospatial Distribution: A geospatial map was used to plot EV registrations based on their latitude and longitude, visualizing the geographic concentration of EVs across Washington.

**Section 4: Linear Regression Model**

To predict the electric range of a vehicle, a linear regression model was developed.

**Model Objective:**The goal was to predict a vehicle's 'Electric Range' using other features from the dataset.

**Features (Independent Variables):**The primary features selected to predict the electric range were \*'Model Year' and 'Base MSRP'\*. Other variables like 'Make' could also be included.

**Handling Categorical Variables:**To incorporate a categorical variable like 'Make' into the model, \*One-Hot Encoding\* was used. This technique converts the categorical data into a numerical format that the regression algorithm can process.

**Model Performance (R² Score):** The model's accuracy was evaluated using the R² score, which indicates the proportion of the variance in the electric range that is predictable from the independent variables. For instance, an R² score of 0.49 means that approximately 49% of the variability in electric range can be explained by the model's features.

I**nfluence of MSRP:** The model's coefficient for 'Base MSRP' reveals its influence on the electric range. A positive coefficient indicates that, on average, a higher MSRP is associated with a longer electric range.

**Model Improvement:** Several steps can be taken to improve the model's accuracy, including \*feature engineering\* (creating new predictive features), \*handling outliers\*, and checking for multicollinearity among predictor variables.

**Predictive Capability:** This trained model can be used to \*predict the electric range of new EV models\* by inputting their specifications (e.g., a 2024 model year with a $50,000 MSRP).

**Conclusion**

This analysis of the Washington State EV dataset provides valuable insights into the regional adoption trends, popular vehicle models, and key factors influencing EV specifications. The data cleaning process ensured the reliability of the findings, while visualizations effectively highlighted trends such as the dominance of certain models and the concentration of EVs in specific counties. The linear regression model demonstrated a method for predicting electric range, offering a practical tool for forecasting and analysis. Key findings indicate that EV adoption is rapidly growing, with a strong correlation between vehicle price and electric range.

**THE END**