

Dataset Information

This dataset shows the Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) that are currently registered through the Washington State Department of Licensing (DOL).

Column Descriptions:

Column Name	Description
VIN (1-10)	First 10 characters of the Vehicle Identification Number (unique identifier for a vehicle)
County	The county where the vehicle is registered (e.g., King, Kitsap)
City	The city where the vehicle is registered (e.g., Seattle, Poulsbo)
State	The state of registration (WA - Washington)
Postal Code	The ZIP code where the vehicle is registered
Model Year	The year the vehicle was manufactured (e.g., 2019, 2020, 2023)
Make	The brand of the vehicle (e.g., Tesla, Hyundai)
Model	The specific model of the vehicle (e.g., Model 3, Model Y, Ioniq 5)
Electric Vehicle Type	The type of EV (Battery Electric Vehicle - BEV)
Clean Alternative Fuel Vehicle (CAFV) Eligibility	Whether the vehicle qualifies for clean energy incentives
Electric Range	The estimated range of the vehicle in miles on a full charge
Base MSRP	Manufacturer's Suggested Retail Price (MSRP) for the base model (0 means data is missing)
Legislative District	The legislative district where the vehicle is registered
DOL Vehicle ID	Unique ID assigned by the Department of Licensing (Washington)
Vehicle Location	GPS coordinates (longitude, latitude) of the vehicle's registered location
Electric Utility	The energy provider for the area (e.g., Puget Sound Energy, City of Seattle)
2020 Census Tract	The Census tract for demographic and geographic analysis

Assignment Questions

1. Data Cleaning Questions:

- How many missing values exist in the dataset, and in which columns?
- How should missing or zero values in the Base MSRP and Electric Range columns be handled?
- Are there duplicate records in the dataset? If so, how should they be managed?
- How can VINs be anonymized while maintaining uniqueness?
- How can Vehicle Location (GPS coordinates) be cleaned or converted for better readability?

2. Data Exploration Questions:

- What are the top 5 most common EV makes and models in the dataset?
- What is the distribution of EVs by county? Which county has the most registrations?
- How has EV adoption changed over different model years?
- What is the average electric range of EVs in the dataset?
- What percentage of EVs are eligible for Clean Alternative Fuel Vehicle (CAFV) incentives?
- How does the electric range vary across different makes and models?
- What is the average Base MSRP for each EV model?
- Are there any regional trends in EV adoption (e.g., urban vs. rural areas)?

3. Data Visualization Questions:

- Create a bar chart showing the top 5 EV makes and models by count.
 - Use a heatmap or choropleth map to visualize EV distribution by county.
 - Create a line graph showing the trend of EV adoption by model year.
 - Generate a scatter plot comparing electric range vs. base MSRP to see pricing trends.
 - Plot a pie chart showing the proportion of CAFV-eligible vs. non-eligible EVs.
 - Use a geospatial map to display EV registrations based on vehicle location.
4. Linear Regression Model Questions:
- How can we use Linear Regression to predict the Electric Range of a vehicle?
 - What independent variables (features) can be used to predict Electric Range? (e.g., Model Year, Base MSRP, Make)
 - How do we handle categorical variables like Make and Model in regression analysis?
 - What is the R² score of the model, and what does it indicate about prediction accuracy?
 - How does the Base MSRP influence the Electric Range according to the regression model?
 - What steps are needed to improve the accuracy of the Linear Regression model?
 - Can we use this model to predict the range of new EV models based on their specifications?

Assignment Submission Instructions

1. General Guidelines:

- Ensure your responses are **clear, concise, and well-structured**.
- Include **explanations, code snippets, charts, and interpretations** where necessary.
- Use appropriate **data cleaning, exploration, visualization, and regression techniques** to answer the questions.
- Submit your assignment in one of the following formats:
 - **Jupyter Notebook (.ipynb)** – If using Python for analysis.
 - **Word Document (.docx) or PDF (.pdf)** – If writing an analytical report with visualizations.

2. Submission Format:

- **Upload the URL of GitHub/Google Drive only** and ensure open access.
- Your submission should include the following:
 - 📄 **Title Page:** Assignment title, your name, date, and course details.
 - 📄 **Introduction:** A brief summary of the dataset and the objective of the analysis.
 - 📄 **Section 1: Data Cleaning** – Identify and handle missing values, duplicates, and inconsistencies.
 - 📄 **Section 2: Data Exploration** – Answer key questions about trends, distributions, and insights.
 - 📄 **Section 3: Data Visualization** – Provide charts and graphs with interpretations.
 - 📄 **Section 4: Linear Regression Model** – Explain the model, features used, and evaluation metrics.
 - 📄 **Conclusion:** Summarize key findings and insights from your analysis.
 - 📄 **Appendix (if needed):** Any additional information, references, or raw code.

3. How to Upload & Share Your Assignment:

- 🔗 **GitHub:**
1. Create a new repository on **GitHub** (name it `EV-Data-Analysis`).
 2. Upload your assignment files.
 3. Click on **Settings > Manage Access** and ensure the repository is public.
 4. Copy the repository URL and include it in your submission.
- 🔗 **Google Drive:**
1. Upload your files to **Google Drive** in a folder named `EV-Data-Analysis`.
 2. Right-click the folder, select **"Share"**, and set access to **"Anyone with the link can view."**
 3. Copy the shared link and include it in your submission.

4. Submission Deadline & Grading Criteria:

- **Deadline:** [Specify Deadline]
- **Evaluation Criteria:**

Criteria	Weight (%)
Data Cleaning & Preparation	25%
Exploratory Data Analysis	25%
Visualizations & Insights	20%
Regression Model & Interpretation	20%
Clarity & Presentation	10%

