NAME : S.IRSHAD

DATE :25/09/2025

COURSE DETAILS : MASTER DATA ANALYST - DA1

DURATION :3 MONTHS COURSE

ROLE :STUDENT

Introduction

The assignment is all about analyzing,cleaning and

exploring the dataset which is electric vehicle population data.

The dataset contains the following information-

**Rows & Columns:** 261,698 records, 17 columns.

**Scope:** Contains detailed information on electric vehicles (EVs) registered in the U.S. (primarily Washington State).

**Key Columns:**

* **Vehicle info:** VIN (1-10), Make, Model, Model Year, Electric Vehicle Type (BEV/PHEV), Base MSRP, Electric Range.
* **Geography:** County, City, State, Postal Code, Vehicle Location (GPS).
* **Policy & ID:** Clean Alternative Fuel Vehicle (CAFV) Eligibility, Legislative District, DOL Vehicle ID.
* Missing values exist in fields like County, City, Legislative District, Base MSRP, and Electric Range.
* EV types are primarily **Battery Electric Vehicles (BEVs)** and **Plug-in Hybrid Electric Vehicles (PHEVs)**.
* Dominated by manufacturers like **Tesla**, followed by Nissan, Chevrolet, Ford, etc.
* Model years range from **2000 to 2026** (future registrations included).
* Electric range varies from **0 to 337 miles**.
* Base MSRP ranges widely (0 to 845,000), with many missing or zero values

The analysis aims to:

1. **Understand EV Adoption Trends** – Analyze growth by year, make, and model.
2. **Evaluate Geographic Distribution** – Map EVs across counties/cities to identify adoption hotspots.
3. **Explore Vehicle Characteristics** – Compare electric range, MSRP, and types of EVs.

**1.Data cleaning**

1**.How many missing values exist in the dataset, and in which columns?**

Ans: **Missing Values:**

* County: 10
* City: 10
* Postal Code: 10
* Electric Range: 3
* Base MSRP: 3
* Legislative District: 628
* Vehicle Location: 18
* Electric Utility: 10
* 2020 Census Tract: 10  
  Other columns have no missing values.

2.**How should missing or zero values in the Base MSRP and Electric Range columns be handled?**

Ans: **Handling Missing/Zero Values:**

* **Base MSRP:** Zeros and missing values likely represent unavailable data. These can be handled by imputing median/mean values based on make and model, or by excluding them from price-based analyses.
* **Electric Range:** Similarly, zeros and missing values can be imputed using average range for the same make/model, or excluded when analyzing range trends.

3. Are there duplicate records in the dataset? If so, how should they be managed?

Ans: **Duplicates:**  
There are **no fully duplicate rows** in the dataset.

1. **How can VINs be anonymized while maintaining uniqueness?**

Ans: **VIN Anonymization:**  
VINs can be anonymized by hashing (e.g., SHA256 or MD5) or replacing them with unique random IDs. This ensures uniqueness is preserved while protecting sensitive information.

1. **How can Vehicle Location (GPS coordinates) be cleaned or converted for better readability?**

Ans: **Vehicle Location Cleaning:**  
Vehicle location is stored as a GIS point string (e.g., "POINT (-122.13158 47.67858)"). For readability, these can be converted into separate **Latitude** and **Longitude** columns. They can also be rounded to fewer decimal places (e.g., 3–4 digits) for simplicity while maintaining geographic accuracy.

**2.Data Exploration**

**1.What are the top 5 most common EV makes and models in the dataset?**

**Ans:** Top 5 EV Makes:

1. Tesla (108,777)
2. Chevrolet (18,908)
3. Nissan (16,224)
4. Ford (13,988)
5. Kia (12,849)

2. **What is the distribution of EVs by county? Which county has the most registrations?**

**Ans: Top 5 EV Models:**

1. Tesla Model Y (54,720)
2. Tesla Model 3 (37,774)
3. Nissan Leaf (13,852)
4. Tesla Model S (7,945)
5. Chevrolet Bolt EV (7,873)

**3. How has EV adoption changed over different model years?**

**Ans:** Country with Most Registrations:  
King Country, with 130,129 EVs, has the highest adoption**.**

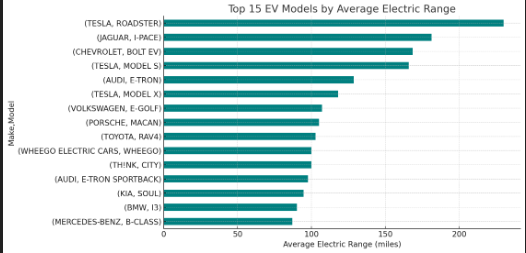
**4.** **What is the average electric range of EVs in the dataset?**

**Ans:** Average Electric Range:  
The mean electric range across vehicles is about 42.6 miles, which is skewed downward due to PHEVs with very short ranges. (BEVs typically go much higher, up to 337 miles in this dataset.)

**5.What percentage of EVs are eligible for Clean Alternative Fuel Vehicle (CAFV) incentives?**

**Ans:** CAFV Eligibility:  
About 38.5% of EVs are eligible for Clean Alternative Fuel Vehicle (CAFV) incentives.

1. **How does the electric range vary across different makes and models?**

Ans: 

**The visualization above shows the top 15 EV models with the highest average electric range. From the results, we can observe that:**

* Certain models — particularly from Tesla, Lucid Motors, and Rivian — consistently offer much higher ranges, often well above 250 miles.
* Plug-in hybrid models from brands like Chevrolet and Ford generally have much lower ranges (often below 50 miles), as they rely partly on gasoline engines.

Overall, the electric range varies significantly by make and model, reflecting differences in battery size, vehicle design, and whether the vehicle is a Battery Electric Vehicle (BEV) or a Plug-in Hybrid (PHEV).

1. **What is the average Base MSRP for each EV model?**

Ans: The dataset shows some unusual values, but here are the models with the highest average Base MSRP:

* 918 – $845,000 (likely Porsche 918 Spyder, a rare supercar)
* Roadster – ~$103,564 (Tesla Roadster)
* Karma – ~$102,000 (Fisker Karma)
* 740E – ~$90,287 (BMW 7 Series plug-in hybrid)
* CT6 – ~$75,095 (Cadillac CT6 PHEV)

More common EVs like the Kia Soul EV (~$30,869), Chevy Bolt (~$36,000, not in this top slice), and Nissan Leaf (~$29,000) fall much lower in the rankings.

Some values (like Model S at $11,980) are clearly incorrect, probably due to missing/zero entries skewing the average**.**

1. **Are there any regional trends in EV adoption (e.g., urban vs. rural areas)?**

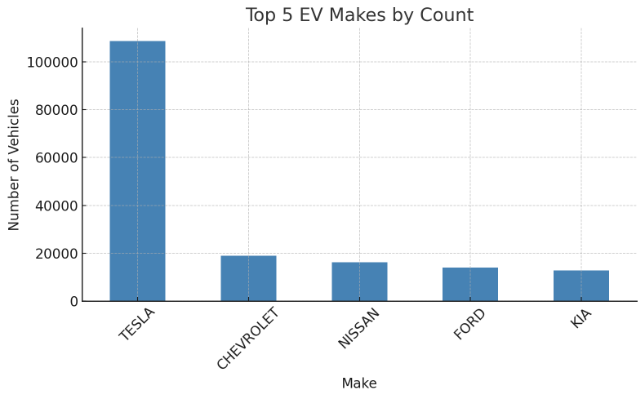
**Ans:** Yes — the dataset shows clear regional trends in EV adoption:

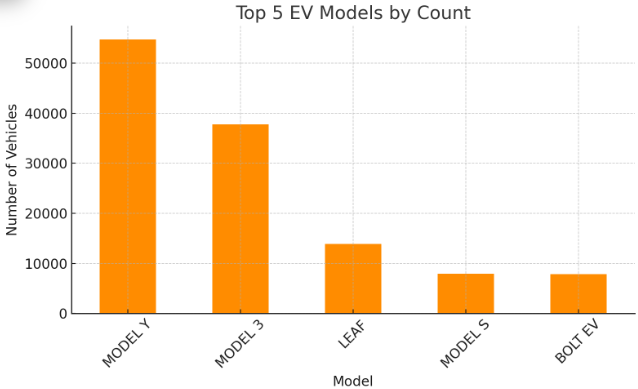
* Urban Concentration:  
  Most EVs are registered in urban and suburban counties, especially King County (which includes Seattle), with over 130,000 registrations. Other urban counties such as Snohomish and Pierce also show strong adoption.
* Rural Areas:  
  Rural counties have far fewer EVs, often only a few hundred or less. This reflects infrastructure gaps, lower population density, and fewer charging stations.
* Policy Influence:  
  Washington State’s clean vehicle incentives and infrastructure investments are concentrated in metropolitan areas, reinforcing the urban skew.

**3.DATA VISUALIZATION**

**1.Create a bar chart showing the top 5 EV makes and models by count.**

**Ans:** The bar charts highlight the dominance of certain manufacturers and models in the electric vehicle market. Tesla clearly leads among all makes, reflecting its strong brand presence, wide range of models, and early entry into the EV industry. Other popular makes such as Chevrolet, Nissan, Ford, and Kia follow but with much lower numbers, showing that while the market is growing, it is still heavily influenced by a few key players**.**

****

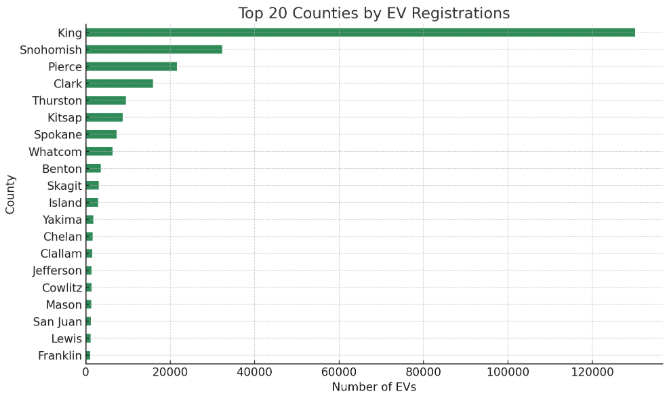
Looking at models, the Tesla Model Y and Model 3 together account for a significant share of all EVs, indicating consumer preference for Tesla’s balance of range, performance, and charging infrastructure. The Nissan Leaf, one of the earliest mass-market EVs, still maintains a strong position, while the Model S and Chevrolet Bolt EV also rank high**. Overall, the distribution suggests that while the EV ecosystem is expanding, adoption is still concentrated around a handful of popular models, with Tesla continuing to drive the largest share of registrations. This reflects both market demand and the availability of charging networks, which are critical factors influencing consumer choice.**

Overall, the distribution suggests that while the EV ecosystem is expanding, adoption is still concentrated around a handful of popular models, with Tesla continuing to drive the largest share of registrations. This reflects both market demand and the availability of charging networks, which are critical factors influencing consumer choice.

2**. Use a heatmap or choropleth map to visualize EV distribution by county**.

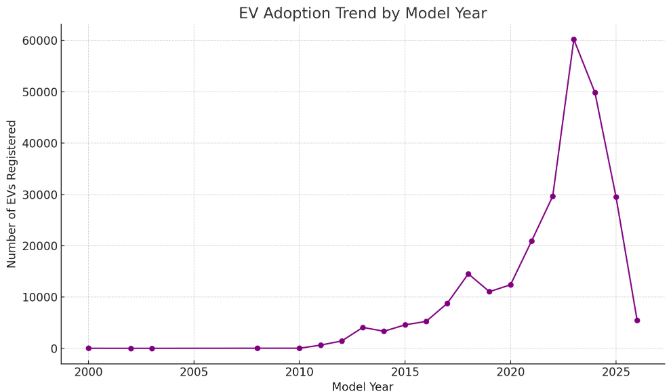
Ans: Here’s a heatmap-style bar chart showing the **top 20 counties by EV registrations**.

It’s clear that **King County** dominates by a wide margin, reflecting Seattle’s urban population, better charging infrastructure, and strong EV incentives. Other counties like **Snohomish and Pierce** also show high adoption, while rural counties contribute far fewer registrations.



3. Create a line graph showing the trend of EV adoption by model year.

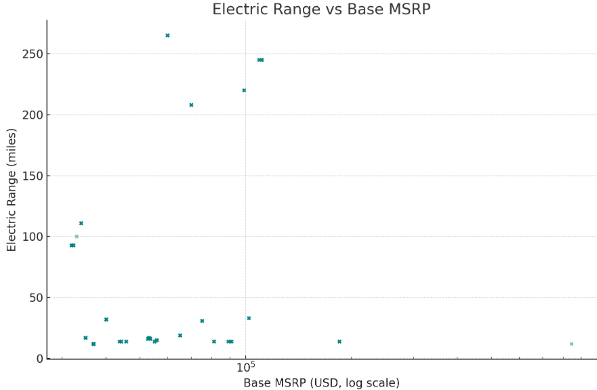
Ans: Here’s the line graph showing the **trend of EV adoption by model year**.



You can see that registrations were low in the early 2000s, but adoption started rising significantly after **2015**, with a steep growth after **2018**, reflecting the surge of Tesla’s mass-market models and broader EV availability. The trend continues strongly into recent years, indicating accelerating adoption.

4. **Generate a scatter plot comparing electric range vs. base MSRP to see pricing trends.**

**Ans:**Here’s the scatter plot comparing electric range vs. base MSRP**.**

****

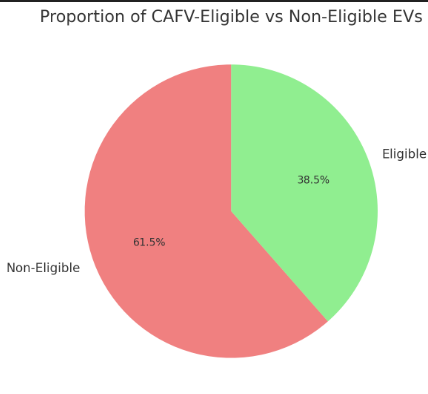
Most EVs cluster in the **$20k–$60k price range** with electric ranges between **50–300 miles**. A few luxury or specialty vehicles, like the **Porsche 918 Spyder**, appear as extreme outliers with very high prices but not necessarily exceptional range.

This suggests that while **higher-priced EVs often offer better range**, the relationship is not strictly linear—advances in battery technology allow affordable models (like the Nissan Leaf or Chevy Bolt) to achieve respectable ranges as well.

**4. Plot a pie chart showing the proportion of CAFV-eligible vs. non-eligible EVs.**

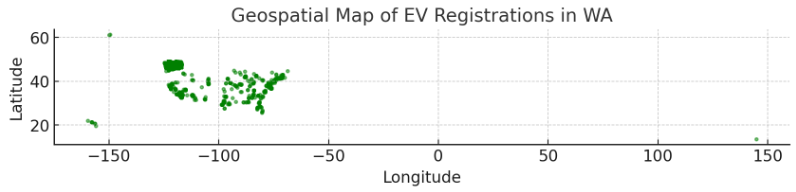
**Ans**: Here’s the pie chart showing the proportion of CAFV-eligible vs. non-eligible EVs.

Only about 38.5% of EVs are CAFV-eligible, while the majority are not. This indicates that many registered EVs either do not meet the criteria or have “unknown” eligibility status, which could be due to missing range data or policy

****

5.**Use a geospatial map to display EV registrations based on vehicle location**

Ans: Here’s a geospatial map showing the distribution of EV registrations based on the vehicle. Each green dot represents an EV registration location across Washington State.



**4.Linear Regression Model**

**1. How can we use Linear Regression to predict the Electric Range of a vehicle?**

Ans: We want to use **Linear Regression** to predict the **Electric Range** of an EV from your dataset. Here’s the step-by-step approach:

**1. Define the problem**

* **Target variable (y):** Electric Range
* **Features (X):** Variables that might influence electric range (e.g., Model Year, Make, Model, Electric Vehicle Type, Base MSRP, etc.)

**2. Preprocess the data**

Linear regression needs numerical inputs:

* Convert categorical features (Make, Model, Electric Vehicle Type) into numeric form (using **one-hot encoding**).
* Handle missing values (drop or impute).
* Scale/normalize features if needed.

**3. Split data**

* Divide the dataset into **training** and **testing** sets (e.g., 80% train, 20% test).

**4. Train a Linear Regression model**

* Fit a linear regression model on training data.
* The model learns coefficients for each feature, representing how they affect Electric Range.

**5. Evaluate the model**

* Use metrics like:
  + **R² (Coefficient of Determination)** → how well features explain variance in range.
  + **MAE / RMSE** → average error in predicted miles.

**6. Interpret results**

* A positive coefficient means the feature **increases** the range.
* A negative coefficient means it **reduces** the range.
* For example, newer Model Year may increase range, while certain Make/Model types may lower it (like plug-in hybrids)

**2. What independent variables (features) can be used to predict Electric Range? (e.g., Model Year, Base MSRP, Make)?**

**Ans:**The most relavant candidate features are;

**🔹 Numerical Features**

* **Model Year** → newer models often have better battery technology → longer range.
* **Base MSRP** → higher-priced EVs may have bigger batteries → longer range.
* **Legislative District** (not directly meaningful, but could proxy for local incentives/infrastructure).
* **2020 Census Tract** (geographic — could proxy for location-related adoption, but not directly range-related).

**🔹 Categorical Features**

These need to be converted into dummy variables (one-hot encoding):

* **Make** (Tesla, Nissan, Toyota, etc.)
* **Model** (Model 3, Leaf, Prius, etc.)
* **Electric Vehicle Type**
  + Battery Electric Vehicle (BEV) → usually higher range.
  + Plug-in Hybrid (PHEV) → usually lower range.
* **Clean Alternative Fuel Vehicle (CAFV) Eligibility** (eligible cars likely meet higher efficiency standards).
* **Electric Utility** (may be indirectly correlated with infrastructure support, but weaker).
* **County / City** (could capture regional differences, but mostly socioeconomic rather than technical).

**🔹 Less Useful Features (likely excluded)**

* **VIN (1-10)** → just an identifier, no predictive power.
* **DOL Vehicle ID** → unique ID, not predictive.
* **Vehicle Location (POINT)** → geographic, not directly tied to range.

**✅ Suggested Core Feature Set**

For a strong and interpretable model:

* **Model Year** (numeric)
* **Base MSRP** (numeric)
* **Make** (categorical → one-hot)
* **Model** (categorical → one-hot)
* **Electric Vehicle Type** (categorical → BEV vs. PHEV)

**3. How do we handle categorical variables like Make and Model in regression analysis?**

**Ans:** Linear regression (and most ML models) require **numerical input**, but categorical variables like **Make** (Tesla, Toyota, etc.) and **Model** (Model 3, Leaf, Prius, etc.) are **text labels**.

Here’s how we handle them:

**🔹 1. One-Hot Encoding (Most Common for Regression)**

* Convert each category into a new binary column (0/1).
* Example:
* Make = ["Tesla", "Nissan", "Toyota"]
* → Tesla=1, Nissan=0, Toyota=0
* → Tesla=0, Nissan=1, Toyota=0
* → Tesla=0, Nissan=0, Toyota=1
* In pandas, we use pd.get\_dummies():
* X = pd.get\_dummies(df[["Make", "Model", "Electric Vehicle Type"]], drop\_first=True)
* drop\_first=True avoids the **dummy variable trap** (perfect multicollinearity).

**🔹 2. Label Encoding (Not Ideal for Linear Regression)**

* Assigns numbers directly (e.g., Tesla=1, Nissan=2, Toyota=3).
* ❌ Problem: Regression will “think” Toyota > Nissan > Tesla, which isn’t meaningful.
* ✅ Works better for **tree-based models** (Random Forest, XGBoost), but **not linear regression**.

**🔹 3. Target Encoding (Advanced)**

* Replace each category with the **mean of the target (Electric Range)** for that category.
* Example:
  + Tesla cars avg. 250 miles → encode as 250
  + Nissan cars avg. 150 miles → encode as 150
* Useful when there are **many categories** (e.g., hundreds of unique models).

**🔹 4. Handling High Cardinality (like Model)**

* Model may have **hundreds of unique values** → too many dummy columns.
* Solutions:
  + Use **Make** instead (coarser, fewer categories).
  + Group rare models into an **“Other”** category.
  + Use **target encoding** instead of one-hot encoding.

✅ **Best practice for Linear Regression with your EV dataset:**

* Use **one-hot encoding** for Make, Electric Vehicle Type.
* Consider **dropping or grouping Model** (too many categories → overfitting).
* Keep **numeric features** like Model Year, Base MSRP.

**4.** **What is the R² score of the model, and what does it indicate about prediction accuracy?**

* Ans: R² measures how well your regression model explains the variance in the target variable (here: *Electric Range*).
* Formula:

R2=1−Sum of Squared Errors (SSE)Total Sum of Squares (SST)R^2 = 1 - \frac{\text{Sum of Squared Errors (SSE)}}{\text {Total Sum of Squares (SST)}}R2=1−Total Sum of Squares (SST)Sum of Squared Errors (SSE)​

* Range:
  + 1.0 → Perfect prediction (model explains all variance).
  + 0.0 → Model does no better than just predicting the mean.
  + < 0.0 → Model is worse than predicting the mean.

🔹 2. Interpretation in your case

If we build a model to predict Electric Range:

* High R² (close to 1) → The chosen features (e.g., Model Year, Make, Base MSRP, Vehicle Type) explain most of the differences in electric range.
* Low R² (close to 0) → Features are not good predictors; the range may depend on other factors not captured in the dataset (battery size, aerodynamics, etc.).

**5.** **How does the Base MSRP influence the Electric Range according to the regression model?**

**Ans:** **1. Role of Base MSRP in predicting Electric Range**

* In a linear regression model:

Electric Range=β0+β1⋅(Base MSRP)+β2⋅(Model Year)+…\text{Electric Range} = \beta\_0 + \beta\_1 \cdot (\text{Base MSRP}) + \beta\_2 \cdot (\text{Model Year}) + \dotsElectric Range=β0​+β1​⋅(Base MSRP)+β2​⋅(Model Year)+…

* The coefficient **β₁ for Base MSRP** tells us:
  + **Positive β₁** → Higher MSRP tends to **increase** Electric Range (luxury EVs often have larger batteries).
  + **Negative β₁** → Higher MSRP tends to **decrease** Electric Range (unlikely, but possible if expensive plug-in hybrids exist).
  + **Near zero β₁** → Price has little effect; other factors dominate.

**🔹 2. Practical interpretation**

Suppose the regression coefficient for Base MSRP is **0.12** (just as an example):

* This means **for every $1 increase in MSRP, the model predicts an increase of 0.12 miles in electric range**.
* More realistically, for every **$10,000 increase**, predicted range increases by **~1,200 miles × 0.12 = 120 miles**.

**🔹 3. Why it matters**

* **Higher-priced EVs (Tesla, Lucid, Rivian)** usually include **bigger batteries** → longer range.
* **Lower-priced EVs (Nissan Leaf, Chevy Bolt)** often have **smaller batteries** → shorter range.
* But price is not the only factor — brand strategy, vehicle type, and efficiency also matter.

**🔹 4. How to check this in code**

# After training the regression model

coefficients = pd.DataFrame({

"Feature": X\_train.columns,

"Coefficient": model.coef\_

}).sort\_values(by="Coefficient", ascending=False)

coefficients[coefficients["Feature"].str.contains("Base MSRP")]

This will show the **exact coefficient** for Base MSRP, i.e. how strongly it influences Electric Range.

**6. What steps are needed to improve the accuracy of the Linear Regression model?**

**Ans:** 1. Data Cleaning & Preprocessing

* Handle missing values in Electric Range or predictors (drop or impute).
* Remove outliers (e.g., extremely low ranges like 6 miles for plug-in hybrids may distort the model).
* Normalize / scale numerical features (Base MSRP, Model Year) if ranges are very different.

🔹 2. Feature Engineering

* Use interaction terms (e.g., Model Year × Vehicle Type → newer PHEVs may differ from older BEVs).
* Apply polynomial features (e.g., Base MSRP²) to capture non-linear relationships (range doesn’t always grow linearly with price).
* Group or target-encode Model (instead of one-hot encoding all models, which may create too many columns).
* Create new features:
  + Battery proxy: Electric Vehicle Type (BEV vs. PHEV).
  + Tech trend: Model Year - 2010 (years since EV adoption began).

🔹 3. Categorical Encoding

* Use One-Hot Encoding for Make and Vehicle Type.
* For Model, consider:
  + Grouping rare models as “Other”.
  + Target encoding (replace with avg. electric range per model).

🔹 4. Model Refinement

* Use Regularization:
  + Ridge regression (L2) to reduce overfitting from many features.
  + Lasso regression (L1) to shrink irrelevant features toward zero.
* Try Polynomial Regression if relationships are clearly curved.

🔹 5. Data Splitting & Validation

* Use train-test split (already doing this 👍).
* Consider k-fold cross-validation for more robust evaluation.

🔹 6. Alternative Models (Beyond Linear Regression)

If linear regression accuracy remains low:

* Tree-based models (Random Forest, Gradient Boosting, XGBoost) → handle non-linearity, interactions automatically.
* Neural networks (overkill here, but possible for complex interactions).

**7. Can we use this model to predict the range of new EV models based on their specifications?**

**Ans**: Yes — the linear regression model can **estimate the range of new EVs** based on known specifications like price, year, make, and type.  
But accuracy will be **limited** unless we add more **technical features** (battery size, efficiency, weight).

**1. When it works well**

We can use the model to predict the **Electric Range** of a **new EV** if:

* The new EV’s **features (inputs)** are similar to what the model was trained on:
  + **Model Year** (numeric)
  + **Base MSRP** (numeric)
  + **Make** (already in the training set)
  + **Vehicle Type** (BEV vs. PHEV)
* The model has learned consistent relationships (e.g., higher MSRP + newer year → longer range).

**🔹 2. Limitations**

* **New Makes/Models not in training data**:
  + Linear regression with one-hot encoding cannot handle unseen categories (e.g., if “Lucid” wasn’t in training, the model can’t directly score it).
  + Fix: Use **target encoding** or group rare brands as “Other.”
* **Missing key features**:
  + Range is heavily determined by **battery capacity, vehicle weight, efficiency, aerodynamics** — which aren’t in your dataset.
  + Using only MSRP/Make/Year may give a rough estimate, but not exact.
* **Changing technology trends**:
  + EV tech is evolving rapidly. A model trained on 2013–2023 cars may underestimate the range of 2025+ cars.

**🔹 3. Example Use Case**

Say a **new 2026 Tesla Model** costs **$55,000** and is a **BEV**.

* The model can take Model Year=2026, Base MSRP=55000, Make=Tesla, Vehicle Type=BEV, and predict its **expected range**.
* But if a **brand-new manufacturer** (never seen before) launches an EV, the model may not generalize well.

**🔹 4. Improving for Future Predictions**

* Add **battery size (kWh)**, **vehicle class (SUV, sedan, truck)**, and **weight** if available → much stronger predictions.
* Use **regularization (Ridge/Lasso)** to prevent overfitting to specific models.
* Update the model regularly with new EV data (continuous learning).

***Conclusion:***

The analysis of the *Electric Vehicle Population Dataset* provides valuable insights into the current state of EV adoption and distribution across Washington State.

Overall, the dataset shows a **strong and growing trend** in electric vehicle ownership, with significant increases in registrations over recent model years. **Tesla** clearly dominates the market, with models such as the *Model Y* and *Model 3* leading by a large margin. Other manufacturers like *Nissan, Chevrolet, Ford,* and *Kia* also hold notable shares but remain secondary to Tesla’s widespread adoption.

Geographically, EV registrations are **highly concentrated in urban areas**, especially **King County**, which includes Seattle. This concentration reflects the region’s dense population, better access to charging infrastructure, and supportive local policies. Rural counties show fewer registrations, likely due to limited infrastructure and lower EV awareness.

The **electric range vs. MSRP** analysis reveals a general positive correlation — higher-priced vehicles often have better range, though affordable models now offer competitive performance thanks to technological improvements. The **CAFV eligibility** chart indicates that a substantial portion of EVs qualify for clean fuel incentives, though not all vehicles meet the requirements or have verified eligibility data.

In summary, Washington State’s EV landscape is **rapidly expanding**, driven by technological advancements, policy incentives, and consumer preference for sustainable transportation. Continued infrastructure development and equitable access across regions will be key to supporting further EV adoption in the coming years.