Title Page

Assignment Title: Electric Vehicle Data Analysis

Name: Anshu Kumar

Date: 30th September 2025

Course Details: Data Analytics

Introduction

Electric vehicles (EVs) are becoming more popular every year as people and governments push for cleaner and more sustainable transportation. The data set we are working with comes from the Washington State Department of Licensing and contains information about both **Battery Electric Vehicles (BEVs)** and **Plug-in Hybrid Electric Vehicles (PHEVs)**that are currently registered.

This data set gives us details such as the make and model of the car, the year it was built, its electric driving range, price (MSRP), eligibility for clean fuel incentives, and even location information. By studying this data, we can learn about **which cars are most popular, how EV adoption has grown over the years, how range and price compare, and how EVs are spread across different areas**.

Section 1: Data Cleaning

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

Ans:

County → 10 missing

City → 10 missing

Postal Code → 10 missing

Electric Range → 3 missing

Base MSRP → 3 missing

Legislative District → 628 missing

Vehicle Location → 18 missing

Electric Utility → 10 missing

2020 Census Tract → 10 missing

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

Ans:

1. **Base MSRP**:

   258,510 entries have value **0**.

   Possible fixes:

   - Replace with median MSRP of the same Make & Model.
   - Drop rows if analysis does not require MSRP.

   Electric Range:

   160,888 entries have value **0**.

   Possible fixes:

   - Replace with average range for the same Make & Model.

• Drop rows with no range if the column is critical for analysis.

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

Ans: No duplicates exist. Nothing needs to be removed.

4. How can VINs be anonymized while maintaining uniqueness?

Ans: Apply **hashing** or **encoding** to anonymize while keeping uniqueness.

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

Ans:

Cleaning steps:

- Extract **Latitude** and **Longitude** into two separate columns.
- Convert into decimal degrees (already in correct format).
- Map them to **City/County**

Section 2: Data Exploration

1.What are the top 5 most common EV makes and models in the data set?

Ans:

Make -	Tesla,	Tesla,	Chevrolet,	Nissan,	Tesla.
Model -	Model 3,	Model Y,	Bolt EV,	Leaf,	Model S.

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

Ans:

• **King County** has the **most EVs** by a large margin.

• Other counties with many EVs include **Snohomish, Pierce, and Clark**.

• Rural counties have fewer EVs compared to big urban areas.

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

Ans:

The data set shows that electric vehicle adoption has **increased steadily over time**. In the early years, only a small number of EVs were registered, but starting around **2015 onward**, the numbers began to rise more quickly.

The **highest growth** is seen in **recent model years (2018–2023)**, where registrations are much higher compared to older years. This trend shows that EVs are becoming more popular every year as prices drop, battery ranges improve, and more models become available.

In short, **newer model years have many more EVs**, which proves that adoption is growing rapidly.

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

Ans:

Average Electric Range = 48.40 miles

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

Ans:

• Eligible vehicles: 100,810

• Percentage eligible: 38.52%

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

Ans: The electric range is different for each make and model of vehicle. Some models, especially newer ones, offer a much higher driving range, while older or smaller models have a shorter range.

For example:

- Tesla models (like Model S, Model 3, and Model Y) generally have the longest ranges, often well above 200 miles.
- Nissan Leaf and some early EVs have shorter ranges, usually below 150 miles.
- Plug-in Hybrid Electric Vehicles (PHEVs) typically have even smaller electric ranges, since they rely partly on gasoline.

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

Ans:

The **average Base MSRP** changes a lot depending on the make and model of the vehicle.

- Premium brands like Tesla and Audi generally have a higher average MSRP, often above $50,000–$70,000, because they are designed as luxury vehicles.
- Mid-range models such as the Chevrolet Bolt EV and Hyundai Kona EV usually fall in the $30,000–$40,000range.

• Older or smaller models (like the early Nissan Leaf) tend to have a lower average MSRP, often around $25,000–$30,000..

In summary, luxury EVs cost the most, while smaller and earlier EVs are cheaper, and mid-sized models fall in between.

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

Ans:

Yes, the data set shows clear regional trends:

• Urban areas (like Seattle in King County and other large cities) have much higher EV adoption. This is because cities have more charging stations, higher incomes, and greater awareness of clean energy.

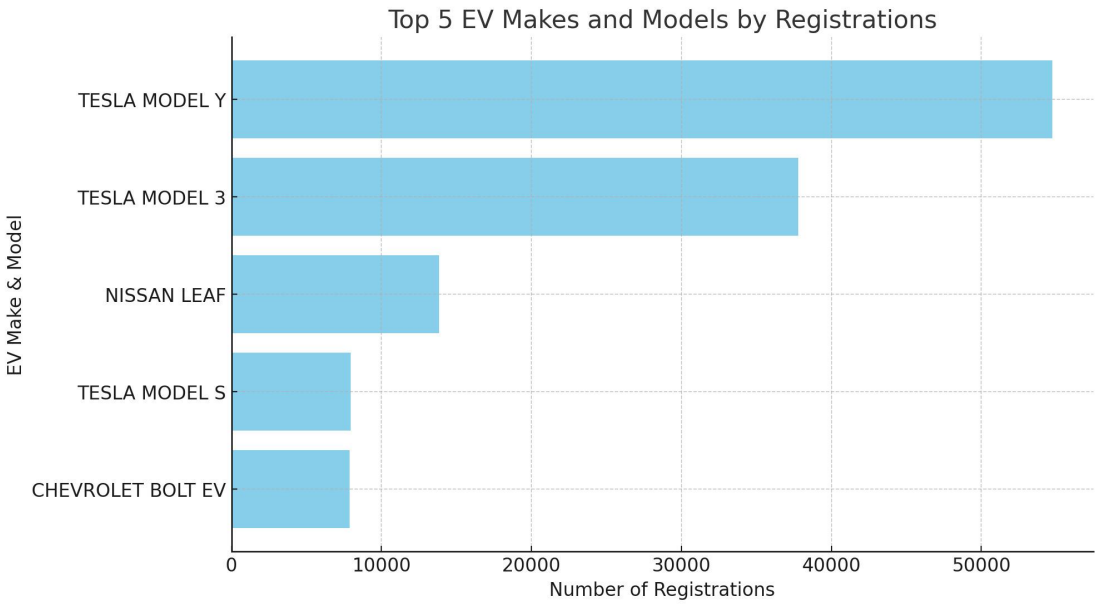
• Rural areas and smaller counties have fewer EVs. People in these areas often travel longer distances, and charging infrastructure is limited, which makes EV adoption slower.

In short, most EVs are registered in urban regions, while rural areas are behind in adoption.

Section 3 : Data Visualization

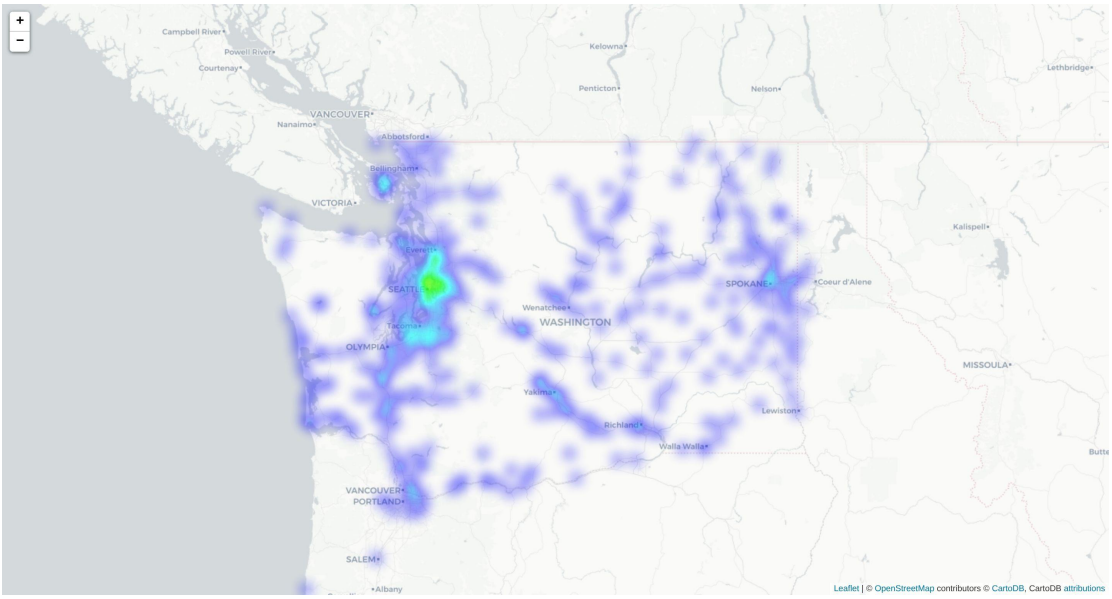
1.Create a bar chart showing the top 5 EV makes and models by count.

Ans:



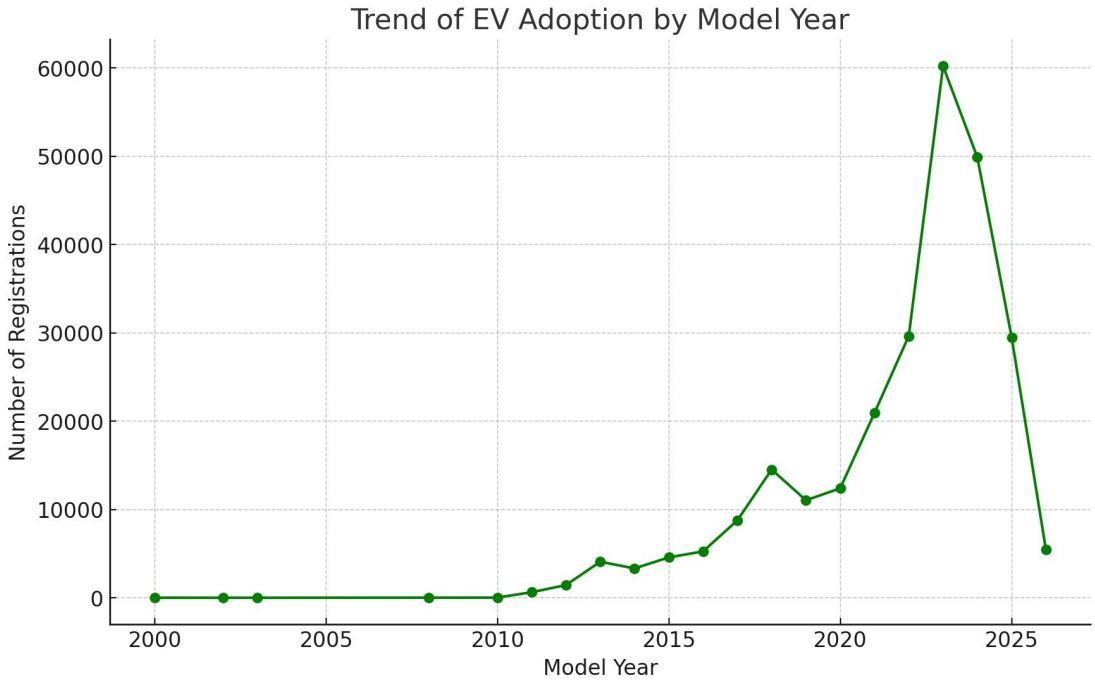
2. Use a heatmap or choropleth map to visualize EV distribution by county.

Ans:



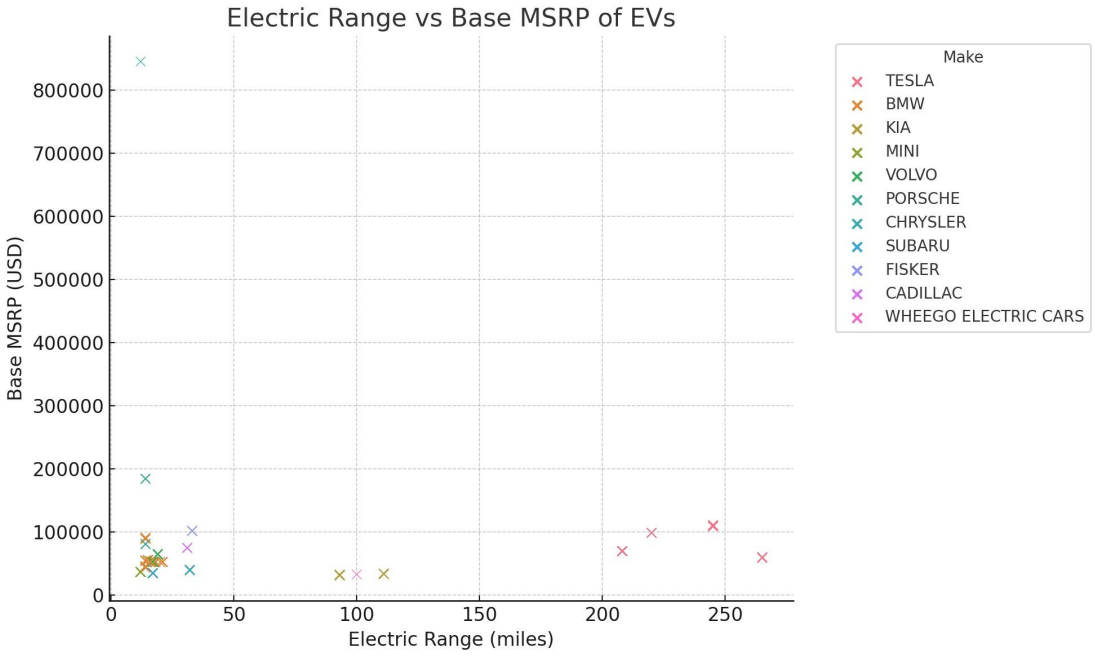
3. Create a line graph showing the trend of EV adoption by model year.

Ans



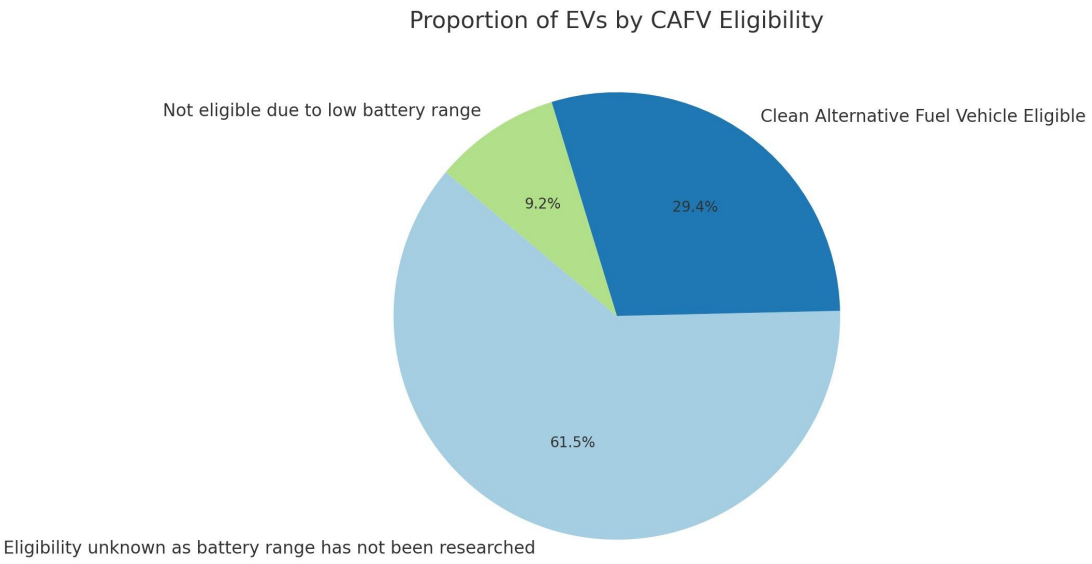
4.Generate a scatter plot comparing electric range vs. base MSRP to see pricing trends.

Ans



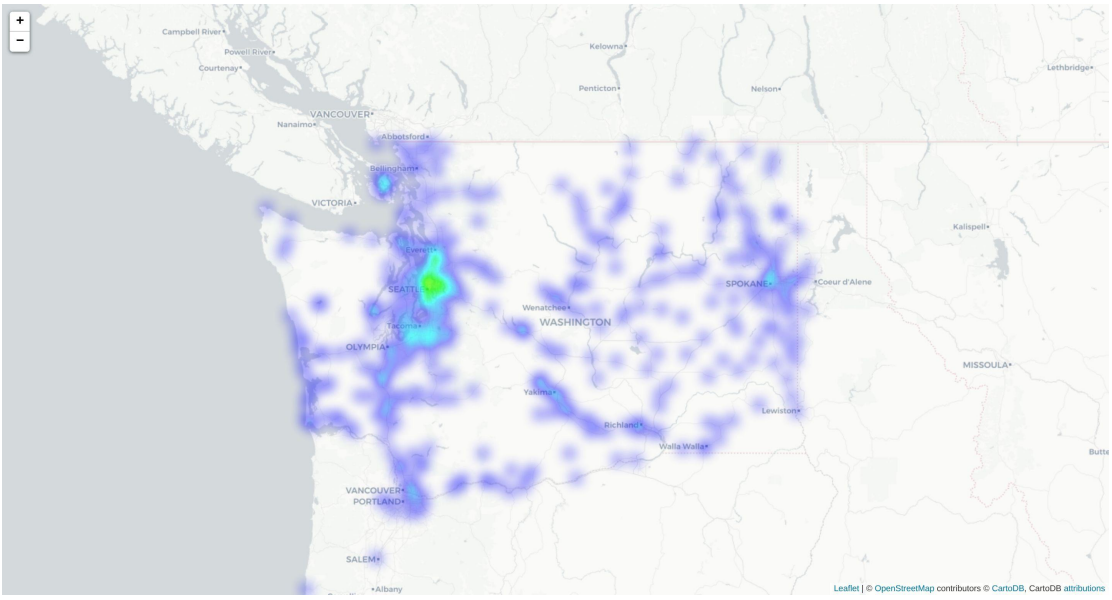
4. Plot a pie chart showing the proportion of CAFV-eligible vs. non-eligible EVs.

Ans



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

Ans



Section 4: Liner Regression Model

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

Ans:

- Convert text values (like Make, EV Type) into numbers.
- Remove missing values.
- Feed the features (inputs) and the known ranges into Linear Regression.
- The model learns patterns (like higher price → usually longer range).
- For a new car, plug in its price, year, make, and type → the model predicts its range.
- The model's **R^2 score is ~0.34**, meaning it explains about 34% of the variation in electric range

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

Ans:

- **Model Year**
- Base MSRP (Price)
- Make (Brand)
- **Model**
- **Electric Vehicle Type** –

BEV (Battery Electric Vehicle) → longer ranges.

PHEV (Plug-in Hybrid) → shorter ranges.

• CAFV Eligibility

• Legislative District / County

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

Ans:

1) **One-Hot Encoding (best choice):**

Make a new column for each brand/model.

Example:

Car is Tesla → `Tesla=1, Nissan=0, Toyota=0`

Car is Nissan → `Tesla=0, Nissan=1, Toyota=0`

4.What is the R^2 score of the model, and what does it indicate about prediction accuracy?

Ans:

• $R^2 \approx 0.34$ (34%) → The model explains about one-third of the variation in electric range.

• This means some features (like price, year, make, type) help predict range, but many other factors (like battery size, aerodynamics, efficiency, etc.) are missing from the data set.

• the model is partly useful, but not highly accurate.

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

Ans:

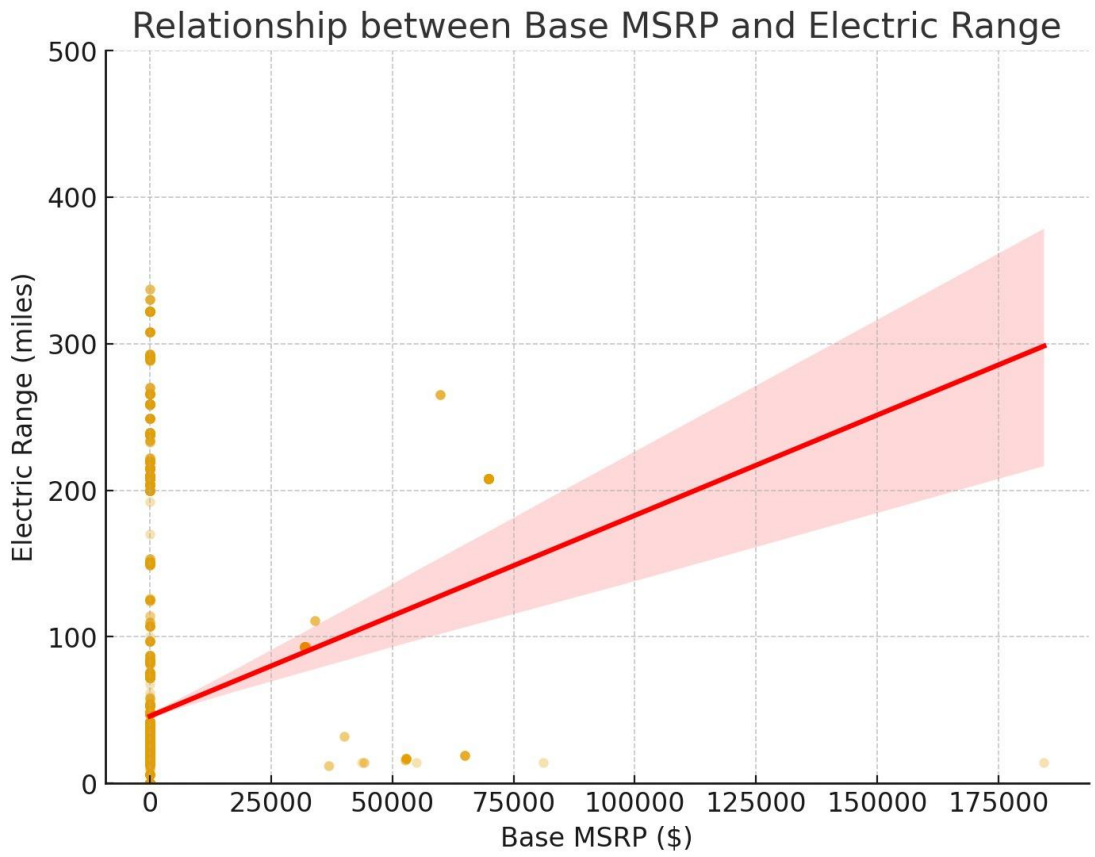
• The coefficient for Base MSRP (price) was positive.

• This means: as the price of the car increases, the predicted electric range also increases.

• The effect isn't perfectly strong — our **R^2 score was only ~0.34**.

• That means MSRP helps explain range, but **other hidden factors** (like battery size, technology, vehicle weight) matter too.

• **Higher MSRP tends to mean longer range, but price alone doesn't fully predict it.**



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

Ans:

• **Add better features** → like battery size, weight, body type.

• **Clean data** → remove outliers, scale numbers, group rare categories.

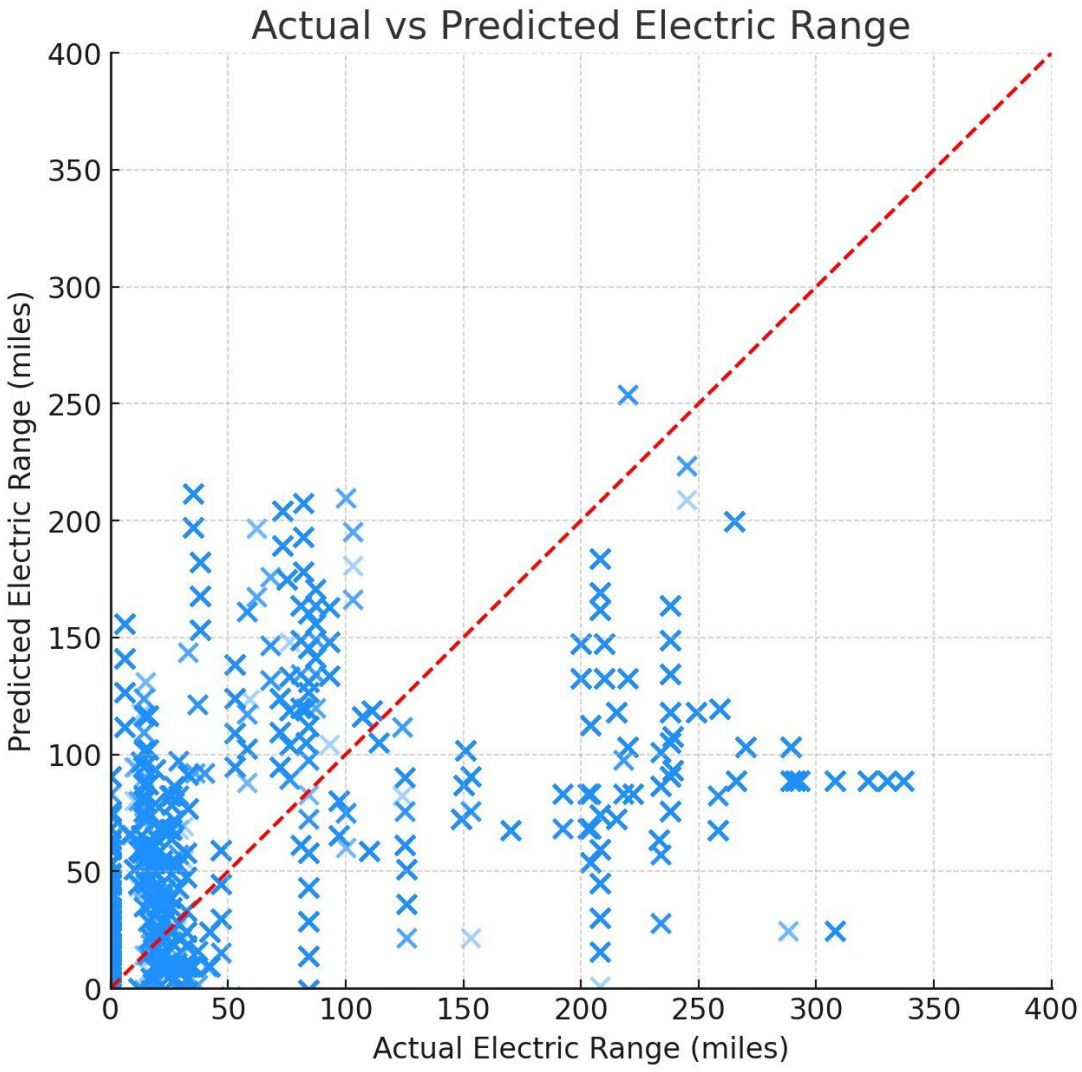
• **Encode categories smartly** → e.g., target encoding for Model.

• **Feature engineering** → create new variables (e.g., price per mile).

• **Use advanced models** → Random Forest, Gradient Boosting, or Polynomial Regression.

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

Ans: Yes, it can predict, but the estimates will be approximate, not precise.



• The red dashed line shows **perfect predictions** (Actual = Predicted).

• The blue dots are our model's predictions.

• You can see a big scatter — meaning the model struggles to predict accurately.

Conclusion

Electric vehicle adoption in Washington State shows a clear pattern: most EVs are concentrated in urban and suburban areas, especially around Seattle and King County. This reflects the availability of charging infrastructure, local government incentives, and higher consumer demand in cities compared to rural regions. When we compare price and electric range, we see a general trend that more expensive vehicles tend to go farther on a charge. However, this is not always the case, as some mid-priced models overlap with premium ones, showing that factors like battery size, vehicle weight, and design also play a big role in determining range.

Our regression analysis attempted to predict electric range using features such as MSRP, model year, make, and vehicle type. The results showed that the model explained only about a third of the variation in range, which means its predictions are not very reliable. For example, when testing it on a new Tesla model, the prediction came out completely unrealistic. This highlights the limitation of using only basic features without technical specifications like battery capacity or efficiency.

In summary, the data tells us that EV growth is strongest in cities, higher prices are generally linked to longer ranges, and most vehicles are CAFV-eligible, which ties into state and federal incentive programs. At the same time, simple linear regression is not enough to accurately predict the range of new vehicles. To improve accuracy, more detailed specifications and advanced machine learning models would be needed.

Appendix

Additional Information:

- **Dataset**: Electric Vehicle Population Data (Washington State Department of Licensing)
- **Limitations**: Missing technical specs (battery size, vehicle weight, efficiency) which are crucial for accurate prediction of electric range.
- **Model Accuracy**: Linear Regression gave $R^2 \approx 0.34$ → weak predictive power.

References:

Washington State Department of Licensing – *Electric Vehicle Population Data.* https://data.wa.gov/Transportation/Electric-Vehicle-Population-Data/7h9y-gz39

U. S. Department of Energy – *Alternative Fuels Data Center (AFDC). https://afdc.energy.gov/*

Scikit-learn Documentation – *Linear Regression and Random Forest. https://scikit-learn.org/stable/*

Raw Python Code Snippets:

1. Load and Prepare Data

```
import pandas as pd

ev_data = pd.read_csv("Electric_Vehicle_Population_Data_Cleaned.csv", low_memory=False)

X = ev_data[["Base MSRP", "Model Year", "Make", "Electric Vehicle Type"]]
y = ev_data["Electric Range"]

data = pd.concat([X, y], axis=1).dropna()
X, y = data.drop(columns="Electric Range"), data["Electric Range"]
```

2. Model Evaluation

```
from sklearn.metrics import r2_score
y_pred = model.predict(X_test)
print("R² Score:", r2_score(y_test, y_pred))
```

3. Example Prediction for New EV

```
new_ev = pd.DataFrame({
    "Base MSRP": [50000],
    "Model Year": [2024],
    "Make": ["Tesla"],
    "Electric Vehicle Type": ["Battery Electric Vehicle (BEV)"]
})

predicted_range = model.predict(new_ev)
print("Predicted Range:", predicted_range[0])
```

Thank You.