🔋 Electric Vehicle Population Dataset Analysis

1. Introduction

The rapid adoption of electric vehicles (EVs) has led to a global shift toward sustainable and environmentally friendly transportation. This project focuses on analyzing data related to Electric Vehicle registrations, characteristics, and environmental impact using Big Data technologies, primarily PySpark, to uncover patterns and insights that can guide policy makers, manufacturers, and consumers.

---

2. Objective of the Project

The main goal of this project is to perform Big Data analysis on a large dataset of electric vehicles to understand:

- The growth trends in electric vehicle adoption.
- The relationship between key variables such as vehicle type, make, model, range, and cost.
- The distribution of EVs across cities and states.
- How vehicle characteristics (like range or MSRP) vary by manufacturer or region.

Through this analysis, the project aims to demonstrate how Big Data analytics using PySpark can extract meaningful insights from large-scale automotive datasets efficiently.

---

3. About the Dataset

Dataset Name: *Electric Vehicle Population Dataset*
File Provided: Electric_Vehicle_Polution_Dataset.csv

This dataset contains records of registered electric vehicles, including information about the make, model, location, price, range, and more.

Columns Description:

Column Name	Description
VIN_1_10	Partial Vehicle Identification Number (unique vehicle reference)
County	County of vehicle registration
City	City where the vehicle is registered
State	State of registration
Postal_Code	ZIP or postal code

Column Name	Description
Model_Year	Manufacturing year of the vehicle
Make	Vehicle manufacturer (e.g., Tesla, Nissan, BMW)
Model	Model name of the vehicle
Electric_Vehicle_Type	Type of EV (e.g., Battery Electric Vehicle (BEV) or Plug-in Hybrid (PHEV))
Clean_Alternative_Fuel_Vehicle_CAFV_Eligibility	Eligibility for clean fuel programs
Electric_Range	Distance (in km) the EV can travel on a full charge
Base_MSRP	Manufacturer's Suggested Retail Price (approximate market price)
Legislative_District	Political district of registration
DOL_Vehicle_ID	Unique identifier for Department of Licensing
Vehicle_Location	Latitude and longitude coordinates
Electric_Utility	Name of the local electricity provider
2020_Census_Tract	Geographic identifier used for demographic analysis

---

4. Tools and Technologies Used

Category	Tools / Frameworks
Big Data Processing	Apache Spark (PySpark)
Programming Language	Python
Visualization	Matplotlib, Seaborn
Environment	Jupyter Notebook
Data Storage Format	CSV file
Optional Tools	Pandas (for easy plotting and summary after Spark processing)

---

5. Data Analysis Steps

1. **Load the dataset** into PySpark DataFrame.
2. **Clean and sanitize** column names for ease of processing.

3. **Handle missing or inconsistent data**.
4. **Perform exploratory data analysis (EDA)** to understand variable distributions.
5. **Analyze relationships** between columns such as price, make, model year, and range.
6. **Generate visual insights** using graphs and correlation matrices.
7. **Interpret results** to draw meaningful conclusions about EV adoption and market trends.

---

6. Key Insights

- **EV Growth Trends:**
  There has been a steady increase in electric vehicle registrations over the past few years, with newer models dominating recent registrations.
- **Popular Manufacturers:**
  Tesla leads in registrations, followed by Nissan, Chevrolet, and Ford.
- **EV Type Distribution:**
  Battery Electric Vehicles (BEVs) make up the majority of the dataset (~70%), while Plug-in Hybrid Electric Vehicles (PHEVs) account for ~30%.
- **Geographic Trends:**
  Urban counties and cities have higher EV adoption rates compared to rural areas. King County (Seattle area) shows the highest EV concentration.
- **Model Popularity:**
  Tesla Model 3, Model Y, and Nissan Leaf are among the most registered EV models.
- **Relationship Analysis:**
  - Newer vehicles are more likely to be BEVs.
  - Cities with higher income and population density show higher EV adoption rates.

7. Conclusion

This project demonstrates the use of **PySpark** for processing and analyzing large datasets efficiently. Through data-driven exploration and visualization, it provides insights into:

- Electric vehicle trends in various regions,
- Price vs. performance relationships, and
- The overall evolution of the electric vehicle ecosystem.

These findings can support decision-making for:

- **Government agencies** (for policy planning),
- **Manufacturers** (for product improvement), and
- **Consumers** (for choosing efficient EV models).