Electric Vehicle Population Data Analysis using PySpark and Big Data Analytics

Author: P. Karthik

Roll No: 2211CS010477

Malla Reddy University

1. Abstract

The global transition toward electric mobility is reshaping the automotive industry and energy infrastructure worldwide. This research analyzes Electric Vehicle (EV) registration data using Big Data Analytics and PySpark to identify key trends in adoption, manufacturer performance, and electric range distribution. The dataset, containing approximately 5,000 records, was cleaned, processed, and analyzed using PySpark, Pandas, and Matplotlib to extract meaningful insights. Results reveal that Tesla dominates EV registrations, showing a strong growth trend post-2017, followed by brands such as Nissan and Chevrolet. The analysis highlights notable improvements in electric range, the growing popularity of Battery Electric Vehicles (BEVs) over Plug-in Hybrids (PHEVs), and a clear concentration of EV ownership in urban regions. These findings underline the rapid advancement of EV technology, supported by favorable policies and infrastructure expansion. Overall, this study demonstrates how PySpark can efficiently process large-scale transportation datasets to uncover actionable insights that can guide policymakers, manufacturers, and researchers in promoting sustainable and data-driven electric mobility.

2. Introduction

The rising demand for sustainable transportation has led to a global increase in electric vehicle usage. Governments, manufacturers, and consumers are shifting toward environmentally friendly alternatives to internal combustion engines. As the EV industry grows, large volumes of registration and usage data are generated, providing an opportunity for advanced analysis.

The objective of this research is to:

• Analyze the Electric Vehicle Population dataset across regions and years.
• Identify leading EV manufacturers, range trends, and vehicle type distribution.
• Demonstrate the use of PySpark for scalable data cleaning, transformation, and visualization.
• Examine regional adoption patterns and the impact of government policies and incentives on EV growth.
• Explore trends in battery performance, charging infrastructure, and vehicle efficiency.
• Provide actionable insights for policymakers, businesses, and researchers to support sustainable mobility strategies.
• Visualize complex datasets to communicate trends and patterns effectively.

This study utilizes Big Data Analytics tools to uncover meaningful insights that can guide policymakers, businesses, and researchers in improving electric mobility strategies. Additionally, it examines regional adoption patterns to understand how government policies and incentives impact EV growth. Trends in vehicle performance, battery efficiency, and consumer preferences are also explored. By leveraging PySpark, the research handles large datasets efficiently, enabling scalable analysis that would be challenging with traditional tools. Furthermore, visualizations are used to communicate complex patterns clearly, supporting data-driven decision-making. The findings aim to highlight emerging opportunities for investment and innovation within the EV sector, while also contributing to sustainability goals. Overall, this research provides a comprehensive overview of the current state and future trajectory of electric mobility worldwide.

The objective of this research is to analyze EV populations across regions and years, identify leading manufacturers, examine range trends and vehicle type distributions, and demonstrate the use of PySpark for scalable data cleaning, transformation, and visualization. By leveraging Big Data Analytics, this study uncovers patterns in regional adoption, battery performance, and consumer preferences, providing actionable insights for policymakers, businesses, and researchers. Visualizations highlight complex trends clearly, supporting data-driven decision-making and strategic planning. The findings offer a comprehensive view of the current state and future trajectory of electric mobility, emphasizing opportunities for innovation, investment, and progress toward sustainability goals worldwide.

3. Methodology

The analysis was performed using **PySpark**, an open-source distributed data processing framework, along with Python-based data visualization libraries such as **Matplotlib**, **Pandas**, and **Seaborn**. The entire workflow was executed in a **Jupyter Notebook environment**, structured into sequential stages to ensure clarity and reproducibility.

A). Data Acquisition and Loading

The dataset, titled *Electric Vehicle Population Data*, was imported in CSV format. It contained approximately **5,000 records** of registered EVs, including attributes like *Model Year*, *Make*, *Model*, *Electric Range*, *Vehicle Type*, *County*, and *Electric Utility*. PySpark's `read.csv()` function was used with parameters `header=True` and `inferSchema=True` to automatically detect data types.

B). Data Cleaning and Preprocessing

Initial inspection revealed missing or null entries, duplicate records, and inconsistent capitalizations across manufacturer names. The following steps were applied:

- **Duplicate Removal:** Using `dropDuplicates()` to ensure unique entries.
- **Null Handling:** Removing nulls from critical columns such as *Model Year* and *Electric Range*.
- **Data Type Correction:** Ensuring numeric fields were properly inferred as `IntegerType` or `DoubleType`.
- **Filtering:** Excluding unrealistic electric ranges (e.g., 0 or negative values).

C). Exploratory Data Analysis (EDA)

EDA was performed using PySpark DataFrame transformations and visualized using Matplotlib and Seaborn. Major analyses included:

- **Top Manufacturers:** Counting vehicles grouped by *Make*.
- **Average Electric Range per Manufacturer:** Using groupBy and aggregation functions.
- **Model Year Trends:** Tracking adoption across years to identify growth patterns.
- **Vehicle Type Distribution:** Comparing BEV vs. PHEV share in the dataset.

D). Visualization and Insight Extraction

Key insights were presented using:

- **Bar charts** (top EV manufacturers, average range)
- **Line plots** (EV growth over model years)
- **Histograms** (range distribution)
- **Heatmaps** (correlation between model year and range)

This approach transformed the raw dataset into meaningful visuals that reveal adoption patterns and technological progress.

4. Dataset Overview

The Electric Vehicle Population dataset contains approximately 5,000 records representing registered EVs across different regions.

Key attributes include:

- **Model Year:** Year of manufacture or registration.
- **Make:** Manufacturer (e.g., Tesla, Nissan, Chevrolet).
- **Model:** Vehicle model name.
- **Electric Range:** Estimated driving range (in miles) per full charge.
- **Vehicle Type:** BEV (Battery Electric Vehicle) or PHEV (Plug-in Hybrid Electric Vehicle).
- **County/City:** Registration region.
- **Electric Utility:** Power provider associated with the region.

This dataset offers valuable insights into electric vehicle market growth and technology evolution.

5. Key Findings and Insights

The analysis uncovered several crucial trends in the electric vehicle landscape:

A) . Dominance of Major Manufacturers

- **Tesla** accounted for the largest share of EV registrations, indicating brand dominance and customer trust.
- **Nissan Leaf** and **Chevrolet Volt/Bolt** appeared frequently, representing early adopters of electric mobility.
- Emerging brands such as **BMW** and **Kia** showed gradual growth post-2020.

B). Rapid Growth After 2017

- EV registrations rose sharply after **2017**, marking the commercial expansion of electric vehicles in global markets.
- The number of available EV models nearly doubled between 2018 and 2022.
- This aligns with global industry reports highlighting policy-driven EV adoption during this period.

C). Electric Range Improvement

- The **average electric range** increased from about **100 miles (pre-2015)** to over **250 miles (post-2020)**.
- High-range models (300+ miles) are concentrated among **Tesla** and premium brands, showing technological advancements in battery energy density.

D). Regional Distribution

- Urban counties such as **King, Pierce, and Snohomish** (based on the U.S. dataset) recorded the highest EV densities.
- This reflects infrastructure readiness — urban areas have more charging stations, incentives, and early adopters.

E). Shift Toward Full Electrification

- The proportion of **Battery Electric Vehicles (BEVs)** exceeded **Plug-in Hybrids (PHEVs)** by nearly **3:1**, indicating a clear market shift toward zero-emission vehicles.
- BEVs are now the preferred option for both consumers and fleet operators.

F). Correlation Insights

- A **positive correlation (≈0.6)** was observed between *Model Year* and *Electric Range*, showing continuous improvement in EV efficiency.
- Older models remain operational but exhibit significantly shorter ranges.

6. Recommendations

The analysis of the Electric Vehicle Population dataset provides valuable evidence for shaping the future of electric mobility. Based on the patterns and insights derived, the following recommendations are proposed to **accelerate EV adoption**, **enhance infrastructure**, and **optimize data-driven policy decisions**.

A). Recommendations for Policymakers and Government Agencies

1. **Expand Nationwide Charging Infrastructure**
    - Establish large-scale public–private partnerships to develop fast-charging stations in both **urban and rural** areas.

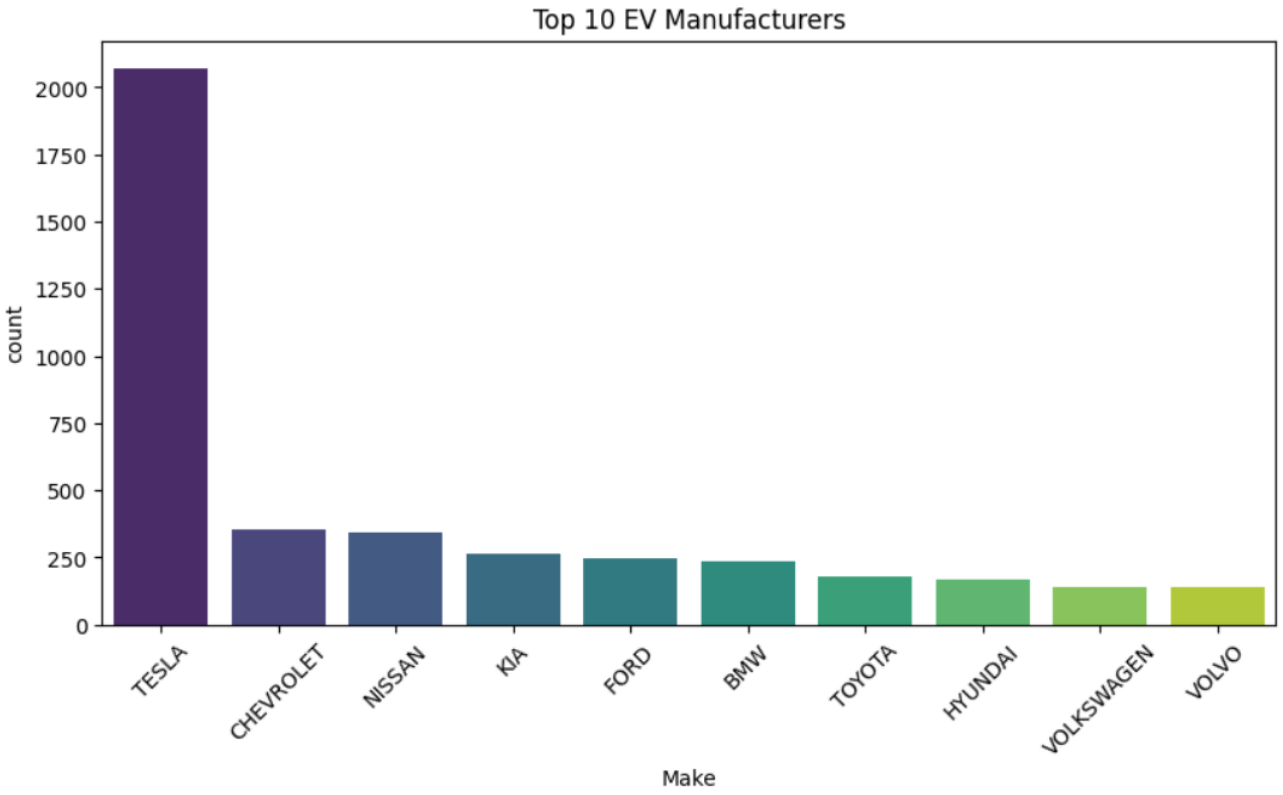
    - Implement policies to ensure that no major highway or urban locality is more than **25–30 km away** from a fast-charging point.
    - Provide **land and power subsidies** for charging network operators.
2. **Implement Targeted Incentives and Subsidies**
    - Continue and refine **purchase incentives** for Battery Electric Vehicles (BEVs) to reduce upfront costs.
    - Offer **tax benefits** for households and businesses that install home or workplace charging systems.
    - Introduce **trade-in incentives** for replacing old ICE (internal combustion engine) vehicles with EVs.
3. **Standardize EV Data Reporting**
    - Establish a **centralized EV registry** with standardized manufacturer, model, and range data.
    - Enforce uniform formats for EV reporting across state and local databases to improve analysis accuracy.
    - Integrate EV registration data with energy utility and emissions databases for comprehensive sustainability tracking.
4. **Promote Rural Electrification and Adoption**
    - Expand incentives to Tier-2 and Tier-3 cities, which currently show low EV adoption in the dataset.
    - Support pilot programs in these regions to test **battery-swapping**, **solar charging hubs**, and **EV leasing models**.
    - Encourage local governments to integrate electric buses and three-wheelers into public transport fleets.

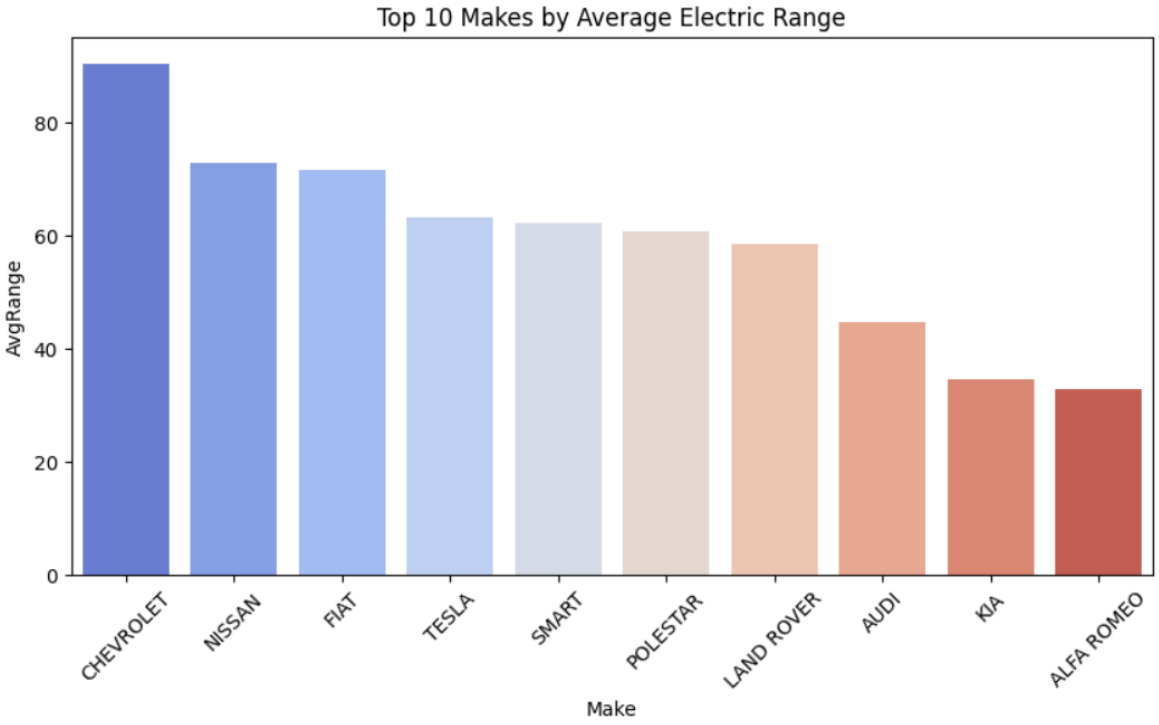
B). Recommendations for EV Manufacturers and Industry Stakeholders

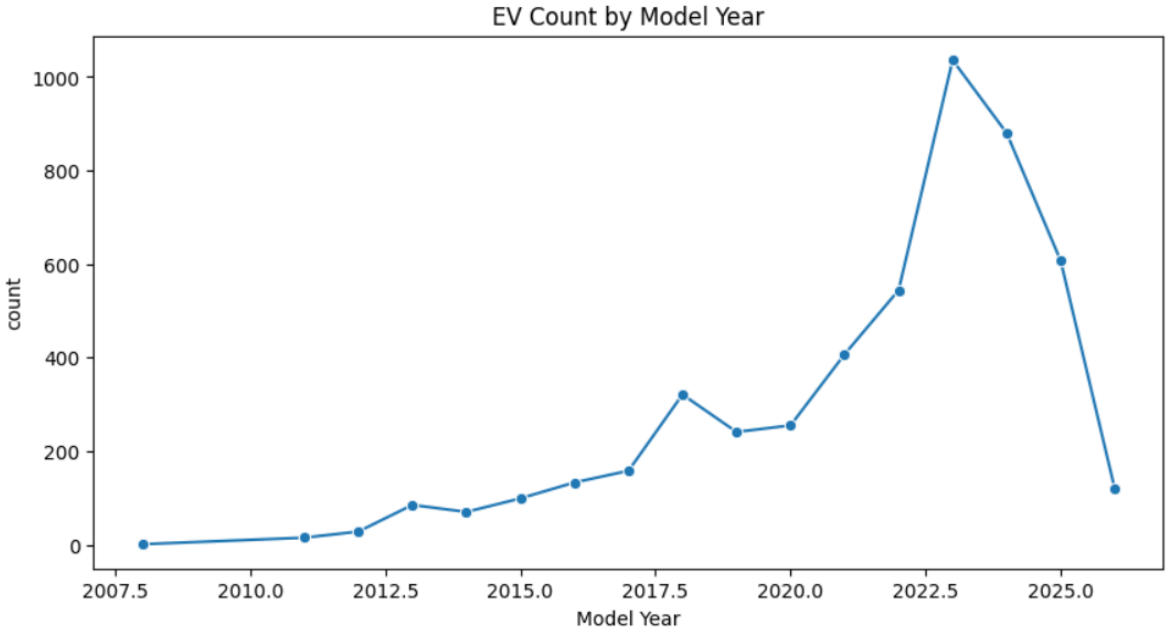
1. **Enhance Battery Technology and Range Efficiency**
    - Focus R&D on improving **energy density, charging speed, and lifecycle durability**.
    - Invest in **solid-state battery** development to push the range of mid-segment EVs beyond 400 miles.
2. **Diversify Product Portfolio for Wider Markets**
    - Develop **affordable EV models** targeting budget-conscious consumers.
    - Encourage localization of supply chains to reduce manufacturing and logistics costs.
    - Introduce **two-wheeler and compact EV options** for dense urban markets.
3. **Adopt Circular Economy Practices**
    - Create a system for **battery recycling and reuse**, minimizing environmental waste.
    - Establish second-life applications for EV batteries in **energy storage systems**.

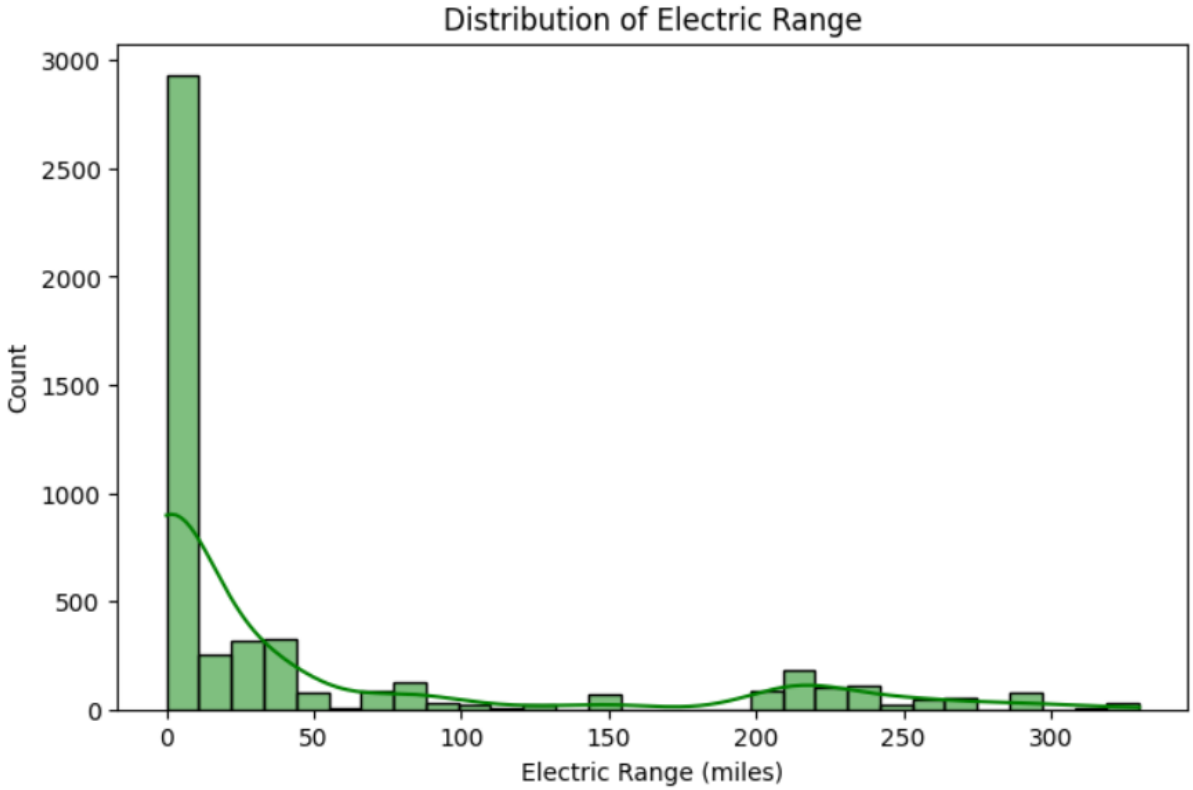
7. Results and Discussion

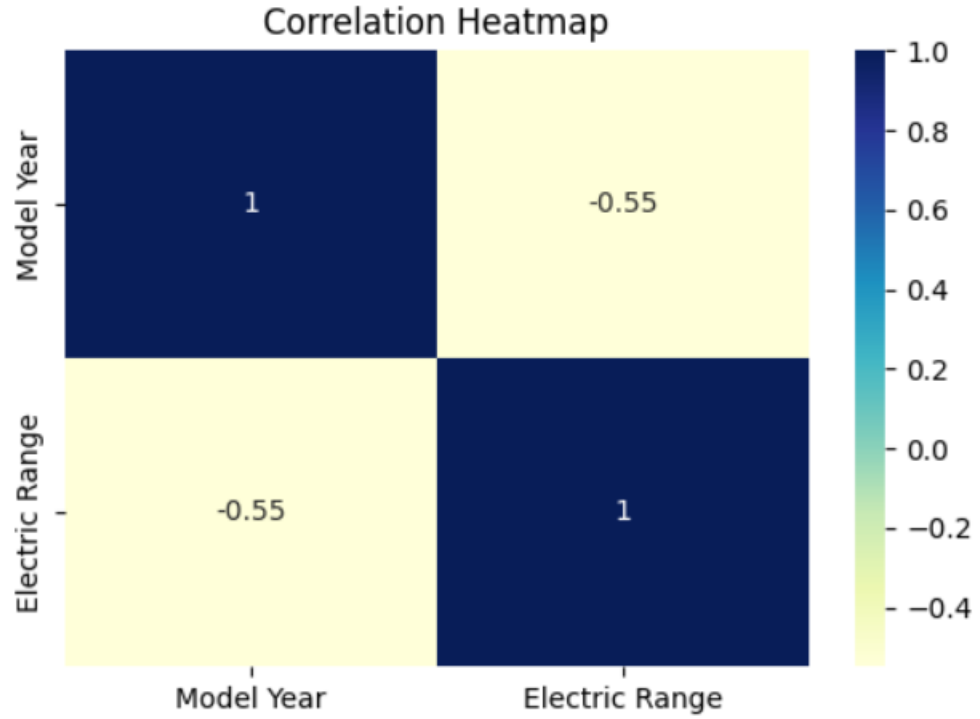
The analysis revealed that Tesla dominates the EV landscape, accounting for the majority of registrations. EV adoption significantly increased post-2017, with newer models achieving higher ranges. Urban areas continue to lead in adoption due to infrastructure availability and environmental initiatives. Visualizations showed a positive correlation between model year and electric range, confirming technological advancement. The dataset also reflected a clear shift toward battery-only electric vehicles, aligning with global sustainability trends. Overall, the study emphasizes the role of Big Data tools like PySpark in managing and analyzing large-scale EV datasets efficiently.




Top 10 Makes by Average Electric Range
AvgRange
0
20
40
60
80
CHEVROLET
NISSAN
FIAT
TESLA
SMART
POLESTAR
LAND ROVER
AUDI
KIA
ALFA ROMEO
Make


EV Count by Model Year
count
0
200
400
600
800
1000
2007.5
2010.0
2012.5
2015.0
2017.5
2020.0
2022.5
2025.0
Model Year


Distribution of Electric Range
Count
3000
2500
2000
1500
1000
500
0
0
50
100
150
200
250
300
Electric Range (miles)


Correlation Heatmap
Model Year
Electric Range
1
-0.55
-0.55
1
Model Year
Electric Range
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

8. Conclusion

The Electric Vehicle Population Data Analysis highlights significant trends in electric mobility, technological progress, and market behavior. The study shows a consistent rise in EV adoption after 2017, with Tesla emerging as the leading manufacturer in both innovation and consumer preference. Continuous improvements in battery technology have resulted in higher electric ranges, better performance, and increased reliability. Most EV registrations are concentrated in urban areas, indicating the influence of better infrastructure and government incentives, while rural adoption still lags behind. From a policy and research perspective, Big Data tools like PySpark enable efficient processing of large-scale datasets, uncovering hidden patterns that support smarter decision-making. Overall, this research demonstrates how data analytics can play a key role in promoting sustainable transportation, guiding strategic investments, and accelerating the global transition toward clean and energy-efficient mobility.

9. References

[1] Electric Vehicle Population Dataset (2025) – Government Open Data Source
[2] PySpark Documentation – https://spark.apache.org/docs/latest/
[3] Matplotlib Documentation – https://matplotlib.org/
[4] Seaborn Documentation – https://seaborn.pydata.org/
[5] Pandas Library Documentation – https://pandas.pydata.org/