Exploratory Data Analysis of Electric Vehicle Population

Project progress report in partial fulfilment of the requirement for the award of the degree of

Master of Computer Applications

Submitted By

ARPAN BISWAS	**University Enrollment No. 12023006015046**
ASHUTOSH SAHA	**University Enrollment No. 12023006015047**
AVIK SARKHEL	**University Enrollment No. 12023006015050**
MAYUKH DUTTA	**University Enrollment No. 12023006015053**
SUDIPTA BISWAS	**University Enrollment No. 12023006015008**

Under the guidance of

Prof. KAUSTUV BHATTACHARJEE

Department of Computer Applications

INSTITUTE OF ENGINEERING & MANAGEMENT, KOLKATA
NEWTOWN
(School of University of Engineering and Management, Kolkata)

University Area, Plot No. III – B/5, New Town,

Action Area – III, Kolkata – 700160.

UNIVERSITY OF ENGINEERING & MANAGEMENT
(Established by Act XXV of 2014 of Govt. of West Bengal & recognized by UGC, Ministry of HRD, Govt. of India)
University Area, Plot No. III-B/5, Main Arterial Road, New Town, Action Area-III, Kolkata - 700 160, W.B., India
Admission Office : 'ASHRAM', GN-34/2, Salt Lake Electronics Complex, Kolkata - 700 091, W.B., India

Ph. (Office) : 91 33 2357 2969
: 91 33 2357 7649
Admissions : 91 33 2357 2059
Fax : 91 33 2357 8302
E-mail : vc@uem.edu.in
Website : www.uem.edu.in

CERTIFICATE

This is to certify that the project titled **Exploratory Data Analysis of Electric Vehicle Population** submitted by **Arpan Biswas (University Enrollment No. 12023006015046), Ashutosh Saha (University Enrollment No. 12023006015047)**, **Avik Sarkhel (University Enrollment No. 12023006015050), Mayukh Dutta (University Enrollment No. 12023006015053)**, **Sudipta Biswas (University Enrollment No. 12023006015008)** students of INSTITUTE OF ENGINEERING AND MANAGEMENT, NEWTOWN, a school of UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA, in partial fulfilment of the requirement for the Degree of Master of Computer Applications, is a Bonafide work carried out by them under the supervision and guidance of Prof. Kaustuv Bhattacharjee during 3rd Semester of the academic session of 2024 - 2025. The content of this report has not been submitted to any other university or institute. I am glad to inform that the work is entirely original and its performance is found to be quite satisfactory.

_____

Prof. Kaustuv Bhattacharjee

HOD

Department of Computer Applications

UEM, Kolkata

_____

Prof. Kaustuv Bhattacharjee

HOD, Department of Computer Applications

UEM, Kolkata

ACKNOWLEDGEMENT

We would like to take this opportunity to thank everyone whose cooperation and encouragement throughout the ongoing course of this project remains invaluable to us.

We are sincerely grateful to our guide **Prof. Kaustuv Bhattacharjee** of the Department of Computer Applications, UEM, Kolkata, for his wisdom, guidance and inspiration that helped us to go through with this project and take it to where it stands now.

We would also like to express our sincere gratitude to **Prof. Kaustuv Bhattacharjee**, HOD, Department of Computer Applications, UEM, Kolkata and all other departmental faculties for their ever-present assistant and encouragement.

Last but not the least, we would like to extend our warm regards to our families and peers who have kept supporting us and always had faith in our work.

Arpan Biswas

Ashutosh Saha

Avik Sarkhel

Mayukh Dutta

Sudipta Biswas

Exploratory Data Analysis of Electric Vehicle Population

Kaustuv Bhattacharjee[1], Arpan Biswas[1], Ashutosh Saha[1], Avik Sarkhel[1], Mayukh Dutta[1], Sudipta Biswas[1]

[1] University of Engineering and Management, Kolkata
[1] Institute of Engineering and Management, Newtown

Abstract

This project focuses on analyzing a dataset of Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) registered in Washington State. Initially containing 200,049 records and 17 parameters, the dataset was cleaned and reduced to 182,137 records and 11 key columns for analysis. Key manipulations included manually filling missing values for columns like 'Electric Range (MILE)' and adding new ones, such as 'Electric Utility Type' and 'Urban/Rural' classifications, to enrich insights.

The analysis was conducted on Google Colab using Python libraries like Pandas, NumPy, and Matplotlib. Two analytical perspectives were adopted: a holistic dataset-wide analysis and a company-wise analysis focusing on the top five EV makers. Questions were developed to explore various parameters, and visualizations were created to simplify insights.

Significant findings include key trends in EV adoption, battery range, and pricing across manufacturers and counties. Future work aims to create an interactive dashboard for county-wise and maker-wise analyses.

Keywords: Electric Vehicles (EVs), Battery Electric Vehicles (BEVs), Plug-in Hybrid Electric Vehicles (PHEVs), Washington State, Data Analysis

1. Introduction

1.1 Background

The field of electric vehicles (EVs) is becoming increasingly important as the world looks for more sustainable and environmentally-friendly transportation solutions. With advancements in technology and growing awareness of climate change, electric vehicles have emerged as a crucial component of reducing carbon emissions from transportation. Electric and plug-in hybrid electric vehicles (PHEVs) have become more popular, offering a clean alternative to traditional gasoline or diesel-powered vehicles. However, there remains a need for further analysis and insights to optimize EV adoption, understand customer preferences, and enhance EV infrastructure.

1.2 Objective

This project aims to analyze a comprehensive dataset of electric vehicles (BEVs and PHEVs) registered through the Washington State Department of Licensing (DOL). The primary objective is to provide valuable insights about EV sales, consumer behavior, and manufacturer performance. Through this analysis, we aim to identify trends, explore the relationship between various factors (such as electric range, vehicle type, and electric utility), and offer actionable recommendations for EV manufacturers, policy-makers, and other stakeholders in the EV industry.

1.3 Scope

The scope of this project covers the entire dataset of EVs registered in Washington State, which initially included 200,049 records and 17 parameters. After cleaning and pre-processing, the dataset has been reduced to 200,043 records with 11 parameters. Our analysis primarily focuses on two viewpoints:

Whole Dataset-Wise Analysis: Analyzing the dataset without any restrictions to provide an initial overview of EV sales, electric utility types, and customer distribution based on different parameters.

Company-Wise Analysis: Focusing on the top 5 EV makers in terms of sales and examining how their vehicles perform across various counties in Washington State. This involves analyzing the top-selling electric vehicles

(EVs) by different manufacturers, as well as visualizing key insights to help organizations, EV manufacturers, and policymakers understand the performance and behavior of EVs in the market.

The project's analysis will help us understand the key factors influencing EV adoption, which will contribute to future product development, policy-making, and market strategies. By making the dataset more accessible, we aim to provide useful insights for EV manufacturers, policymakers, and other stakeholders to make informed decisions, improve products, and contribute to a greener transportation future.

1.4 Significance

This project is significant as it offers a comprehensive, data-driven analysis of EV sales in Washington State, providing critical insights into EV market trends, consumer preferences, and the effectiveness of different electric utilities and regions. With growing concerns about climate change, carbon emissions, and the transition to clean energy, this analysis can help guide EV manufacturers, policymakers, and other stakeholders in making better decisions for promoting EV adoption and improving electric vehicle infrastructure. By visualizing and summarizing the key findings, our project aims to support organizations and EV manufacturers in understanding EV performance metrics, optimizing EV offerings, and ultimately contributing to a more sustainable and environmentally-friendly transportation system.

2. Literature Survey

The document *"Data Analysis and its Importance" by Manpreet Kaur Bhatia* emphasizes the critical role of data analysis in today's data-driven economy, highlighting the exponential growth of data, particularly unstructured data, which poses challenges for organizations. It outlines how effective data analysis is essential for informed decision-making across various sectors, including management, retail, healthcare, and education, ultimately enhancing productivity and creating significant value. The paper also addresses common challenges faced by data analysts, such as data quality and representativeness, which can affect the outcomes of their analyses. Overall, it underscores that data analysis is vital for navigating modern market complexities and achieving strategic business objectives.

The paper *"The use of Big Data Analytics in healthcare" by Batko and Slezak* provides a comprehensive overview of the current state of Big Data Analytics (BDA) in healthcare, particularly within Polish medical facilities. It begins with a literature review that highlights the evolution of BDA, emphasizing its significance in transforming healthcare practices by enabling data-driven decision-making. The authors discuss the complexity of the healthcare system, which involves various stakeholders and is governed by strict regulations, thus necessitating a shift from traditional approaches to more collaborative, data-centric models. The study identifies the types of data utilized by medical facilities, both structured and unstructured, and examines the maturity of these institutions in leveraging BDA for improved patient outcomes and operational efficiency. The findings underscore the potential benefits and challenges of implementing BDA in healthcare, particularly in the context of evolving patient care paradigms and the ongoing demands posed by public health crises, such as the Covid-19 pandemic.

3. Problem Statement

3.1 Definition

The project addresses the challenge of analyzing a large dataset of Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) registered through the Washington State Department of Licensing. The dataset contains several complexities, such as missing data, misclassifications, and complications in analyzing specific attributes like electric utility types and vehicle variants. The goal of the project is to clean and analyze the data to generate actionable insights that can be used by organizations or EV makers to improve their products and services.

3.2 Relevance

This problem is highly relevant as it involves understanding the adoption and performance of electric vehicles in Washington State, an important aspect of the transition to clean energy. By solving these issues, the project aims to provide an accurate and comprehensive view of the EV market. The insights generated will help stakeholders, including manufacturers, government bodies, and consumers, make informed decisions based on accurate data.

3.3 Scope of The Problem

The project focuses on analyzing the EV dataset, which includes a variety of parameters such as electric range, MSRP, make, county, and others. Specifically, it addresses issues related to missing data, misclassifications, and the complex nature of some attributes (such as electric utility types). The analysis will also provide a company-wise breakdown of top EV makers and perform county-wise analysis. The scope excludes broader global EV trends and focuses on Washington State and the vehicles within the dataset.

3.4 Challenges

Several challenges are anticipated in solving this problem:

Data Integrity: A significant portion of the data was missing, requiring manual research to fill gaps, which may introduce inaccuracies or delays.

Misclassification of Vehicles: Some vehicles were misclassified, particularly with regard to vehicle type (e.g., BEVs being marked as PHEVs), requiring manual correction.

Complex Attribute Handling: The ‘Electric Utility’ column contains both single and multiple utility names, which complicates the analysis. A new column was added to simplify the analysis, but this required additional processing.

Model Variants: Some vehicles had multiple variants (PHEV and BEV), and it was challenging to ensure that the correct variant was analyzed, particularly when searching for MSRP and electric range data.

Limited Model Year Data: Since the dataset was collected in 2024, some newer model years (e.g., 2024 and 2025) had limited data, impacting the analysis of recent trends.

4. Proposed Solution

4.1 Approach

- Dataset preprocessing to handle null values and inconsistencies.
- Whole-dataset and manufacturer-wise analyses for insights.

4.2 Design & Architecture

- Utilized Google Colab for data analysis and visualization, leveraging its integrated tools for collaborative work.
- Added new columns, "Urban/Rural" and "Electric Utility Type", to classify counties and simplify utility analysis based on ownership types.
- Inserted values into the "Base MSRP" and "Electric Range (MILE)" columns through extensive web research to correct missing or erroneous data.
- Edited the CAFV eligibility column by manually researching EV models and their ranges (Records with researched electric ranges were categorized as either "Clean Alternative Fuel Vehicle Eligible (Manually Researched)" or "Not eligible due to low battery range (Manually Researched))".
- Fixed classification errors in the "Electric Vehicle Type" column, such as reassigning specific Plug-in Hybrid Electric Vehicles (PHEVs) as Battery Electric Vehicles (BEVs) after verifying their specifications.
- Dropped records with unknown utility types to ensure cleaner analysis.

- Clustered data by model year, vehicle type, and electric range to identify trends in EV adoption and technological advancements over time.

4.3 Tools & Technologies

- Pandas: Used for data manipulation, cleaning, and exploratory data analysis.
- Matplotlib: Created various visualizations, including bar charts and pie charts, to represent trends and distributions.
- NumPy: Performed calculations such as percentages and averages for numerical data analysis.
- Python: Served as the core programming language for implementing all data processing, analysis, and visualization tasks.
- Google Colab Notebook: Provided an interactive environment for collaboration, enabling team members to edit and share the analysis seamlessly.
- MS Excel: Used initially to preprocess and filter the dataset, removing irrelevant parameters and simplifying the data for analysis.

4.4 Workflow

1. Data cleaning in Excel and Python.
2. Question-based analysis at dataset and manufacturer levels.
3. Visualized findings using pie charts and bar graphs.

5. Experimental Setup and Result Analysis

5.1 Setup

- **Environment**: Google Colab for cloud-based analysis and collaboration.
- **Libraries**: Pandas (data manipulation), Matplotlib (visualizations), NumPy (calculations).
- **Dataset**: EV dataset from Washington DOL, preprocessed to include 182,137 records across 11 parameters.

5.2 Data Collection

- Preprocessing involved removing null values, manually correcting "Base MSRP" and "Electric Range" data, and adding derived columns such as "Urban/Rural" and "Electric Utility Type."

5.3 Key Results

- **Top Manufacturers**: Tesla (44%), Chevrolet (7%), Nissan (7%).
- **Top Counties**: King (56%), Snohomish (13%), Pierce (8%).
- **Vehicle Types**: BEVs (79%) dominate over PHEVs (21%).
- **Electric Range**: Snohomish (240 miles) and King (231 miles) lead in mean ranges.

5.4 Visualizations

- **Pie Charts**: Highlighted top counties and manufacturers.
- **Bar Charts**: Showed EV type distribution, electric ranges by county, and utility provider contributions.
- **Line Charts**: Depicted advancements in BEV ranges and pricing trends over model years.

6. Conclusion

6.1 Summary of Work

This project analyzed the electric vehicle (EV) population in Washington State, focusing on adoption patterns, preferences, and trends. The analysis involved cleaning and preprocessing the dataset, correcting missing values, categorizing electric utilities, and exploring key metrics such as county-level EV sales, manufacturer dominance, CAFV eligibility, vehicle types, and electric ranges. Relationships between electric range, MSRP, and model years were also explored.

6.2 Key Findings

1. **EV Adoption Patterns:**
   - King County led with 56% of EV sales, followed by Snohomish (13%) and Pierce (8%).
   - Tesla accounted for 44% of EVs.
2. **Vehicle Types and Preferences:**
   - BEVs made up 79%, while PHEVs were 21%.
   - Snohomish County had the highest mean electric range at 240 miles.
3. **CAFV Eligibility:**
   - 90% of EVs were CAFV-eligible after manual corrections.
4. **Utility Analysis:**
   - Puget Sound Energy Inc. and the City of Tacoma served 40% of EVs, emphasizing the role of infrastructure.
5. **Trend Analysis:**
   - BEVs showed increased electric range due to battery advancements, while PHEVs remained stable.
   - BEVs had higher MSRPs, reflecting better technology and range.

6.3 Significance

The project offers valuable insights into Washington State's EV market, emphasizing Tesla's dominance and the importance of utility infrastructure. The findings can inform policies to boost EV adoption and infrastructure development.

6.4 Limitations

Data Quality: Missing or inaccurate values required manual corrections.

Geographical Scope: The analysis focused only on Washington State.

Unexplored Parameters: Factors like charging infrastructure and consumer income were not analyzed.

Temporal Data: The dataset did not account for future trends.

7. Future Scope

7.1 Possible Enhancements:

The project can be expanded by utilizing additional parameters like recharge time (zero to full), number of airbags, safety ratings, and vehicle types (e.g., sedan, SUV, hatchback, convertible). These additions would provide a deeper insight into consumer preferences and vehicle performance. The *Base MSRP* column, which is critical for price analysis, needs to be completed by the data provider for more accurate analysis. Future work could also focus on integrating interactive dashboards for dynamic exploration and applying machine learning models for trend prediction.

7.2 Real-World Applications:

EV companies could professionally use this data analysis, especially when a comprehensive company-wise analysis is completed. Insights gained from the study can inform product development, particularly in enhancing performance and safety features. Policymakers can leverage this information to promote EV adoption and infrastructure.

7.3 Challenges for Future Research:

Challenges include gaps in data, such as missing or zero values in critical columns like *Base MSRP* and *Electric Range (MILE)*, which impacted this analysis. Increasing the dataset size could add complexity, providing a more comprehensive understanding but requiring more sophisticated tools and algorithms for analysis.

References

- Bhatia, M. K. & Institute of Innovation, Technology and Management, Guru Gobind Singh Indraprastha University, New Delhi, India-110058. (2017). Data analysis and its importance. International Research Journal of Advanced Engineering and Science, 2(1), 166–168. https://irjaes.com/wp-content/uploads/2020/10/IRJAES-V2N1P58Y17.pdf
- Batko, K., & Ślęzak, A. (2022). The use of Big Data Analytics in healthcare. Journal of Big Data, 9(1). https://doi.org/10.1186/s40537-021-00553-4
- Washington State Office of Financial Management. (2017). Based on Washington State Office of Financial Management, April 2017. https://doh.wa.gov/sites/default/files/legacy/Documents/Pubs/609003.pdf
- Wikipedia contributors. (2024, October 22). Wheego LiFe. Wikipedia. https://en.wikipedia.org/wiki/Wheego_LiFe\
- https://www.edmunds.com/
- https://www.caranddriver.com/
- https://www.kbb.com/
- https://www.solarwa.org/utilities_washington_state
- https://d3n8a8pro7vhmx.cloudfront.net/solarwa/pages/54/attachments/original/1513141161/Washington_Electric_Utilities-large_map.pdf?1513141161