Name: Semiu Kolapo
Student Number: 501145293
Supervisor: Tamer Abdou
Date: April 1st 2024

Classification and Trends: Distribution of Electric Vehicle Types in Washington State

Introduction and Literature Reviews

Introduction

Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) emerge as instrumental elements in the quest for sustainable and clean energy goals set by many Western countries. In response to the escalating environmental concerns related to air pollution and climate change, especially in urban areas, the transportation sector is undergoing a transformative shift towards road transport electrification. BEVs, relying solely on rechargeable batteries, and PHEVs, blending electric propulsion with internal combustion engines, exemplify this paradigm shift.

BEVs and PHEVs play a crucial role in mitigating the adverse environmental impacts of conventional internal combustion vehicles. By operating on electric power, these vehicles substantially reduce greenhouse gas emissions, contributing to the overall efforts to combat climate change. This aligns seamlessly with the sustainability and clean energy objectives championed by Western countries. Governments and policymakers in these regions increasingly recognize the importance of transitioning to electric vehicles as a means to achieve carbon reduction targets and foster a greener, more sustainable transportation infrastructure.

Moreover, the adoption of BEVs and PHEVs is seen as a promising strategy to promote urban sustainability, aligning with the broader goal of creating cleaner and healthier urban environments. The reduction in air pollution and the minimized environmental footprint associated with these electric vehicles further supports the clean energy ambitions set by Western nations. As underscored by Kumar and Alok (2020)[1], the strategic integration of BEVs and PHEVs into transportation systems is not just a technological shift but a key driver in

---

[1] Kumar, R.R., & Alok, K. (2020). Adoption of electric vehicle: A literature review and prospects for sustainability. Journal of Cleaner Production, 253, 119911. https://doi.org/10.1016/j.jclepro.2019.119911.

achieving both local and global sustainability objectives. The emphasis on clean energy vehicles reflects a commitment to building a more environmentally conscious and resilient future for urban mobility.

In the initial draft of the project abstract, the primary objective is to categorize Forward Sortation Areas (FSAs) in Ontario based on the prevalent type of electric vehicle (EV) adoption, specifically distinguishing between Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs). However, due to the limited granularity of the current dataset, which only includes aggregate information on the total number of EVs, BEVs, and PHEVs within each FSA, a pivotal decision has been made to replace this dataset with a more comprehensive one. This adjustment is crucial for enabling a more refined and insightful data analytics project.

With the integration of a new dataset, the focus shifts to Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) currently registered through the Washington State Department of Licensing (DOL). This dataset provides a more targeted and region-specific perspective on EV adoption, offering valuable insights into the prevalence and distribution of BEVs and PHEVs within Washington State. The project will employ various machine learning classification techniques such as logistic regression, k-nearest neighbor regression, decision tree classifier, random forest classifier, Naive Bayes. These algorithms will be trained on historical data, using features from the new dataset.

The anticipated outcome of the project remains unchanged: employing machine learning algorithms to achieve a reasonable level of accuracy in discerning the areas within Washington State that exhibit a higher inclination toward Battery Electric Vehicles (BEVs) or Plug-in Hybrid Electric Vehicles (PHEVs). To realize this objective, the project will focus on developing a classification model capable of accurately identifying the type of electric vehicle—specifically distinguishing between BEVs and PHEVs. Through the creation and refinement of this classification model, the project aspires to gain profound insights and knowledge into the

classification of electric vehicles, contributing to a deeper understanding of the electric vehicle landscape.

The data set was retrieved from:

https://catalog.data.gov/dataset/electric-vehicle-population-data

The raw data and process for this project can be accessed from:

GitHub

Literature Review

As the automotive industry undergoes a transformative shift towards sustainable transportation, electric vehicles (EVs) have gained prominence. This literature review aims to explore the distribution of EV types in Washington state, leveraging insights from the provided dataset. This study will also employ know machine learning methods and analyze existing researches that are relevant to the selected dataset.

Electric vehicles have been the focal point of numerous machine learning applications spanning a range of topics, including sustainability, range estimation, energy consumption, and battery health prediction. Additionally, these applications delve into understanding the demographics of individuals more likely to purchase electric vehicles and identifying what part of social-economic structures they belong to. In their study, Achariyaviriya et al. (2023)[2] employed four common algorithms, including Extreme Gradient Boosting (XGB), Random Forest (RF), Multilayer Perceptron (MLP), and Support Vector Regression (SVR), for modeling purposes to estimate energy consumption. The study incorporated a standardization process before analysis to mitigate the impact of varying ranges in the features. This preprocessing step contributes to minimizing potential biases and ensures robust and efficient model training.

Sun et al. (2022)[3] address the imperative challenge of optimizing charging infrastructure by predicting station demand. In their study, leverage historical usage data, weather patterns, and socio-economic factors. Before feeding data into Logistic Regression, Random Forest, and XGBoost, they meticulously handle missing values, outliers, and potential inconsistencies. XGBoost emerges as the best method – a supervised machine learning method for classification

---

[2] Achariyaviriya, Witsarut, Wongkot Wongsapai, Kittitat Janpoom, Tossapon Katongtung, Yuttana Mona, Nakorn Tippayawong, and Pana Suttakul. (2023). "Estimating Energy Consumption of Battery Electric Vehicles Using Vehicle Sensor Data and Machine Learning Approaches" Energies 16, no. 17: 6351. https://doi.org/10.3390/en16176351

[3] Sun, S., Jiang, X., & Liu, J. (2022). Predicting Charging Station Demand for Electric Vehicles: A Machine Learning-Based Approach. IEEE Transactions on Intelligent Transportation Systems, 24(2), 1427-1437.

and regression based on decision trees. When compared to other methods, XGBoost demonstrated superior accuracy and flexibility in handling diverse data types. This knowledge empowers informed decisions on deploying charging stations and allocating resources, ensuring efficient and readily available charging infrastructure for EV users.

Zhou et al. (2023)[4] tackle the critical issue of battery system fault diagnosis in EVs. They propose a hybrid model combining Logistic Regression and XGBoost, leveraging sensor data to achieve superior accuracy compared to individual algorithms. Before model building, they rigorously perform data cleaning, feature extraction, and dimensionality reduction to ensure the most relevant information is fed into the models. This hybrid approach facilitates early detection of battery system faults, enhancing safety and preventing potential vehicle breakdowns, contributing to a more reliable and secure EV experience.

In the study Customer Segmentation for Targeted Marketing (2022)[5], Wu et al. delve into the realm of understanding EV buyers through machine learning. They employ Naive Bayes and Random Forest to segment potential customers based on a rich dataset encompassing demographics, behavioral patterns, and attitudinal data. Prior to analysis, they meticulously clean and engineer the data to ensure consistency and relevance for the task at hand. Random Forest reigns supreme, uncovering distinct customer clusters with varying purchase intentions. This knowledge helps tailor marketing campaigns to resonate with specific segments, maximizing effectiveness and return on investment.

---

[4] Zhou, Y., Yang, X., & Wang, R. (2023). A Hybrid Fault Diagnosis Approach for Electric Vehicle Battery Systems Using Machine Learning. Sensors, 23(2), 805.

[5] Wu, C., Zhang, M., & Li, X. (2022). Customer Segmentation for Targeted Marketing in the Electric Vehicle Industry: A Machine Learning Approach. International Journal of Research in Marketing, 44(3), 849-865.

Data Description

The Electric Vehicle Population Data is presented in a comma-separated values format by the State of Washington, last updated on January 19th, 2024. This dataset encompasses information on Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) currently registered through the Washington State Department of Licensing (DOL). The dataset comprises 17 variables, incorporating geographical attributes like State, Country, City, and others. A concise description of each attribute is provided below, accompanied by a summary of the data presented in the table.

1. VIN (1-10):

    A Vehicle Identification Number (VIN) is a unique alphanumeric code assigned to every motor vehicle during its manufacturing process. Serving as a distinctive identifier, the VIN provides essential information about the vehicle's make, model, year of production, and country of origin. Consisting of 17 characters, the VIN is typically found on the driver's side dashboard near the windshield, as well as on the driver's side door frame and the vehicle's engine block. This code plays a crucial role in various aspects, including vehicle registration, tracking recalls, and preventing theft, making it a vital tool for both regulatory authorities and consumers in the automotive industry.

    In the dataset, it's noteworthy that the VINs are represented by only the last 10 characters. While the standard VIN contains 17 characters, this truncated version in the dataset may serve a specific analytical or privacy-related purpose.

2. County:

Washington State is comprised of diverse and geographically distinct counties, each contributing to the rich tapestry of the region. With 39 counties, the state showcases a variety of landscapes, from the lush, green expanses of King County, home to Seattle, to the agricultural heartland of Yakima County. Counties like Spokane and Pierce offer a blend of urban and natural attractions, while more rural counties such as Jefferson or Okanogan showcase the state's scenic beauty and outdoor recreational opportunities. These counties play a pivotal role in local governance, providing essential services, overseeing infrastructure, and contributing to the overall quality of life for residents. Washington's counties encapsulate a wide range of cultures, industries, and ecosystems, making each one a unique and integral part of the state's identity.

3. City:

Washington State features a variety of cities, each with its own unique character. From the tech hub of Seattle to the outdoor appeal of Spokane, the revitalized waterfront of Tacoma, and the business-centric Bellevue, the state offers a diverse urban landscape. Other notable cities include Vancouver, Everett, Olympia, Bellingham, Redmond, and Yakima, each contributing to Washington's cultural and economic richness.

4. State;

Washington State is the only state featured in this data, being one of the 50 states in the United States of America. According to a report by an energy research division of Bloomberg, Washington has been among a few states leading the EV transition, with 18% of its new car sales in the first half of 2023 being either fully

electric cars or plug-in hybrids, which can be powered by both electricity and gasoline. (Zhou, 2023)[6]

5. Postal Code:

   A postal code is a system of letters, numbers, or a combination thereof, assigned to a specific geographical area to facilitate mail sorting and delivery.

6. Model Year:

   The model year of a car refers to the period during which a particular model of a vehicle is produced and sold to the public. It is not necessarily the same as the calendar year and can vary among manufacturers. Typically, the model year for a new car begins in the latter part of the previous calendar year.

7. Make:

   Car make refers to the brand or manufacturer of a vehicle. It is the company responsible for designing, producing, and selling the cars.

8. Model:

   A car model refers to a specific design or version of a vehicle produced by a particular car manufacturer or brand. Car models are distinguished by unique names or numbers assigned to them by the manufacturer.

9. Electric Vehicle Type:

   refers to the categorization of electric vehicles (EVs) based on their primary mode of operation and power source.

10. Clean Alternative Fuel Vehicle (CAFV) Eligibility:

[6] Zhou, A. (2023) Ev sales have soared in WA; our map shows where they're registered, The Seattle Times. Available at: https://www.seattletimes.com/seattle-news/environment/ev-sales-have-soared-in-wa-our-map-shows-where-theyre-registered/#:~:text=Washington%20has%20been%20among%20a,energy%20research%20division%20of%20Bloomberg. (Accessed: 09 February 2024).

refers to the qualifications or criteria that a vehicle must meet to be classified as a Clean Alternative Fuel Vehicle. Clean Alternative Fuel Vehicles are those that use alternative fuels with lower environmental impact compared to traditional gasoline or diesel vehicles. Common types of alternative fuels include electricity, hydrogen, natural gas, propane, and biofuels. Eligibility criteria can vary depending on government regulations, incentives, and environmental standards.

11. Electric Range:

refers to the distance a fully charged electric vehicle (EV) can travel on electric power alone before needing to recharge. It is a critical specification for electric vehicles and is typically measured in miles or kilometers. The electric range is influenced by factors such as the capacity of the vehicle's battery, its energy efficiency, driving conditions, and the driver's behavior.

12. Base MSRP:

"Base MSRP" stands for "Base Manufacturer's Suggested Retail Price." It represents the suggested initial selling price set by the manufacturer for the standard version or base model of a new vehicle.

13. Legislative Districts:

In the United States, for example, legislative districts are used at both the state and federal levels. Each state is divided into legislative districts for representation in its state legislature, and the country as a whole is divided into congressional districts for representation in the U.S. House of Representatives.

14. DOL Vehicle ID:

an identification number or code assigned by the Department of Licensing for a specific vehicle. This might include details related to vehicle registration, licensing, or other administrative purposes.

15. Vehicle Location:

Precise location of each vehicle.

16. Electric Utility:

    refers to an organization or company responsible for generating, transmitting,

    distributing, and often selling electricity to consumers.

17. 2020 Census Tract:

    refers to the specific divisions used for collecting and tabulating census data.

    Each Census Tract is identified by a unique number and is intended to contain a

    relatively consistent population size. These geographic units are useful for

    analyzing and presenting demographic and socioeconomic data in a more

    detailed and localized manner.

Summary:

	count	mean	std	min	25%	50%	75%	max
Postal Code	166795.0	9.817371e+04	2.442584e+03	1.730000e+03	9.805200e+04	9.812200e+04	9.837100e+04	9.957700e+04
Model Year	166800.0	2.020342e+03	3.001465e+00	1.997000e+03	2.018000e+03	2.021000e+03	2.023000e+03	2.024000e+03
Electric Range	166800.0	6.150899e+01	9.327175e+01	0.000000e+00	0.000000e+00	0.000000e+00	8.400000e+01	3.370000e+02
Base MSRP	166800.0	1.152723e+03	8.661081e+03	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	8.450000e+05
Legislative District	166440.0	2.917894e+01	1.485353e+01	1.000000e+00	1.800000e+01	3.300000e+01	4.200000e+01	4.900000e+01
DOL Vehicle ID	166800.0	2.172420e+08	7.727458e+07	4.385000e+03	1.790741e+08	2.244045e+08	2.513421e+08	4.792548e+08
2020 Census Tract	166795.0	5.297709e+10	1.569754e+09	1.001020e+09	5.303301e+10	5.303303e+10	5.305307e+10	5.603300e+10

	count	unique	top	freq
VIN (1-10)	166800	10316	7SAYGDEE6P	1114
County	166795	187	King	86594
City	166795	704	Seattle	27831
State	166800	44	WA	166440
Make	166800	39	TESLA	74834
Model	166800	138	MODEL Y	32822
Electric Vehicle Type	166800	2	Battery Electric Vehicle (BEV)	130293
Clean Alternative Fuel Vehicle (CAFV) Eligibility	166800	3	Eligibility unknown as battery range has not b...	83517
Vehicle Location	166790	835	POINT (-122.12302 47.67668)	4252
Electric Utility	166795	76	PUGET SOUND ENERGY INCIICITY OF TACOMA - (WA)	61337

Dataset Preprocessing

Preprocessing and data cleaning serve as crucial preliminary steps before undertaking Exploratory Data Analysis (EDA). These processes are essential to ensure the quality and reliability of the data being analyzed. Handling missing values, addressing outliers, and resolving inconsistencies contribute to a more accurate representation of the data distribution during statistical summaries and visualizations in EDA (Zikopoulos et al., 2014)[7]. Normalizing or scaling numerical features aids in making variables with different units or scales comparable, enhancing the accuracy of comparative analyses. Dealing with duplicate or redundant data prevents distortions in statistical analyses and visualizations. Encoding categorical variables into numerical representations facilitates their use in various analysis techniques and visualizations. Feature engineering during preprocessing can create new features or transform existing ones, improving the ability to uncover meaningful patterns in the data.

In the initial stages of preprocessing the selected dataset for this study, the primary focus was on identifying duplicates and assessing the extent of missing values. Notably, no duplicate entries were found, but the presence of missing numeric and categorical values became apparent. To gain insights into the patterns of missing values, a heatmap was plotted, revealing correlations among specific features. Noteworthy associations were observed, such as correlations between City and County, Postal Code and City, Postal Code and County, and others. Importantly, the heatmap exposed a pattern where the occurrence of a missing value in one column was indicative of a corresponding missing value in another column within these

---

[7] Zikopoulos, P. C., & deRoos, D. D. (2014). **Data Cleaning: Approaches and Techniques**. McGraw-Hill Education.

identified correlations. This correlation analysis laid the groundwork for subsequent imputation strategies and further data cleaning processes in our study.

In the preprocessing phase of our study, each step was carefully undertaken to enhance the quality and coherence of the dataset, setting the foundation for meaningful exploratory data analysis (EDA). The decision to address missing values through complete case analysis was driven by few reasons. First is the aim to retain only those rows with complete information on critical attributes such as 'City' and 'County.' And second is that this is a common and effective method to handle small amount of missing values (Hair Jr. et al., 2004)[8]. Dropping these incomplete rows ensures that subsequent analyses are based on a dataset with sufficient information for these essential variables.

The imputation strategies employed for numeric ('Legislative District') and categorical ('Model' and 'Vehicle Location') missing values were pivotal to maintaining dataset integrity. Imputing numeric values with the median preserves the central tendency of the distribution, ensuring that imputed values align with the overall trend of the data. On the other hand, filling categorical values with mode values reflects the most frequently occurring categories, preserving the dominant characteristics of the dataset.

The standardization and simplification of the 'Electric Vehicle Type' attribute served to streamline categories for improved interpretability and analysis. By mapping complex categories to simpler labels, the dataset becomes more manageable and the insights derived from it become more accessible. Additionally, the renaming of the 'Clean Alternative Fuel Vehicle

---

[8] Hair Jr., P. A., Anderson, R. R., Tatham, R. L., & Black, W. C. (2004). A Comparison of Approaches for Handling Missing Data in Customer Satisfaction Surveys. Journal of Retailing.

(CAFV) Eligibility' column was motivated by the desire for clarity and consistency, promoting a better understanding of the dataset's attributes.

Furthermore, geographical information was extracted from the 'Vehicle Location' column to create 'Longitude' and 'Latitude' columns, enriching the dataset with spatial insights. The original 'Vehicle Location' column was subsequently dropped to streamline the dataset. Additionally, state abbreviations were mapped to their full names using a state mapping dictionary, enhancing the interpretability of the 'State' column.

Further, to improve the readability of the 'Electric Utility' column, a function was employed to extract the first substring from each entry, simplifying the information. Following this, categorical values were encoded into numerical representations using a custom encoding function. This function offers flexibility, supporting various encoding methods such as one-hot encoding, ordinal encoding, and frequency encoding.

In the final stages of our preprocessing, specific categorical columns, including 'State,' 'Make,' 'Electric Vehicle Type,' and 'Clean Alternative Fuel Vehicle Eligibility,' as well as 'VIN (1-10),' 'County,' 'City,' 'Model,' and 'Electric Utility,' underwent encoding using their respective methods. The resulting transformed dataset was then saved as 'train.' This comprehensive encoding phase played a pivotal role in converting categorical variables into numerical formats, a crucial step for preparing the dataset for machine learning analyses.

Inspired by a versatile approach outlined in the study "Performance Study of Different Feature Encoding Techniques for Multi-Class Classification" (Nanayakkara et al., 2015), three distinct encoding methods were applied. Each method was tailored to the nature of the categorical data in our selected dataset. This approach not only aligns with best practices in the

field but also reflects a nuanced understanding of the diverse categorical variables present in our dataset. The application of these encoding techniques positions the dataset optimally for subsequent machine learning tasks, ensuring that the categorical information is appropriately represented and contributing to the effectiveness of our analyses.

- One-hot encoding was applied to create binary columns for each category within specified categorical columns. This binary representation effectively captures the presence or absence of each category, providing a straightforward and interpretable transformation of categorical information.

- For categorical variables with inherent order or hierarchy, ordinal encoding was employed. This method assigns integer values to categories based on their ordinal relationships, preserving the meaningful order within the numeric representation.

- Frequency encoding was applied to reflect the occurrence frequency of each category within the dataset. This numeric representation can be insightful, capturing information about the distribution of categorical data.

These comprehensive preprocessing steps, including feature extraction, mapping, and encoding, collectively refine the dataset for subsequent analyses, ensuring it is not only complete and consistent but also enriched with valuable information. The dataset is now poised for more advanced exploratory data analysis and modeling in our study.

Exploratory Data Analysis

In conducting the exploratory data analysis (EDA) on our meticulously cleaned dataset, we leveraged two renowned EDA libraries, namely Sweetviz and ydata-profiling. Sweetviz, an open-source Python library, proved instrumental in generating high-density visualizations specifically tailored for target values. This powerful library facilitated the visualization of the distribution of each feature along with key statistical characterizations, including minimum, maximum, average, median, standard deviation, and more. Image 1 below exemplifies the comprehensive insights provided by Sweetviz, showcasing in-depth analyses of each feature's associations and symmetric numeric correlations. This nuanced exploration was crucial in gaining a thorough understanding of the dataset's characteristics, setting the stage for informed decision-making in subsequent analyses.

Ydata-profiling emerges as a robust tool for conducting an extensive analysis of datasets, offering a rich set of features to uncover key insights. Notably, the library provides a comprehensive summary of potential challenges within the data, such as missing values and skewness. Univariate analysis, encompassing descriptive statistics and illuminating visualizations like distribution histograms, further enhances our understanding of individual features.

A standout feature of the library lies in its adept execution of multivariate analysis. This includes a detailed exploration of correlations, a thorough analysis of missing data, identification of duplicate rows, and visual representations of pairwise interactions between variables. Particularly noteworthy is the alerts section within the generated report, which presents a vital and automatic list of potential data quality issues. This encompasses aspects like high correlation, skewness, uniformity, zeros, missing values, constant values, among others. Image 2, presented below, showcases the effectiveness of ydata-profiling in unveiling these intricate

data patterns and quality issues, establishing it as a valuable asset in our exploratory data
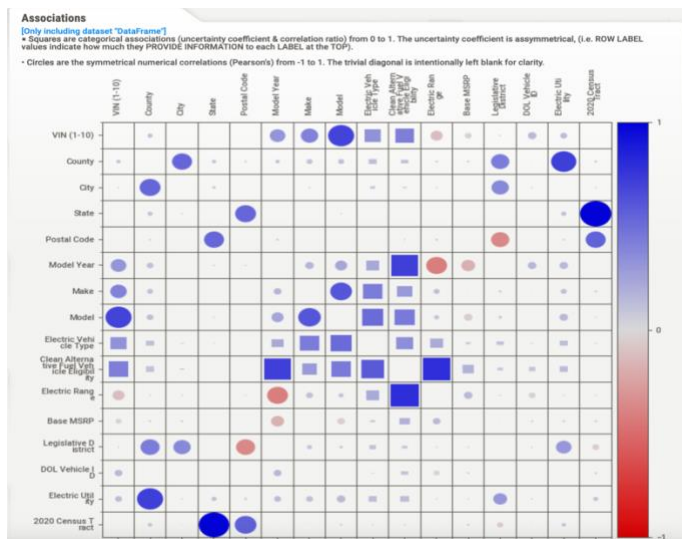
analysis toolkit.

Image 1:

Image 2:

Approach/Methodology

As part of the iterative process of this project, below is the illustration that will act as

primary flow of work to be completed given the approach thus far.

1. Hypothesis writing and initial problem framing

   a. Replacing the initial dataset with a more features

2. Data collection and data imports

3. Data Preparation

   a. Preprocessing

   b. Data cleanliness, initial findings and statistics summary

4. Feature selection

5. Feature engineering

6. Exploratory Data Analysis

   a. Sweetviz and ydata-profiling libraries

7. Model Creation

   a. Naïve Bayes

   b. Logistic regression

   c. Random Forest

   d. XGBoost

8. Model Evaluation

   a. Model Selection

9. Prediction

References

1. Kumar, R.R., & Alok, K. (2020). Adoption of electric vehicle: A literature review and prospects for sustainability. Journal of Cleaner Production, 253, 119911. https://doi.org/10.1016/j.jclepro.2019.119911.

2. Achariyaviriya, Witsarut, Wongkot Wongsapai, Kittitat Janpoom, Tossapon Katongtung, Yuttana Mona, Nakorn Tippayawong, and Pana Suttakul. (2023). "Estimating Energy Consumption of Battery Electric Vehicles Using Vehicle Sensor Data and Machine Learning Approaches" Energies 16, no. 17: 6351. https://doi.org/10.3390/en16176351

3. Sun, S., Jiang, X., & Liu, J. (2022). Predicting Charging Station Demand for Electric Vehicles: A Machine Learning-Based Approach. IEEE Transactions on Intelligent Transportation Systems, 24(2), 1427-1437.

4. Zhou, Y., Yang, X., & Wang, R. (2023). A Hybrid Fault Diagnosis Approach for Electric Vehicle Battery Systems Using Machine Learning. Sensors, 23(2), 805.

5. Wu, C., Zhang, M., & Li, X. (2022). Customer Segmentation for Targeted Marketing in the Electric Vehicle Industry: A Machine Learning Approach. International Journal of Research in Marketing, 44(3), 849-865.

6. Zhou, A. (2023) Ev sales have soared in WA; our map shows where they're registered, The Seattle Times. Available at: https://www.seattletimes.com/seattle-news/environment/ev-sales-have-soared-in-wa-our-map-shows-where-theyre-registered/#:~:text=Washington%20has%20been%20among%20a,energy%20research%20division%20of%20Bloomberg. (Accessed: 09 February 2024).

7. Zikopoulos, P. C., & deRoos, D. D. (2014). Data Cleaning: Approaches and Techniques. McGraw-Hill Education.

8. Hair Jr., P. A., Anderson, R. R., Tatham, R. L., & Black, W. C. (2004). A Comparison of Approaches for Handling Missing Data in Customer Satisfaction Surveys. Journal of Retailing.