Exploring Electric Vehicle Adoption Trends in Washington State

Literature Review

Student: Arunkumar Cherungot-Meppadathkalam

Student No.: 501277792

Github Repository: https://github.com/acherungotTMU/TMU_Capstone

Supervisor: Tamer Abdou

Date of Submission: 07th June 2024

Ryerson University

Table of Contents

Introduction

The rise in electric vehicle (EV) adoption is a key component in the transition to sustainable transportation. This literature review aims to provide a comprehensive analysis of the factors influencing EV adoption in Washington State, focusing on geographic and infrastructure related factors, and to evaluate the use of machine learning models in predicting future EV adoption trends.

1. **What is already known about the topic?**

   Electric vehicle adoption is influenced by a variety of factors, including environmental concerns, government policies, financial incentives, and advancements in EV technology. Socio-economic factors such as income, education, and occupation also play significant roles. Apart from these factors, geographic location and infrastructure such as electric utility provider also plays an important role in deciding the adoption rates of electric vehicles.

2. **Critical analysis of what is already known?**

   Existing studies have established that financial incentives and government policies significantly impact EV adoption. However, there is a gap in understanding the combined effect of geographic and infrastructure related factors, particularly in Washington State. Moreover, the application of machine learning models for predicting EV adoption trends based on these factors is relatively underexplored.

3. **Has Anyone Else Ever Done Anything Exactly the Same?**

   No studies have been found that exactly replicate the current research focus on Washington State using the same dataset and machine learning approaches. However, similar studies have been conducted in other regions, focusing on different aspects of EV adoption.

4. **Has Anyone Else Done Anything That is Related?**

   Although there are no articles or research papers that have conducted this specific type of research, there are related studies that emphasize similar techniques and methodologies. These studies include:

1. Bursac, Z., Gauss, C. H., Williams, D. K., & Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. Source Code for Biology and Medicine, 3(1), 17.

- Summary: This paper discusses strategies for selecting variables in logistic regression models.

- Relevance: It provides a methodology for variable selection that can be applied to our analysis.

- Link: Read here

2. Nowling, R. J. (2015). Categorical Variable Encoding and Feature Importance Bias with Random Forests.

- Summary: Examines how different categorical encoding methods, specifically one-hot and integer encoding, affect feature importance scores in Random Forest models.

- Relevance: The study validates the use of one-hot encoding for categorical variables, ensuring unbiased feature importance scores in Random Forest models, which aligns with and supports the methodological approach.

- Link: Read here

3. Alanazi, R., et al. (2022). Identification and prediction of chronic diseases using machine learning approaches. Journal of Big Data.

- Summary: This paper explores the use of machine learning for predicting chronic diseases.

- Relevance: The methodology for feature engineering and model evaluation can be adapted to our research.

- Link: Read here

4. Da Poian, V., Theiling, B., Clough, L., McKinney, B., Major, J., Chen, J., & Hörst, S. (2023). Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry. *Frontiers in Astronomy and Space Sciences, 10*. - Summary: This paper discusses the significance of outliers in boxplots.

- Relevance: Demonstrates how advanced data analysis techniques can be applied to scientific data to extract meaningful insights

- Link: Read here

5. Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. Behavioral Ecology, 17(4), 688-690.

- Summary: This study highlights the use of the unequal variance t-test.

- Relevance: It offers alternative statistical methods for hypothesis testing.

- Link: Read here

6. Snijders, T. A., & Bosker, R. J. (1999). Multilevel analysis: An introduction to basic and advanced multilevel modeling. Sociological Methods & Research, 28(2), 201-223.

- Summary: This paper introduces multilevel modeling techniques.

- Relevance: Multilevel models can help analyze the hierarchical nature of our geographic data.

- Link: Read here

7. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis. John Wiley & Sons.

- Summary: A comprehensive guide to linear regression analysis.

- Relevance: It provides foundational knowledge for applying linear regression in our study.

-Link: Read here

8. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

- Summary: This article explains the use of Random Forest for classification and regression tasks.

- Relevance: It provides a practical approach to implementing Random Forest in our analysis.

- Link: Read here

9. Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?. Geoscientific Model Development Discussions, 7(1), 1525-1534.

- Summary: This paper compares RMSE and MAE as metrics for model evaluation.

- Relevance: It helps us decide on appropriate metrics for evaluating our predictive models.

- Link: Read here

10. Schneider, C., Dryhurst, S., Kerr, J., Freeman, A., Recchia, G., Spiegelhalter, D., & van der Linden, S. (2021). COVID-19 risk perception: A longitudinal analysis of its predictors and associations with health protective behaviours in the United Kingdom. Journal of Risk Research, 24.

- Summary: This study analyzes the predictors and associations of COVID-19 risk perception.

- Relevance: The statistical methods used can inform our approach to hypothesis testing.

- Link: Read here

11. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 881-892.

- Summary: This paper presents an efficient implementation of the k-means clustering algorithm.

- Relevance: It provides a basis for using k-means clustering to identify patterns in our data.

- Link: Read here

12. Xiaodong Wu, Y., & Zeng, Y. (2019). Using Apriori Algorithm on Students' Performance Data for Association Rules Mining. IEEE International Conference on Data Science and Advanced Analytics (DSAA).

- Summary: This study uses the Apriori algorithm to find associations in student performance data.

- Relevance: It guides the application of the Apriori algorithm for identifying patterns in our EV adoption data.

- Link: Read here

13. Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection—a review and recommendations for the practicing statistician. Biometrical Journal, 60(3), 431-449.

- Summary: This paper reviews various methods for variable selection in statistical modeling.

- Relevance: It offers insights on selecting the most relevant variables for our models.

- Link: Read here

14. Wood, J. D., Dykes, J., Slingsby, A., & Clarke, K. (2007). Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geo visualization mashup. IEEE Transactions on Visualization and Computer Graphics, 13(6), 1176-1183.

- Summary: This paper explores methods for interactive visual exploration of large datasets.

- Relevance: It provides techniques for visualizing geographic patterns in our EV adoption data.

- Link: Read here

15. Qasim, H. M., Ata, O., Ansari, M. A., Alomary, M. N., Alghamdi, S., & Almehmadi, M. (2021). Hybrid feature selection framework for the Parkinson imbalanced dataset prediction problem. *Medicina, 57*(11), 1217.

- Summary: Explores the use of SMOTE for balancing datasets and RFE for feature selection to improve the accuracy of Parkinson's disease predictions.

- Relevance: Demonstrates the effectiveness of combining SMOTE and RFE techniques for handling imbalanced datasets and selecting important features

- Link: Read here

16. Imani, M., & Arabnia, H. R. (2023). Hyperparameter optimization and combined data sampling techniques in machine learning for customer churn prediction: A comparative analysis. *Technologies, 11*(6), 167

- Summary: Explores the application of various machine learning techniques, including hyperparameter tuning and combined data sampling methods, for predicting customer churn in the telecommunications sector.

- Relevance: It improves the predictive accuracy of logistic regression and random forest models.by employing SMOTE

- Link: Read here

These references offer valuable insights and methodologies that are relevant to our research on EV adoption, providing a strong foundation for our analysis.

6. Where Does my Work Fit in With What Has Gone Before?

This research builds on existing studies by focusing specifically on Washington State and using a comprehensive dataset from the Department of Licensing. The application of machine learning models to predict EV adoption based on demographic and socio-economic factors is a novel contribution.

7. Why is this Research Worth Doing in the Light of What Has Already Been Done?

Understanding the factors driving EV adoption in Washington State can help policymakers and businesses develop targeted strategies to promote EV adoption. Predictive models can provide valuable insights into future trends, aiding in infrastructure planning and policy formulation. The analysis will be grouped based on city, county and legislative districts as these variables will be necessary to give an insight to the respective policy makers of the state.

Based on my research, policy makers can have an insight on the growth rate of electric vehicles. From these insights, they can make important decisions related to changes in electric vehicle registration policy, number of new charging stations to be installed, design new plans and policies to boost electric vehicle use in respective location and what changes electric utility companies can make in the future based on the growth rate of electric vehicles.

Descriptive Statistics and Exploratory Data Analysis (EDA)

The dataset includes over 181,000 observations with variables such as VIN, County, City, State, Postal Code, Model Year, Make, Model, Electric Vehicle Type, CAFV Eligibility, Electric Range, Base MSRP, Legislative District, DOL Vehicle ID, Vehicle Location, Electric Utility, and 2020 Census Tract.

Description of Variables:

- VIN (1-10): The first ten characters of the Vehicle Identification Number, which uniquely identifies each vehicle.
- County: This is the geographic region of a state that a vehicle's owner is listed to reside within. Vehicles registered in Washington state may be located in other states.
- City: The city in which the registered owner resides.
- State: This is the geographic region of the country associated with the record. These addresses may be located in other states.
- Postal Code: The 5-digit zip code in which the registered owner resides.
- Model Year: The model year of the vehicle, determined by decoding the Vehicle Identification Number (VIN).
- Make: The manufacturer of the vehicle, determined by decoding the Vehicle Identification Number (VIN).
- Model: The model of the vehicle, determined by decoding the Vehicle Identification Number (VIN).
- Electric Vehicle Type: Indicates whether the vehicle is a Battery Electric Vehicle (BEV) or a Plug-in Hybrid Electric Vehicle (PHEV), distinguishing between different types of electric vehicles.

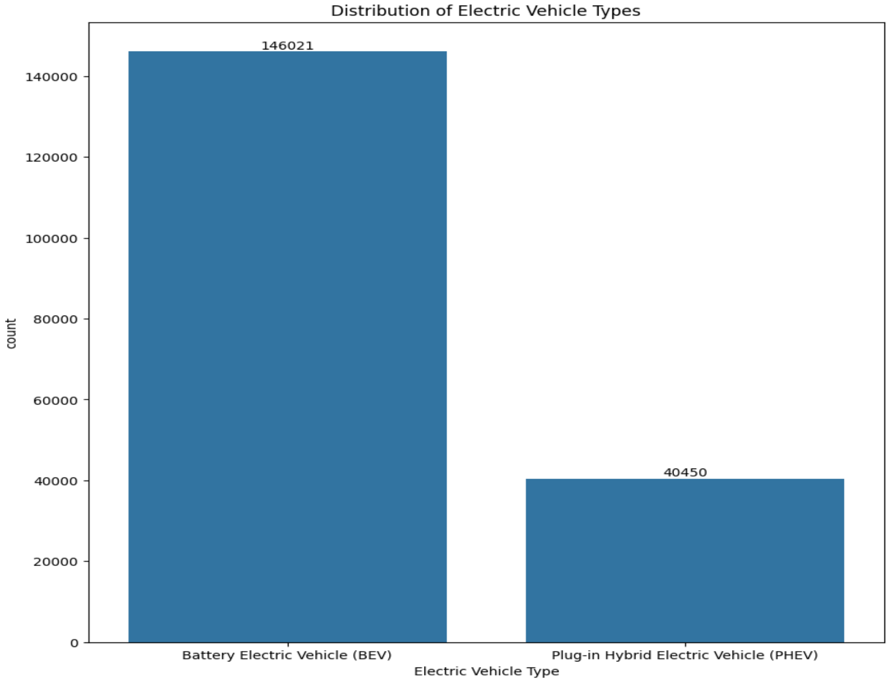
- Clean Alternative Fuel Vehicle (CAFV) Eligibility: This categorizes vehicle as Clean Alternative Fuel Vehicles (CAFVs) based on the fuel requirement and electric-only range requirement in House Bill 2042 as passed in the 2019 legislative session.
- Electric Range: Describes how far a vehicle can travel purely on its electric charge.
- Base MSRP: This is the lowest Manufacturer's Suggested Retail Price (MSRP) for any trim level of the model in question.
- Legislative District: The specific section of Washington State that the vehicle's owner resides in, as represented in the state legislature.
- DOL Vehicle ID: An identification number assigned by the Washington State Department of Licensing, serving as a unique identifier within the licensing system.
- Vehicle Location: The center of the ZIP Code for the registered vehicle.
- Electric Utility: This is the electric power retail service territories serving the address of the registered vehicle. All ownership types for areas in Washington are included: federal, investor owned, municipal, political subdivision, and cooperative. If the address for the registered vehicle falls into an area with overlapping electric power retail service territories, then a single pipe | delimits utilities of same TYPE and a double pipe || delimits utilities of different types.
- 2020 Census Tract: The census tract identifier is a combination of the state, county, and census tract codes as assigned by the United States Census Bureau in the 2020 census, also known as Geographic Identifier (GEOID).

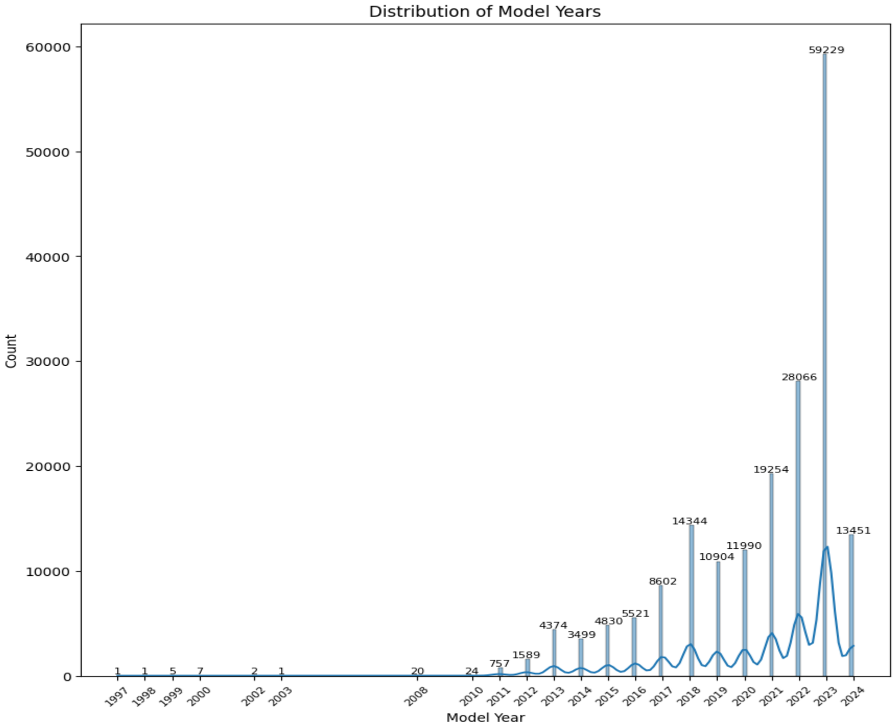
Univariate Analysis and Bivariate Analysis:

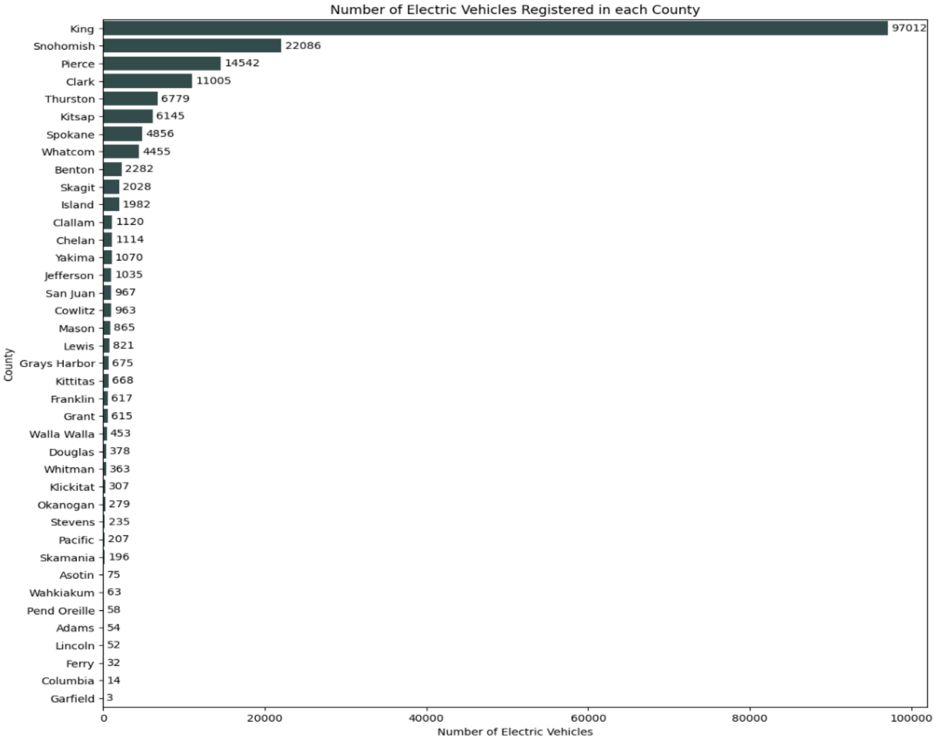
A descriptive statistic including data types of each variable, missing values, correlations and data distribution is done using the y-data profiling library in python. The output, in HTML format is uploaded in GitHub repository.

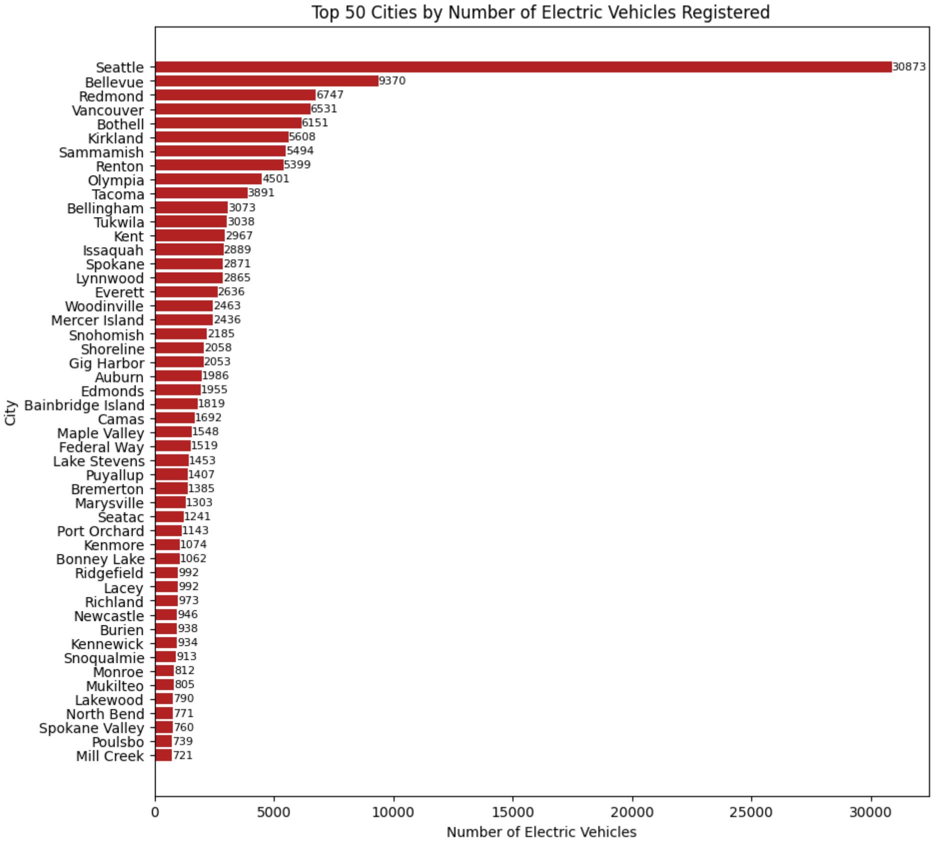
To draw a better picture, visualization tools of python have been used to depict the univariate and bivariate analysis of electric vehicle population data and the same are mentioned below. This will help in understanding the variable fluctuations and relationships:

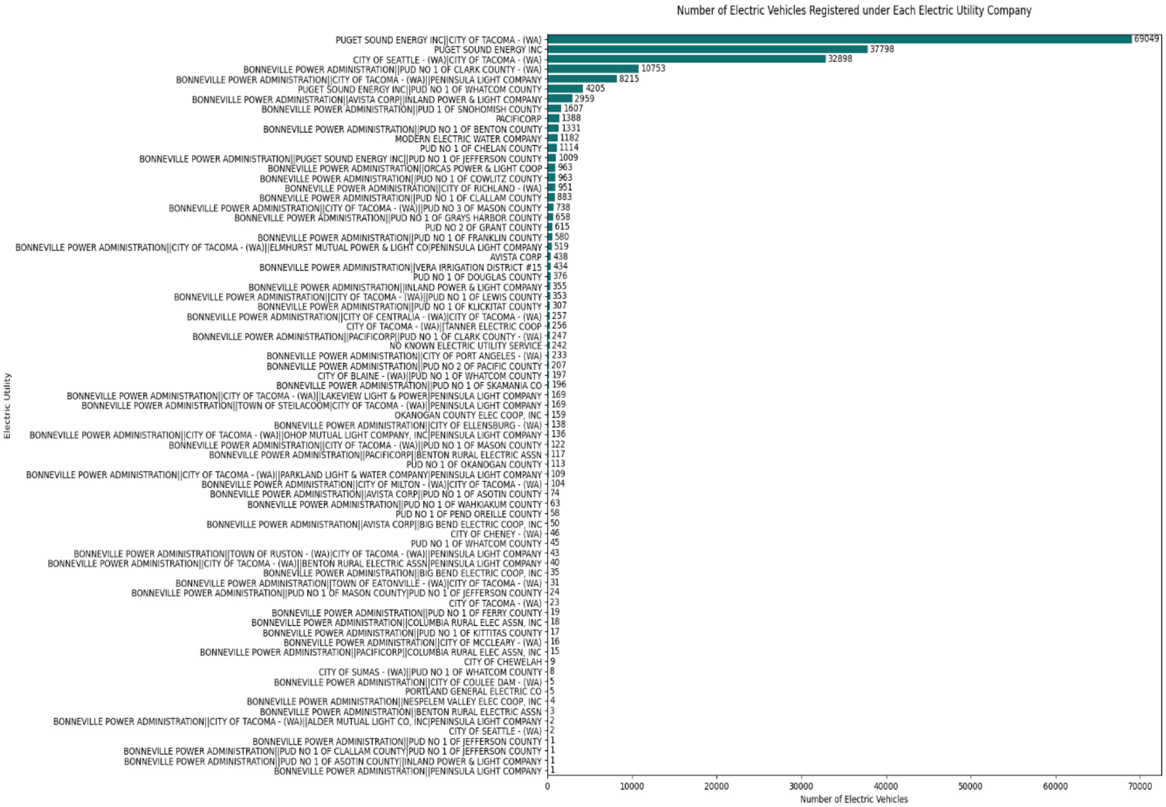
Univariate Analysis:

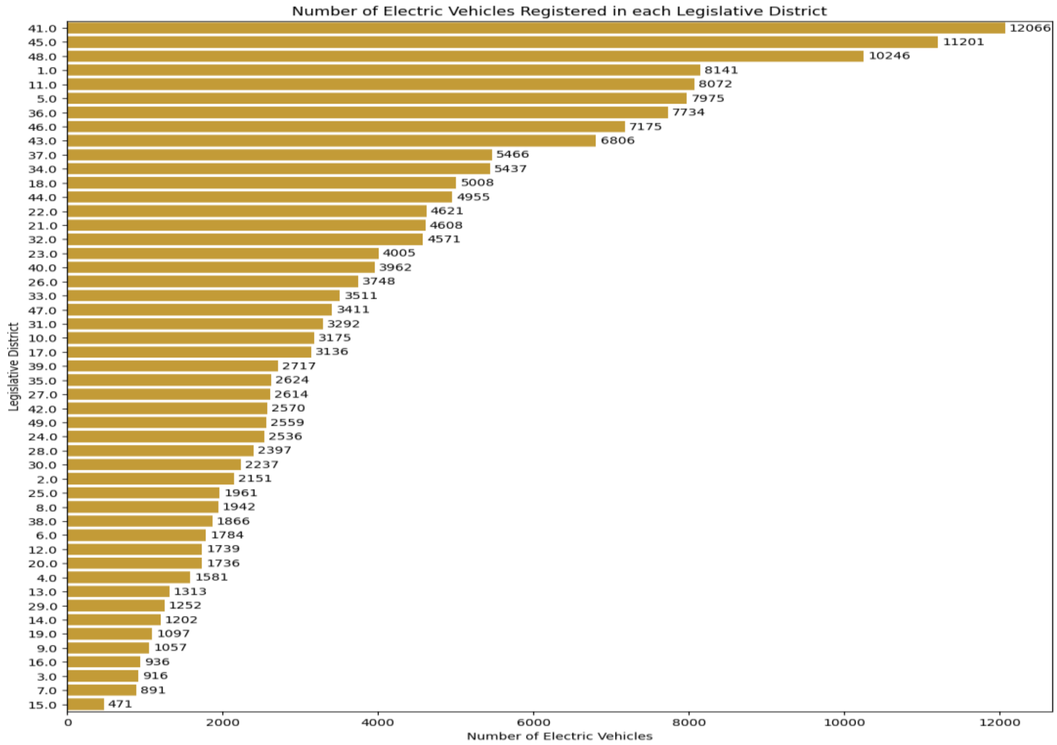



Distribution of Model Years
Count
60000
50000
40000
30000
20000
10000
0
1
1
5
7
2
1
20
24
757
1589
4374
3499
4830
5521
8602
14344
10904
11990
19254
28066
59229
13451
1997
1998
1999
2000
2002
2003
2008
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Model Year

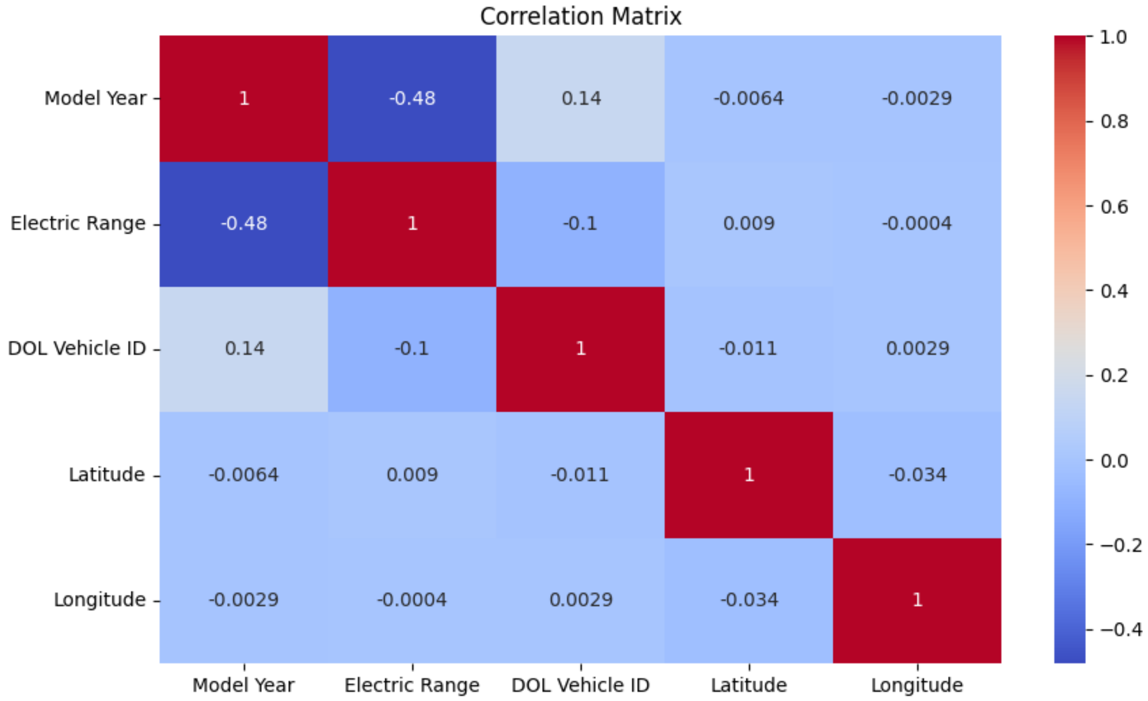

Number of Electric Vehicles Registered in each County
County
King
97012
Snohomish
22086
Pierce
14542
Clark
11005
Thurston
6779
Kitsap
6145
Spokane
4856
Whatcom
4455
Benton
2282
Skagit
2028
Island
1982
Clallam
1120
Chelan
1114
Yakima
1070
Jefferson
1035
San Juan
967
Cowlitz
963
Mason
865
Lewis
821
Grays Harbor
675
Kittitas
668
Franklin
617
Grant
615
Walla Walla
453
Douglas
378
Whitman
363
Klickitat
307
Okanogan
279
Stevens
235
Pacific
207
Skamania
196
Asotin
75
Wahkiakum
63
Pend Oreille
58
Adams
54
Lincoln
52
Ferry
32
Columbia
14
Garfield
3
0
20000
40000
60000
80000
100000
Number of Electric Vehicles


Top 50 Cities by Number of Electric Vehicles Registered
City
Seattle 30873
Bellevue 9370
Redmond 6747
Vancouver 6531
Bothell 6151
Kirkland 5608
Sammamish 5494
Renton 5399
Olympia 4501
Tacoma 3891
Bellingham 3073
Tukwila 3038
Kent 2967
Issaquah 2889
Spokane 2871
Lynnwood 2865
Everett 2636
Woodinville 2463
Mercer Island 2436
Snohomish 2185
Shoreline 2058
Gig Harbor 2053
Auburn 1986
Edmonds 1955
Bainbridge Island 1819
Camas 1692
Maple Valley 1548
Federal Way 1519
Lake Stevens 1453
Puyallup 1407
Bremerton 1385
Marysville 1303
Seatac 1241
Port Orchard 1143
Kenmore 1074
Bonney Lake 1062
Ridgefield 992
Lacey 992
Richland 973
Newcastle 946
Burien 938
Kennewick 934
Snoqualmie 913
Monroe 812
Mukilteo 805
Lakewood 790
North Bend 771
Spokane Valley 760
Poulsbo 739
Mill Creek 721
0
5000
10000
15000
20000
25000
30000
Number of Electric Vehicles


Number of Electric Vehicles Registered under Each Electric Utility Company
Electric Utility
PUGET SOUND ENERGY INC||CITY OF TACOMA - (WA) 69049
PUGET SOUND ENERGY INC 37798
CITY OF SEATTLE - (WA)|CITY OF TACOMA - (WA) 32898
BONNEVILLE POWER ADMINISTRATION||PUD NO 1 OF CLARK COUNTY - (WA) 10753
BONNEVILLE POWER ADMINISTRATION||CITY OF TACOMA - (WA)||PENINSULA LIGHT COMPANY 8215
PUGET SOUND ENERGY INC||PUD NO 1 OF WHATCOM COUNTY 4205
BONNEVILLE POWER ADMINISTRATION||AVISTA CORP||INLAND POWER & LIGHT COMPANY 2959
BONNEVILLE POWER ADMINISTRATION||PUD 1 OF SNOHOMISH COUNTY 1607
PACIFICORP 1388
BONNEVILLE POWER ADMINISTRATION||PUD NO 1 OF BENTON COUNTY 1331
MODERN ELECTRIC WATER COMPANY 1182
PUD NO 1 OF CHELAN COUNTY 1114
BONNEVILLE POWER ADMINISTRATION||PUGET SOUND ENERGY INC||PUD NO 1 OF JEFFERSON COUNTY 1009
BONNEVILLE POWER ADMINISTRATION||ORCAS POWER & LIGHT COOP 963
BONNEVILLE POWER ADMINISTRATION||PUD NO 1 OF COWLITZ COUNTY 963
BONNEVILLE POWER ADMINISTRATION||CITY OF RICHLAND - (WA) 951
BONNEVILLE POWER ADMINISTRATION||PUD NO 1 OF CLALLAM COUNTY 883
BONNEVILLE POWER ADMINISTRATION||CITY OF TACOMA - (WA)||PUD NO 3 OF MASON COUNTY 738
BONNEVILLE POWER ADMINISTRATION||PUD NO 1 OF GRAYS HARBOR COUNTY 658
PUD NO 2 OF GRANT COUNTY 615
BONNEVILLE POWER ADMINISTRATION||PUD NO 1 OF FRANKLIN COUNTY 580
BONNEVILLE POWER ADMINISTRATION||CITY OF TACOMA - (WA)||ELMHURST MUTUAL POWER & LIGHT CO|PENINSULA LIGHT COMPANY 519
AVISTA CORP 438
BONNEVILLE POWER ADMINISTRATION||VERA IRRIGATION DISTRICT #15 434
PUD NO 1 OF DOUGLAS COUNTY 376
BONNEVILLE POWER ADMINISTRATION||INLAND POWER & LIGHT COMPANY 355
BONNEVILLE POWER ADMINISTRATION||CITY OF TACOMA - (WA)||PUD NO 1 OF LEWIS COUNTY 353
BONNEVILLE POWER ADMINISTRATION||PUD NO 1 OF KLICKITAT COUNTY 307
BONNEVILLE POWER ADMINISTRATION||CITY OF CENTRALIA - (WA)|CITY OF TACOMA - (WA) 257
CITY OF TACOMA - (WA)||TANNER ELECTRIC COOP 256
BONNEVILLE POWER ADMINISTRATION||PACIFICORP||PUD NO 1 OF CLARK COUNTY - (WA) 247
NO KNOWN ELECTRIC UTILITY SERVICE 242
BONNEVILLE POWER ADMINISTRATION||CITY OF PORT ANGELES - (WA) 233
BONNEVILLE POWER ADMINISTRATION||PUD NO 2 OF PACIFIC COUNTY 207
CITY OF BLAINE - (WA)||PUD NO 1 OF WHATCOM COUNTY 197
BONNEVILLE POWER ADMINISTRATION||PUD NO 1 OF SKAMANIA CO 196
BONNEVILLE POWER ADMINISTRATION||CITY OF TACOMA - (WA)||LAKEVIEW LIGHT & POWER|PENINSULA LIGHT COMPANY 169
BONNEVILLE POWER ADMINISTRATION||TOWN OF STEILACOOM|CITY OF TACOMA - (WA)||PENINSULA LIGHT COMPANY 169
OKANOGAN COUNTY ELEC COOP, INC 159
BONNEVILLE POWER ADMINISTRATION||CITY OF ELLENSBURG - (WA) 138
BONNEVILLE POWER ADMINISTRATION||CITY OF TACOMA - (WA)||OHOP MUTUAL LIGHT COMPANY, INC|PENINSULA LIGHT COMPANY 136
BONNEVILLE POWER ADMINISTRATION||CITY OF TACOMA - (WA)||PUD NO 1 OF MASON COUNTY 122
BONNEVILLE POWER ADMINISTRATION||PACIFICORP||BENTON RURAL ELECTRIC ASSN 117
PUD NO 1 OF OKANOGAN COUNTY 113
BONNEVILLE POWER ADMINISTRATION||CITY OF TACOMA - (WA)||PARKLAND LIGHT & WATER COMPANY|PENINSULA LIGHT COMPANY 109
BONNEVILLE POWER ADMINISTRATION||CITY OF MILTON - (WA)|CITY OF TACOMA - (WA) 104
BONNEVILLE POWER ADMINISTRATION||AVISTA CORP||PUD NO 1 OF ASOTIN COUNTY 74
BONNEVILLE POWER ADMINISTRATION||PUD NO 1 OF WAHKIAKUM COUNTY 63
PUD NO 1 OF PEND OREILLE COUNTY 58
BONNEVILLE POWER ADMINISTRATION||AVISTA CORP||BIG BEND ELECTRIC COOP, INC 50
CITY OF CHENEY - (WA) 46
PUD NO 1 OF WHATCOM COUNTY 45
BONNEVILLE POWER ADMINISTRATION||TOWN OF RUSTON - (WA)|CITY OF TACOMA - (WA)||PENINSULA LIGHT COMPANY 43
BONNEVILLE POWER ADMINISTRATION||CITY OF TACOMA - (WA)||BENTON RURAL ELECTRIC ASSN|PENINSULA LIGHT COMPANY 40
BONNEVILLE POWER ADMINISTRATION||BIG BEND ELECTRIC COOP, INC 35
BONNEVILLE POWER ADMINISTRATION||TOWN OF EATONVILLE - (WA)|CITY OF TACOMA - (WA) 31
BONNEVILLE POWER ADMINISTRATION||PUD NO 1 OF MASON COUNTY|PUD NO 1 OF JEFFERSON COUNTY 24
CITY OF TACOMA - (WA) 23
BONNEVILLE POWER ADMINISTRATION||PUD NO 1 OF FERRY COUNTY 19
BONNEVILLE POWER ADMINISTRATION||COLUMBIA RURAL ELEC ASSN, INC 18
BONNEVILLE POWER ADMINISTRATION||PUD NO 1 OF KITTITAS COUNTY 17
BONNEVILLE POWER ADMINISTRATION||CITY OF MCCLEARY - (WA) 16
BONNEVILLE POWER ADMINISTRATION||PACIFICORP||COLUMBIA RURAL ELEC ASSN, INC 15
CITY OF CHEWELAH 9
CITY OF SUMAS - (WA)||PUD NO 1 OF WHATCOM COUNTY 8
BONNEVILLE POWER ADMINISTRATION||CITY OF COULEE DAM - (WA) 5
PORTLAND GENERAL ELECTRIC CO 5
BONNEVILLE POWER ADMINISTRATION||NESPELEM VALLEY ELEC COOP, INC 4
BONNEVILLE POWER ADMINISTRATION||BENTON RURAL ELECTRIC ASSN 3
BONNEVILLE POWER ADMINISTRATION||CITY OF TACOMA - (WA)||ALDER MUTUAL LIGHT CO, INC|PENINSULA LIGHT COMPANY 2
CITY OF SEATTLE - (WA) 2
BONNEVILLE POWER ADMINISTRATION||PUD NO 1 OF JEFFERSON COUNTY 1
BONNEVILLE POWER ADMINISTRATION||PUD NO 1 OF CLALLAM COUNTY|PUD NO 1 OF JEFFERSON COUNTY 1
BONNEVILLE POWER ADMINISTRATION||PUD NO 1 OF ASOTIN COUNTY||INLAND POWER & LIGHT COMPANY 1
BONNEVILLE POWER ADMINISTRATION||PENINSULA LIGHT COMPANY 1
0
10000
20000
30000
40000
50000
60000
70000
Number of Electric Vehicles



Bivariate Analysis:



Approach

The approach adopted for the three research questions are as follows:

Cleaning of Data:

1. After carefully summarising the data, 398 observations having the registration details of states other than Washington were removed.’
2. The vehicle location variable having geographical coordinates of the registered vehicle location, has been cleaned and split into two named Latitude and Longitude.
3. As Postal Code, Legislative District and 2020 Census Tract are categorical variables, they have been converted to string data type.
4. Base MSRP variable was removed as 98% of its data have the value 0.
5. No duplicate values were found in the data.

<u>Research Question No.1</u>

Research Question	Definition	Approach
Based on data collected from 2015 to 2023, which covers approximately 5% of Washington State's population, what are the primary factors influencing the decision to purchase BEVs and PHEVs in the state?	This question aims to identify the factors that influence the decision to purchase Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) in Washington State.	STEP I Analyze and visualize the total count of Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) by their model year to provide an insight into the trend of electric vehicle adoption over time. Reference:
		STEP II Using correlation to understand the dataset's structure and relationships before building predictive models, ensuring that the most

		relevant features are selected for further analysis. Reference: Da Poian, V., Theiling, B., Clough, L., McKinney, B., Major, J., Chen, J., & Hörst, S. (2023). Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry. *Frontiers in Astronomy and Space Sciences, 10*.
		STEP III Prepare the data for machine learning, perform feature selection, and identify the most important features for predicting whether an electric vehicle is a Battery Electric Vehicle (BEV) or a Plug-in Hybrid Electric Vehicle (PHEV). **Variables:** Dependent Variable – Electric Vehicle Type (BEV or PHEV) Independent Variable- County, City, Legislative District and Electric Utility company **Testing:** Split the data into 80% training and 20% testing sets

		Reference: Alanazi, R., et al. (2022). Identification and prediction of chronic diseases using machine learning approaches. Journal of Big Data.
		STEP IV Perform feature selection, handle class imbalance, and prepare the data for machine learning model training Reference: Qasim, H. M., Ata, O., Ansari, M. A., Alomary, M. N., Alghamdi, S., & Almehmadi, M. (2021). Hybrid feature selection framework for the Parkinson imbalanced dataset prediction problem. *Medicina, 57*(11), 1217.
		STEP V Perform hyperparameter tuning, train, and evaluate Logistic Regression and Random Forest models on the balanced training set, and then visualize and interpret the results

		Reference: Imani, M., & Arabnia, H. R. (2023). Hyperparameter optimization and combined data sampling techniques in machine learning for customer churn prediction: A comparative analysis. *Technologies, 11*(6), 167

Research Question No.2

Research Question	Definition	Approach
Can we use machine learning to predict the number of new electric vehicle adopters over the next ten years in	This question explores the feasibility of using machine learning models to forecast the adoption of electric vehicles based on	**Feature Engineering:** Create features like the number of BEVs and PHEVs registered per year, city-wise, county-wise, legislative district-wise and electric utility-wise. Reference: Alanazi, R., et al. (2022). Identification and prediction of chronic

Washington State, based on detailed data about residents' geographic locations and the availability of infrastructural facilities, such as electric utility stations?	geographic and infrastructure related data.	diseases using machine learning approaches. Journal of Big Data.
		Variables: Dependent: Number of EVs registered (Future EV Adoption) Independent: County, City, Electric Vehicle Type, Model Year, Electric Utility Company, Legislative District
		Validating the variables: Descriptive Statistics and Visualization: Summary Statistics: Calculate descriptive statistics mean, median, standard deviation, min, max) for each new variable to ensure they are within expected ranges. Distribution Analysis: Use histograms or box plots to visualize the distribution of each variable and identify any outliers or unusual patterns. Reference: Dawson, R. J. (2011). How significant is a boxplot outlier? Journal of Statistics Education, 19(2). Statistical Validation: Hypothesis Testing: Conduct statistical tests (e.g., t-tests, ANOVA) to determine if there are significant differences in EV adoption rates across different geographic or infrastructural categories. Reference:

		Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. Behavioral Ecology, 17(4), 688-690. Variance Analysis: Analyze the variance explained by the new variables to ensure they add meaningful information to the model. Reference: Snijders, T. A., & Bosker, R. J. (1999). Multilevel analysis: An introduction to basic and advanced multilevel modeling. Sociological Methods & Research, 28(2), 201-223.
		Machine Learning: Linear Regression for trend analysis. Random forest for Prediction Reference: Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis. John Wiley & Sons. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.
		Training and Testing: Split the dataset into training (80%) and testing (20%). Reference: Alanazi, R., et al. (2022). Identification and prediction of chronic diseases using machine learning approaches. Journal of Big Data.

		Model Evaluation: Use mean squared error (MSE) and R-squared values to evaluate regression models. Reference: Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?. Geoscientific Model Development Discussions, 7(1), 1525-1534.

Research Question No.3

Research Question	Definition	Approach
How do geographic location and local infrastructure (electric utilities) impact EV adoption in Washington?	This question investigates the influence of geographic factors and the availability of local infrastructure, like electric utilities, on the adoption rates of electric vehicles.	**STEP I** Creating New Features for EV Counts by Geographic and Infrastructure Categories
		STEP II **Validating the variables:** Descriptive Statistics and Visualization: Summary Statistics: Calculate descriptive statistics mean, median, standard deviation, min, max) for

		each new variable to ensure they are within expected ranges. Distribution Analysis: Use bar plots or box plots to visualize the distribution of each variable and identify any outliers or unusual patterns. Reference: Dawson, R. J. (2011). How significant is a boxplot outlier? Journal of Statistics Education, 19(2).
		STEP III Geographic Validation: Mapping: Visualize the geographic distribution of the new variables using maps (e.g., heat maps or choropleth maps) to check for expected spatial patterns. Cross-sectional Consistency: Ensure consistency across different geographic levels (e.g., city-wise data should aggregate correctly to county-wise data).
		STEP IV **Machine Learning:** Multiple Linear Regression to account for geographic dependencies. Reference: Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction

		to linear regression analysis. John Wiley & Sons
		STEP V K-means Clustering to identify patterns and group similar areas. Reference: Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 881-892.
		STEP VI Apriori Algorithm to identify associations between geographic location, electric utilities, and EV adoption patterns. Reference: Xiaodong Wu, Y., & Zeng, Y. (2019). Using Apriori Algorithm on Students’ Performance Data for Association Rules Mining. *IEEE International Conference on Data*

		Science and Advanced Analytics (DSAA).
		STEP VII Perform ANOVA to compare means across different categories of geographic locations.

<u>Methodology for Research Question No.3</u>

K-means Clustering:

- **Reason for Choosing K-means:** K-means is ideal for continuous data and efficiently handles large datasets. It helps us identify patterns and group similar areas based on EV adoption rates and other features.
- **Comparison with K-modes and K-medians:** While K-modes is better suited for categorical data, and K-medians is robust against outliers but computationally intensive, K-means fits best since our features (e.g., latitude, longitude) are mostly continuous.
    - **Reference:** Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 881-892.

Apriori Algorithm:

- **Categorical Variable Handling:** The Apriori algorithm works with categorical data, so we’ll convert continuous variables like, latitude, and longitude into categorical bins.
- **Compatibility with Multiple Linear Regression:** While Apriori helps find associations in categorical data, multiple linear regression will quantify relationships between dependent and independent variables. This mixed-method approach offers a comprehensive analysis.
    - **Reference:** Xiaodong Wu, Y., & Zeng, Y. (2019). Using Apriori Algorithm on Students’ Performance Data for Association Rules Mining. *IEEE International Conference on Data Science and Advanced Analytics (DSAA).*
    -

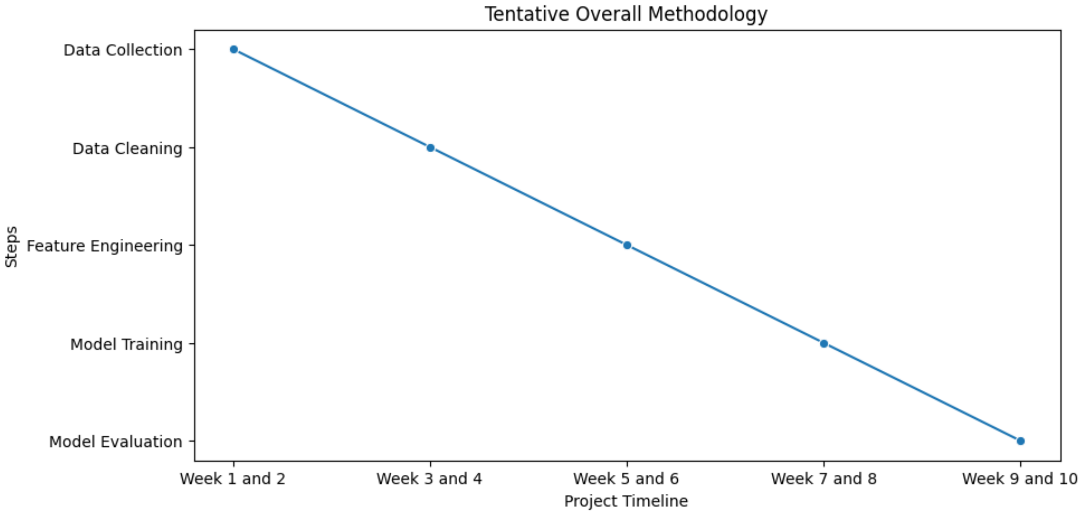
Evaluating Association Rules:

- **Metrics for Evaluation:** We'll use metrics like support, confidence, and lift to evaluate the strength of the association rules generated by the Apriori algorithm.
- **Interpretation:** High values for support, confidence, and lift indicate strong and meaningful associations. Visual tools like association rule graphs will help us interpret and validate these findings.
    - **Reference:** Hahsler, M., Grun, B., Hornik, K., & Buchta, C. (2005). Introduction to arules—a computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, *14*(15), 1-25.

By using these methods, we can effectively analyze the data to uncover patterns and relationships that impact EV adoption, providing a comprehensive understanding of the factors at play.

Tentative overall Methodology

This structured timeline ensures a systematic approach to the project, covering all critical stages from data collection to model evaluation. Each phase is allocated sufficient time to thoroughly address its objectives, ensuring the final model is robust and reliable. This detailed approach will provide a comprehensive understanding of EV adoption trends in Washington State, supporting future strategic planning and policy-making efforts.



References

1. Bursac, Z., Gauss, C. H., Williams, D. K., & Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. Source Code for Biology and Medicine, 3(1), 17. https://scfbm.biomedcentral.com/articles/10.1186/1751-0473-3-17

2. . Nowling, R. J. (2015). Categorical Variable Encoding and Feature Importance Bias with Random Forests. https://rnowling.github.io/machine/learning/2015/08/10/random-forest-bias.html

3. Alanazi, R., et al. (2022). Identification and prediction of chronic diseases using machine learning approaches. Journal of Big Data. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8896926/

4. Da Poian, V., Theiling, B., Clough, L., McKinney, B., Major, J., Chen, J., & Hörst, S. (2023). Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry. *Frontiers in Astronomy and Space Sciences*, *10*. https://www.frontiersin.org/journals/astronomy-and-space-sciences/articles/10.3389/fspas.2023.1134141/full

5. Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. Behavioral Ecology, 17(4), 688-690. Retreived from https://academic.oup.com/beheco/article/17/4/688/215960

6. Snijders, T. A., & Bosker, R. J. (1999). Multilevel analysis: An introduction to basic and advanced multilevel modeling. Sociological Methods & Research, 28(2), 201-223. Reteived from https://www.researchgate.net/publication/44827177_Multilevel_Analysis_An_Introduction_to_Basic_and_Advanced_Multilevel_Modeling

7. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis. John Wiley & Sons. Retrieved from https://statanaly.com/wp-content/uploads/2023/05/IntroductiontoLinearRegressionAnalysisbyDouglasC.MontgomeryElizabethA.PeckG_.GeoffreyViningz-lib.org_.pdf

8. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22. Retreived from https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf

9. Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?. Geoscientific Model Development Discussions, 7(1), 1525-1534. Retrieved from https://gmd.copernicus.org/articles/7/1247/2014/gmd-7-1247-2014.pdf

10. Schneider, C., Dryhurst, S., Kerr, J., Freeman, A., Recchia, G., Spiegelhalter, D., & van der Linden, S. (2021). COVID-19 risk perception: A longitudinal analysis of its predictors and associations with health protective behaviours in the United Kingdom. *Journal of Risk Research*, 24. Retrieved from https://www.researchgate.net/publication/350277760_COVID-19_risk_perception_a_longitudinal_analysis_of_its_predictors_and_associations_with_health_protective_behaviours_in_the_United_Kingdom

11. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(7), 881-892. Retrieved from https://ieeexplore.ieee.org/document/1017616

12. Xiaodong Wu, Y., & Zeng, Y. (2019). Using Apriori Algorithm on Students' Performance Data for Association Rules Mining. *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. Retrieved from https://www.semanticscholar.org/paper/Using-Apriori-Algorithm-on-Students%E2%80%99-Performance-Wu-Zeng/4a5c50c149e77f9240e6e6e4b58e4b2eb939c849

13. Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection—a review and recommendations for the practicing statistician. *Biometrical Journal, 60*(3), 431-449. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1002/bimj.201700067

14. Wood, J. D., Dykes, J., Slingsby, A., & Clarke, K. (2007). Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geovisualization mashup. *IEEE Transactions on Visualization and Computer Graphics, 13*(6), 1176-1183. Retrieved from https://www.researchgate.net/publication/5878327_Interactive_Visual_Exploration_of_a_Large_Spatio-temporal_Dataset_Reflections_on_a_Geovisualization_Mashup

15. Qasim, H. M., Ata, O., Ansari, M. A., Alomary, M. N., Alghamdi, S., & Almehmadi, M. (2021). Hybrid feature selection framework for the Parkinson imbalanced dataset prediction problem. *Medicina*, *57*(11), 1217. Retrieved from https://www.mdpi.com/1648-9144/57/11/1217

16. 16. Imani, M., & Arabnia, H. R. (2023). Hyperparameter optimization and combined data sampling techniques in machine learning for customer churn prediction: A comparative analysis. *Technologies*, *11*(6), 167 https://www.mdpi.com/2227-7080/11/6/167