Exploring Electric Vehicle Adoption Trends in Washington State

Student: Arunkumar Cherungot-Meppadathkalam

Student No.: 501277792

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

Supervisor: Tamer Abdou

Date of Submission: 21st July 2024

Ryerson University

Table of Contents

1. Abstract

1.1 **Background:**

Electric vehicles (EVs) are becoming popular due to concerns about the environment and sustainable transportation. In Washington State, it's important to understand why people are choosing EVs and what factors are driving this choice. This project aims to look at data about EV ownership to figure out why people in Washington are choosing electric cars and what might happen in the future.

1.2 **Problem Statement/ Research Question:**

We're trying to answer three main questions:

1. 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?

2. Can we use machine learning to predict the number of new electric vehicle adopters over the next ten years in Washington State, based on detailed data about residents' geographic locations and the availability of infrastructural facilities, such as electric utility stations?

3. How do geographic location and local infrastructure (electric utilities) impact EV adoption in Washington?

1.3 **Data:**

Dataset Link: https://catalog.data.gov/dataset/electric-vehicle-population-data
License Information: https://opendatacommons.org/licenses/odbl/1-0/

The dataset utilized in this study comprises registered Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) through the Washington State Department of Licensing (DOL). It includes detailed information on vehicle types, ownership demographics, geographic distribution and additional attributes such as vehicle make, model, year, and electric range.

1.4 **Techniques and Tools:**

To address our research questions, we will employ Python programming language. Classification and regression algorithms will be utilized to analyze the data and identify patterns in EV adoption. Additionally, predictive analytics techniques, including time-series analysis and pattern mining, will be applied to forecast future trends in EV adoption rates. Python's versatility and robust ecosystem of data analysis tools will facilitate data preprocessing, model development and visualization of results.

2. Literature Review

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.

2.1 **GitHub Repository**

The source code files for this project can be accessed at the following GitHub repository: https://github.com/acherungotTMU/TMU_Capstone

2.2 **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.

2.3 **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.

2.4 **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.

2.5 **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:

a) 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

b) 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

c) 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

d) 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

e) 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

f) 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

g) 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

h) 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

i) 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

j) 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

k) 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

l) 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

m) 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

n) 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

o) 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

p) 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.

2.6 Where Does my Work Fit in With What Has Done 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.

2.7 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.

2.8 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.

i. **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).

2.9 Approach

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

2.9.1 Cleaning of Data:

Preparing the data is the initial step in our research process. To clean and organize the data, we undertook the following actions:

- After carefully summarising the data, 398 observations having the registration details of states other than Washington were removed.'
- The vehicle location variable having geographical coordinates of the registered vehicle location, has been cleaned and split into two named Latitude and Longitude.
- As Postal Code, Legislative District and 2020 Census Tract are categorical variables, they have been converted to string data type.

- Base MSRP variable was removed as 98% of its data have the value 0.
- No duplicate values were found in the data.

2.9.2 **Research Question No.1**

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.	<u>STEP I</u> 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. <u>STEP II</u> 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. Reference : Alanazi, R., et al. (2022). Identification and prediction of chronic diseases using machine learning
		STEP IV Perform feature selection, handle class imbalance, and prepare the data for machine learning model training. Used

		Random Forest Classifier and Recursive Feature Elimination (RFE). 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.
		<u>STEP V</u> 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

2.9.3 **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 Washington State, based on detailed data about residents' geographic locations and the availability of infrastructural facilities, such as electric utility stations?	This question explores the feasibility of using machine learning models to forecast the adoption of electric vehicles based on geographic and infrastructure related data.	**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 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.

2.9.4 **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)*.

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

a. **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.

b. **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)*.

c. **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.

3. Research and Analysis

3.1 Research Question No.1

Based on the 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?

3.1.1 Main Contribution of the Work

This research provides an in-depth analysis of factors influencing BEV and PHEV adoption in Washington State, employing advanced machine learning techniques and addressing gaps in previous studies by focusing on the latest trends and comprehensive feature analysis.

3.1.2 Methodology

i. Data Preprocessing

a. Focused on data from years 2015 to 2023 to ensure relevance to current electric vehicle adoption trends.
b. Encoded 'Electric Vehicle Type' as 'EV_Type_Encoded' where BEV is 1 and PHEV is 0.
c. Performed one-hot encoding on categorical variables such as County, City, Make, Model, Clean Alternative Fuel Vehicle (CAFV) Eligibility, Electric Utility, Legislative District, and 2020 Census Tract. One-hot encoding is performed to transform categorical variables into a binary format that can be readily used by machine learning algorithms.

ii. Correlation Analysis

a. Correlation Matrix: Calculated the correlation matrix to understand the relationships between numerical features and the target variable (EV_Type_Encoded).
b. Feature Selection Based on Correlation: Identified key features correlated with the target variable to guide further feature selection processes.

iii. Feature Selection

a. **Random Forest Classifier:**

- Trained with 100 trees to determine feature importance. Training the model with 100 trees helps to ensure feature importance scores are robust and stable.
- Identified the top 50 features based on importance scores.

b. **Recursive Feature Elimination (RFE):**

- Used Logistic Regression as the model to select the top 20 features. Logistic Regression provides clear coefficients that indicate the direction and magnitude of the relationship between features and the target variable, making it easy to interpret which features are important.
- Performed SMOTE to balance the training data.

3.1.3 Model Training and Evaluation

a. **Logistic Regression:**

- Performed hyperparameter tuning using GridSearchCV to find the best parameters, trained the model on the balanced dataset, and evaluated its performance using accuracy, confusion matrix, and cross-validation scores.

b. **Random Forest Classifier:**

- Trained on the balanced dataset, evaluated using accuracy, confusion matrix, feature importances, and cross-validation scores.

3.1.4 Findings and Interpretation

a. **Correlation Analysis:**

- Key features correlated with EV_Type_Encoded include Model Year and Electric Range

- Key features correlated with EV_Type_Encoded include Model Year and Electric Range.

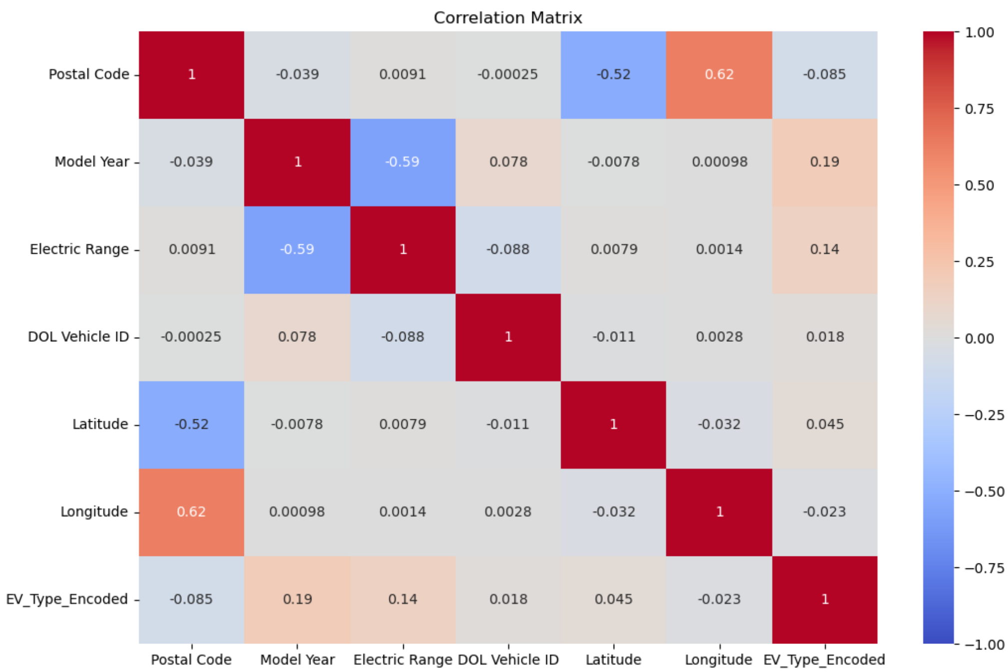


Fig 3.1.1 Correlation Matrix between numerical features and EV_Type (EV_Type_Encoded

Correlation with target variable:

Postal Code -0.084690
Model Year 0.189422
Electric Range 0.143285
DOL Vehicle ID 0.017626
Latitude 0.045268
Longitude -0.023110
Name: EV_Type_Encoded, dtype: float64

Interpretation of Correlation Matrix:

Model Year has the highest positive correlation (0.19) with the target variable. This suggests that newer model years are somewhat associated with BEVs compared to PHEVs.

Electric Range also has a moderate positive correlation (0.143539), indicating that BEVs might have a higher electric range compared to PHEVs.

b. **Feature Importance:**

- Significant features include Electric Range, Clean Alternative Fuel Vehicle (CAFV) Eligibility, Make, and Model.

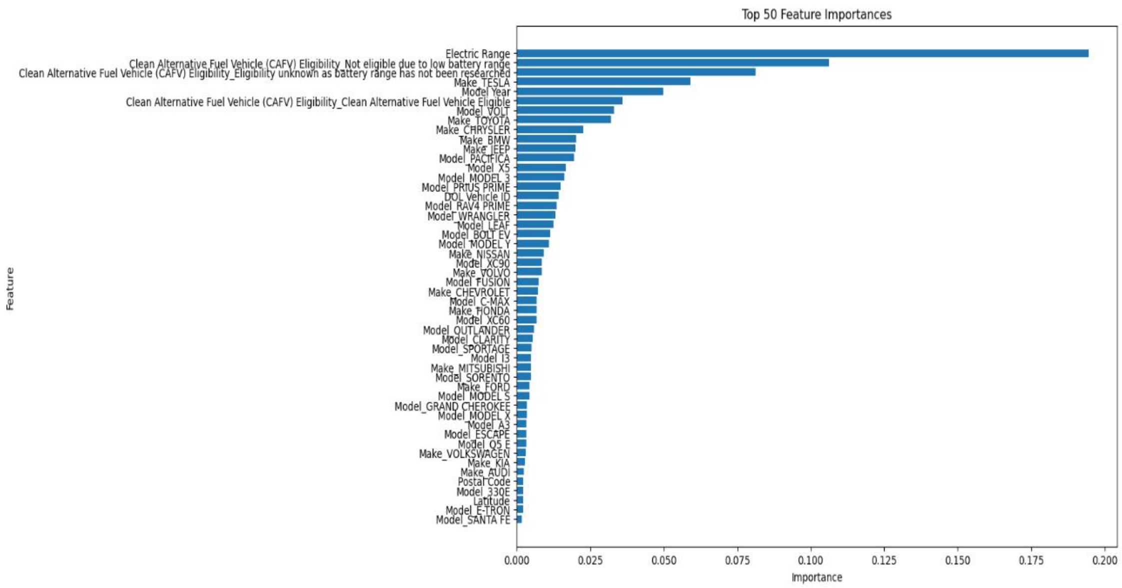


Fig 3.1.2 Top 50 features using Random Forest

Top 20 features by Random Forest:	
Feature	**Importance**
Electric Range	0.1945
Clean Alternative Fuel Vehicle (CAFV) Eligibility_Not elegible due to low battery range	0.106242
Clean Alternative Fuel Vehicle (CAFV) Eligibility unknown as battery range has not been researched	0.081302
Make_TESLA	0.059041
Model Year	0.049786
Clean Alternative Fuel Vehicle (CAFV) Eligibility_Clean Alternative Fuel Vehicle Eligibility	0.035982

Model_VOLT	0.033117
Make_TOYOTA	0.032135
Make_CHRYSLER	0.022602
Make_BMW	0.020247
Make_JEEP	0.019935
Model_PACIFICA	0.01951
Model_X5	0.016716
Model_MODEL 3	0.016316
Model_PRIUS PRIME	0.014884
DOL Vehicle ID	0.014242
Model_RAV4 PRIME	0.013485
Model_WRANGLER	0.013039
Model_LEAF	0.012394
Model_BOLT EV	0.011453

Table 3.1.1 Top 20 features and their corresponding importances using Random forest

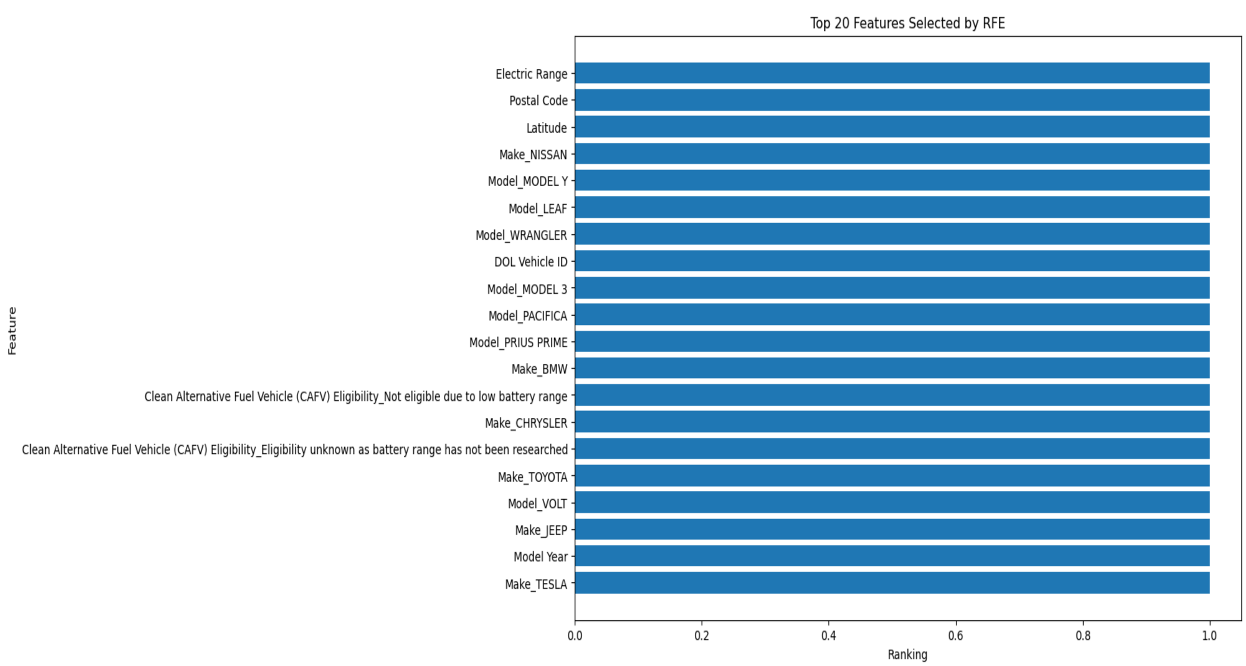


Fig 3.1.3 Top 20 features using RFE-Logistic Regression

Top	20 Features by RFE:
Feature	Ranking
Electric Range	1
Postal Code	1
Latitude	1
Make_NISSAN	1

Model_MODEL Y	1
Model_LEAF	1
Model_WRANGLER	1
DOL Vehicle ID	1
Model_MODEL 3	1
Model_PACIFICA	1
Model_PRIUS PRIME	1
Make_BMW	1
Clean Alternative Fuel Vehicle (CAFV) Eligibility_Not eligible due to low battery range	1
Make_CHRYSLER	1
Clean Alternative Fuel Vehicle (CAFV) Eligibility unknown as battery range has not been researched	1
Make_TOYOTA	1
Model_VOLT	1
Make_JEEP	1
Model Year	1
Make_TESLA	1

Table 3.1.2 Top 20 features and their corresponding importances using RFE-Logistic Regression

Interpretation of the Top 20 Feature using Random Forest and RFE:

The output and the visualization provided show the top 50 features ranked by their importance scores as determined by the Random Forest classifier. Here are the key takeaways and interpretations from this output:

a. **Most Important Features:**

- Electric Range: This feature has the highest importance score, indicating that the electric range of a vehicle is a critical factor in determining whether it is a BEV or PHEV.
- CAFV Eligibility: Various categories of "Clean Alternative Fuel Vehicle (CAFV) Eligibility" are also highly important. This reflects the influence of policy and eligibility criteria on the type of electric vehicle.
- Make (Brand): The make of the vehicle, particularly brands like Tesla, Toyota, BMW, Chrysler, and Jeep, are significant. This suggests that certain brands are more associated with BEVs or PHEVs.
- Model Year: Newer models are more likely to be BEVs, as indicated by the importance of the model year.
- Specific Models: Specific vehicle models like Volt, Pacifica, RAV4 Prime, Prius Prime, Model 3, and others show significant importance, suggesting a strong association with the type of electric vehicle.

b. **Policy and Infrastructure:**

- Clean Alternative Fuel Vehicle (CAFV) Eligibility: The high importance of CAFV eligibility categories underscores the role of policies and incentives in influencing the adoption of BEVs and PHEVs.

c. **Geographical Influence:**

- 2020 Census Tract and City (e.g., City_Tukwila): These features, although lower in importance, still contribute to the model, indicating that location and possibly local policies or infrastructure can impact the type of electric vehicle adopted.

d. **Additional Factors:**

- Features like Latitude and Longitude have lower importance but are still considered, suggesting that geographical location plays a role, though less significant compared to other factors.

3.1.5 <u>**Model Performance:**</u>

- Logistic Regression with SMOTE achieved an accuracy of 99.65% with significant coefficients indicating the influence of features like CAFV Eligibility and Electric Range.

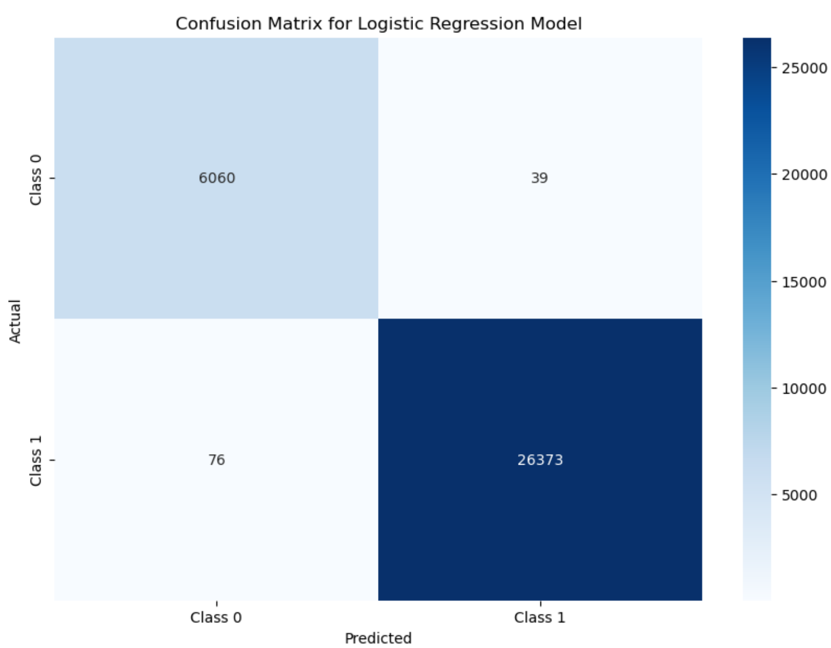


Fig 3.1.4 Confusion Matrix for Logistic Regression Model

- True Positives (Class BEV correctly predicted): 26373
- True Negatives (Class PHEV correctly predicted): 6060
- False Positives (Class BEV incorrectly predicted as Class PHEV): 39
- False Negatives (Class PHEV incorrectly predicted as Class BEV): 76

Mean Cross-Validation Score for Logistic Regression: 0.9976092215739291

- **Random Forest**: Random forest with SMOTE achieved an accuracy of 99.98%, with top features including Electric Range and CAFV Eligibility.

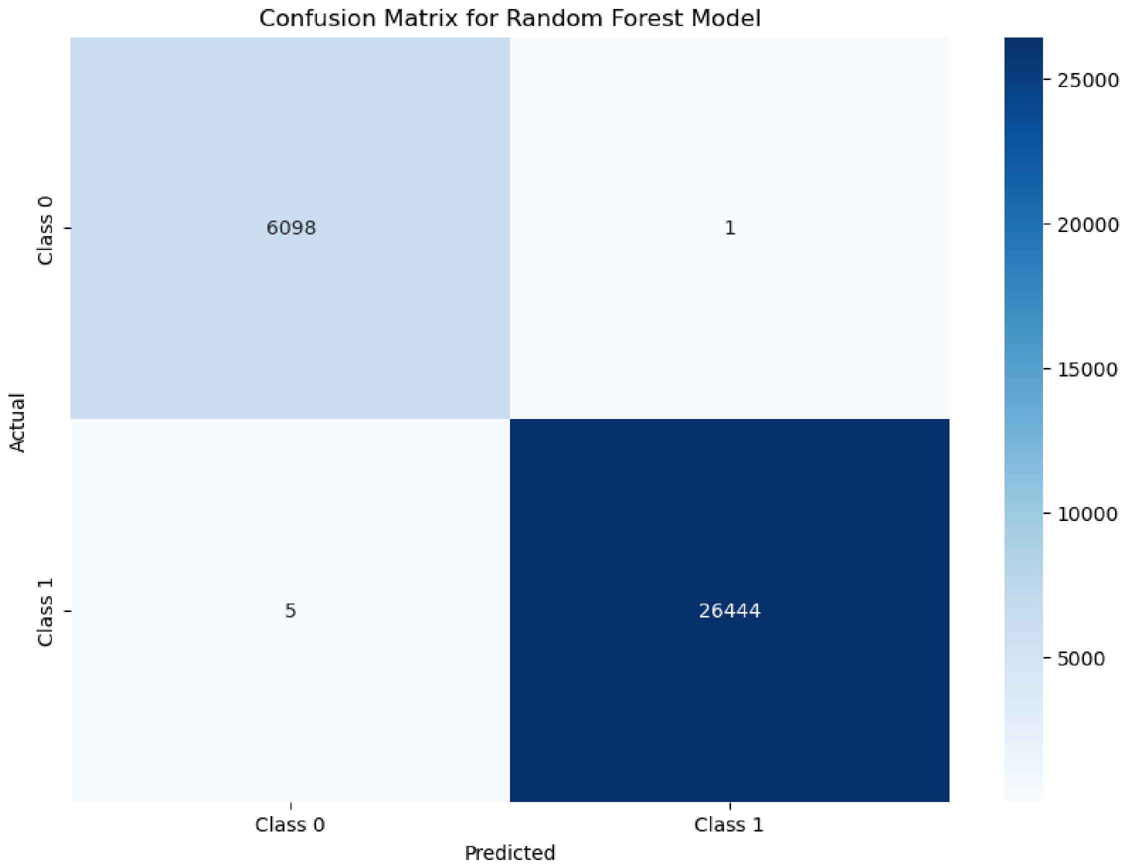


Fig 3.1.5 Confusion Matrix for Random Forest Model

- True Positives (Class BEV correctly predicted): 26444
- True Negatives (Class PHEV correctly predicted): 6098
- False Positives (Class BEV incorrectly predicted as Class PHEV): 1
- False Negatives (Class PHEV incorrectly predicted as Class BEV): 5

Mean Cross-Validation Score for Random Forest: 0.9999241022721881

3.1.6 Model Comparison:

- Both models show high and consistent cross-validation scores, indicating reliable performance across different subsets of the dataset. This suggests that the random forest model is slightly better at capturing the underlying patterns in the data.
- The random forest model slightly outperforms the logistic regression model based on the mean cross-validation scores (0.9999 vs. 0.9976).
- The high performance of both models suggests that they can be effectively used for predicting electric vehicle adoption. However, the slight edge in performance of the random forest model might make it a preferable choice in this specific context.

3.1.7 Shortcomings and Concluding Remarks

- **Limitations**:

  The dataset covers only 5% of Washington State's population, which may not be fully representative. Additionally, external factors like economic incentives and infrastructure development were not considered. Future research should aim to gather more comprehensive data that includes a larger portion of the population and considers additional external factors that can influence electric vehicle adoption. This will help in creating a more representative and accurate model.

- **Conclusion:**

  The study aimed to identify the primary factors influencing the decision to purchase Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) in Washington State, using data collected from 2015 to 2023, which covers approximately 5% of the state's population. The following key findings were derived from the analysis:

  a. **Electric Range:** This feature emerged as the most significant factor influencing the decision to purchase BEVs and PHEVs. Vehicles with higher electric ranges were more likely to be chosen, indicating that range anxiety is a critical consideration for potential buyers.
  b. **Eligibility for Clean Alternative Fuel Vehicle (CAFV) Incentives:** The eligibility for CAFV incentives was a substantial determinant in the decision-making process. Vehicles eligible for these incentives were more likely to be purchased, highlighting the importance of economic benefits and subsidies in promoting electric vehicle adoption.

c. **Make and Model of the Vehicle:** Specific brands and models, particularly those known for their electric vehicle offerings (e.g., Tesla, Nissan LEAF), were more popular among consumers. This suggests that brand reputation and model-specific features play a vital role in consumer preferences.

d. **Model Year:** Newer models were more favored, indicating that technological advancements and improvements in newer vehicles significantly impact purchasing decisions.

e. **Geographic Factors:** While geographic variables like postal code and legislative district showed some correlation with vehicle choice, they were less influential compared to the direct attributes of the vehicles.

3.1.8 References

- 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://doi.org/10.3389/fspas.2023.1134141
- Alanazi, R., et al. (2022). Identification and prediction of chronic diseases using machine learning approaches. Journal of Big Data.
- 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.
- 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

3.2 Research Question No.2

Can we use machine learning to predict the number of new electric vehicle adopters over the next ten years in Washington State, based on detailed data about residents' geographic locations and the availability of infrastructural facilities, such as electric utility stations?

3.2.1 **Main Contribution**

This study advances the understanding of electric vehicle (EV) adoption by leveraging machine learning models to predict future adoption rates based on geographic and infrastructural data. Unlike previous studies that may have focused on demographic or economic factors, this research uniquely integrates detailed geographic and infrastructural data, offering more granular insights into the spatial distribution of future EV adopters.

3.2.2 **Methodology**

a. Data Collection and Preprocessing

Data was collected from EV registration data from the Department of Licensing (DOL) Washington State

The dataset was preprocessed to handle missing values, encode categorical variables, and normalize numerical features.

b. Model Development

The machine learning models that were developed, were:

a. Linear Regression:

- Developed a linear regression model to capture the relationship between geographic, infrastructural data and EV adoption rates.
- Trained and tested using an 80-20 split of the dataset.

b. Random Forest:

- Developed a random forest model to capture complex, non-linear relationships between the variables.
- Trained and tested using an 80-20 split of the dataset.

3.2.3 **Data Analysis Procedures**

a. Statistical Analyses

The models were evaluated using statistical measures to compare their performance:

- Mean Squared Error (MSE)

- R-squared (R^2)

b. **Model Comparison**

The models were compared based on their performance metrics, training and testing times, and robustness to ensure that the best model is selected for predicting EV adoption rates.

3.2.4 Findings and Interpretation

a. **Model Effectiveness**

- Linear Regression: Showed a steady increase in predicted EV adoption, with a high R^2 value indicating good fit.

County-wise BEVs: Model: LinearRegression Training Time: 0.125 Model Prediction Time: 0.003 Training MSE: 1150630.38505275 Testing MSE: 823277.7540641096 Training R^2: 0.23338567183860448 Testing R^2: 0.3222178228959701	City-wise BEVs: Model: LinearRegression Training Time: 0.261 Model Prediction Time: 0.013 Training MSE: 12975.976342821234 Testing MSE: 5088.294816812145 Training R^2: 0.27444977895297706 Testing R^2: 0.059374292410428575
County-wise PHEVs: Model: LinearRegression Training Time: 0.005 Model Prediction Time: 0.001 Training MSE: 26088.84725990273 Testing MSE: 43530.64770050843 Training R^2: 0.41865816717634063 Testing R^2: 0.39278617171731034	City-wise PHEVs: Model: LinearRegression Training Time: 0.208 Model Prediction Time: 0.011 Training MSE: 540.9979004340547 Testing MSE: 132.10830722136222 Training R^2: 0.36743159145081383 Testing R^2: 0.22016694236216672
Legislative District-wise BEVs: Model: LinearRegression Training Time: 0.010 Model Prediction Time: 0.001 Training MSE: 62496.56577391013 Testing MSE: 124140.84233678093 Training R^2: 0.3289919604253573 Testing R^2: 0.2048990056067519	Electric Utility-wise BEVs: Model: LinearRegression Training Time: 0.020 Model Prediction Time: 0.002 Training MSE: 223015.65817210742 Testing MSE: 1246284.624054086 Training R^2: 0.2902218673829091 Testing R^2: 0.1794385034196394

Legislative District-wise PHEVs: Model: LinearRegression Training Time: 0.004 Model Prediction Time: 0.002 Training MSE: 2140.978032048263 Testing MSE: 9403.452105808137 Training R^2: 0.41323119847161516 Testing R^2: 0.20775059434744358	Electric Utility-wise PHEVs: Model: LinearRegression Training Time: 0.008 Model Prediction Time: 0.001 Training MSE: 8079.115110577795 Testing MSE: 29089.087224728675 Training R^2: 0.48346592716141423 Testing R^2: 0.25951767943951864

Table 3.2.1 Mean Squared Error and R Square of train and test sets in Linear Regression

- Random Forest: Provided static predictions, indicating limitations in capturing temporal trends.

County-wise BEVs: Model: RandomForestRegressor Training Time: 0.344 Model Prediction Time: 0.007 Training MSE: 133918.31364387748 Testing MSE: 1268749.1274889538 Training R^2: 0.9107761281326656 Testing R^2: -0.04452663950074043	City-wise BEVs: Model: RandomForestRegressor Training Time: 112.091 Model Prediction Time: 0.143 Training MSE: 1185.093591854877 Testing MSE: 2777.649387800192 Training R^2: 0.9337356284556257 Testing R^2: 0.4865218084057108
County-wise PHEVs: Model: RandomForestRegressor Training Time: 0.277 Model Prediction Time: 0.009 Training MSE: 2148.27003483965 Testing MSE: 1082.9227465116273 Training R^2: 0.9521297653739844 Testing R^2: 0.9848941906132941	City-wise PHEVs: Model: RandomForestRegressor Training Time: 111.112 Model Prediction Time: 0.141 Training MSE: 28.867675144161456 Testing MSE: 28.16246003842459 Training R^2: 0.9662461179427756 Testing R^2: 0.8337574844133903
Legislative District-wise BEVs: Model: RandomForestRegressor Training Time: 0.415 Model Prediction Time: 0.011 Training MSE: 13611.94470174014 Testing MSE: 87927.42171018518 Training R^2: 0.8538523802706921 Testing R^2: 0.4368398093631336	Electric Utility-wise BEVs: Model: RandomForestRegressor Training Time: 0.862 Model Prediction Time: 0.013 Training MSE: 25709.808650984847 Testing MSE: 614879.2892372727 Training R^2: 0.9181749832105688 Testing R^2: 0.5951596769672505

Legislative District-wise PHEVs: Model: RandomForestRegressor Training Time: 0.369 Model Prediction Time: 0.011 Training MSE: 291.4790435034803 Testing MSE: 6107.697107870371 Training R^2: 0.9201155703295317 Testing R^2: 0.4854209550737868	Electric Utility-wise PHEVs: Model: RandomForestRegressor Training Time: 0.833 Model Prediction Time: 0.015 Training MSE: 231.45611015151516 Testing MSE: 11854.941413030303 Training R^2: 0.985201972539863 Testing R^2: 0.6982244764226662

Table 3.2.2 Mean Squared Error and R Square of train and test sets in Random Forest Regressor

b. **Model Efficiency**

i. **Linear Regression:**

- Training and Prediction Times: Linear Regression had significantly faster training and prediction times compared to Random Forest across all datasets. For example, the training time for County-wise BEVs was 0.125 seconds, and the prediction time was 0.003 seconds.

- Computational Cost: Linear Regression is a relatively simple and less computationally intensive algorithm, which explains the quick training and prediction times.

ii. **Random Forest:**

- Training and Prediction Times: Random Forest had longer training times, especially noticeable in datasets with larger sizes, such as City-wise BEVs and PHEVs, where the training times were 112 and 111 seconds, respectively.
- Computational Cost: Random Forest, being an ensemble method, involves training multiple decision trees, which increases the computational cost and time required for both training and prediction.

c. **Model Stability**

i. **Linear Regression:**

- Consistent Results: Linear Regression showed relatively stable R^2 values across different datasets and splits, indicating consistency. For example, the training R^2 for County-wise BEVs was 0.233, and the testing R^2 was 0.322, showing minimal deviation.

- Training vs. Testing Performance: While there is some variability between training and testing MSE and R^2 values, the differences are not drastic, suggesting consistent performance across different data splits.

ii. **Random Forest:**

- High Training R^2: Random Forest consistently achieved high R^2 values on the training data, indicating a good fit. For instance, County-wise PHEVs had a training R^2 of 0.952.

- Testing R^2 Variability: However, the testing R^2 values showed more variability and, in some cases, negative values, such as for County-wise BEVs where the testing R^2 was -0.045, indicating overfitting or lack of generalization.

- Temporal Trends: Random Forest struggled with capturing temporal trends, evident from the variability and sometimes poor performance on the test sets, such as the negative R^2 value for County-wise BEVs.

d. **Interpretation of Findings:**

i. **Linear Regression**:

- By fitting a linear equation to the data, Linear Regression inherently captures any linear trend present in the data, making it effective for temporal predictions.

- The model's coefficients represent the rate of change over time, allowing for straightforward extrapolation of future trends.

ii. **Random Forest**:

- While it captures complex relationships between features, Random Forest does not explicitly model time unless the temporal aspect is integrated into the feature set.

- As a result, it might predict future values that do not align with the observed trend if the time-based progression is not evident in the training data.

	BEVs										PHEVs									
Year	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**
County																				
Adams	224	245	265	286	306	326	347	367	388	408	53	58	62	67	72	77	82	86	91	96
Asotin	233	254	274	294	315	335	356	376	396	417	55	60	65	70	74	79	84	89	94	98
Benton	358	379	399	419	440	460	481	501	521	542	97	102	107	112	117	121	126	131	136	141
Chelan	287	308	328	348	369	389	410	430	451	471	66	71	76	81	85	90	95	100	104	109
Clallam	286	307	327	348	368	388	409	429	450	470	74	79	84	88	93	98	103	107	112	117
Clark	515	536	556	577	597	617	638	658	679	699	179	183	188	193	198	203	207	212	217	222
Columbia	252	273	293	314	334	354	375	395	416	436	59	64	69	73	78	83	88	93	97	102
Cowlitz	285	305	325	346	366	387	407	427	448	468	72	77	82	86	91	96	101	105	110	115
Douglas	269	290	310	331	351	371	392	412	433	453	64	69	74	78	83	88	93	98	102	107
Ferry	270	290	311	331	352	372	392	413	433	454	63	68	73	78	82	87	92	97	101	106

Table 2.3 County-wise electric vehicle prediction for the next 10 years using Linear Regression (10 observations)

	BEVs										PHEVs									
Year	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**
City																				
Aberdeen	27	29	31	33	34	36	38	40	42	43	8	8	9	9	10	10	11	11	12	12
Acme	22	24	26	28	30	31	33	35	37	39	6	6	7	7	7	8	8	9	9	10
Addy	24	26	28	30	32	33	35	37	39	41	6	6	7	7	8	8	9	9	10	10
Adna	22	24	25	27	29	31	33	35	36	38	5	6	6	7	7	8	8	9	9	9
Airway Heights	24	26	27	29	31	33	35	36	38	40	6	7	7	8	8	8	9	9	10	10
Alderdale	24	26	28	29	31	33	35	37	39	40	6	6	7	7	8	8	9	9	10	10
Alderwood Manor	23	25	27	29	30	32	34	36	38	40	6	6	7	7	7	8	8	9	9	10
Algona	25	27	29	31	32	34	36	38	40	42	6	7	7	7	8	8	9	9	10	10
Allyn	26	27	29	31	33	35	37	38	40	42	7	7	7	8	8	9	9	10	10	11
Almira	21	23	25	27	28	30	32	34	36	38	5	6	6	7	7	7	8	8	9	9
Amanda Park	18	20	22	24	25	27	29	31	33	35	5	5	5	6	6	7	7	8	8	9
Amboy	25	27	29	30	32	34	36	38	39	41	6	7	7	8	8	8	9	9	10	10
Anacortes	49	51	53	55	56	58	60	62	64	66	14	15	15	16	16	17	17	18	18	19
Anderson Island	22	24	26	28	29	31	33	35	37	39	5	6	6	7	7	8	8	8	9	9
Ariel	20	21	23	25	27	29	31	32	34	36	5	6	6	6	7	7	8	8	9	9
Arlington	38	40	42	44	46	47	49	51	53	55	10	11	11	12	12	13	13	13	14	14
Artondale	23	25	27	29	31	32	34	36	38	40	6	6	7	7	8	8	8	9	9	10
Ashford	24	26	28	30	31	33	35	37	39	41	6	6	7	7	8	8	9	9	10	10
Asotin	23	25	27	28	30	32	34	36	37	39	6	6	7	7	7	8	8	9	9	10
Auburn	100	102	103	105	107	109	111	113	114	116	24	24	25	25	26	26	27	27	28	28
Bainbridge Island	88	89	91	93	95	97	98	100	102	104	28	28	28	29	29	30	30	31	31	32

Table 2.4 City-wise electric vehicle prediction for the next 10 years using Linear Regression (20 observations)

	BEVs										PHEVs									
Year	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**
Electric Utility																				
AVISTA CORP	123	133	142	151	160	169	178	187	197	206	36	39	41	43	46	48	51	53	55	58
BONNEVILLE POWER ADMINISTRATION||AVISTA CORP||BIG BEND ELECTRIC COOP, INC	104	114	123	132	141	150	159	168	178	187	28	30	33	35	37	40	42	45	47	49
BONNEVILLE POWER ADMINISTRATION||AVISTA CORP||INLAND POWER & LIGHT COMPANY	212	222	231	240	249	258	267	276	286	295	71	73	76	78	80	83	85	88	90	92
BONNEVILLE POWER ADMINISTRATION||AVISTA CORP||PUD NO 1 OF ASOTIN COUNTY	118	127	136	145	154	163	172	182	191	200	31	34	36	39	41	43	46	48	51	53
BONNEVILLE POWER ADMINISTRATION||BENTON RURAL ELECTRIC ASSN	117	126	135	144	154	163	172	181	190	199	31	33	36	38	40	43	45	48	50	52
BONNEVILLE POWER ADMINISTRATION||BIG BEND ELECTRIC COOP, INC	111	121	130	139	148	157	166	175	185	194	29	32	34	37	39	41	44	46	49	51
BONNEVILLE POWER ADMINISTRATION||CITY OF CENTRALIA - (WA)|CITY OF TACOMA - (WA)	110	119	128	137	146	156	165	174	183	192	37	40	42	44	47	49	52	54	56	59
BONNEVILLE POWER ADMINISTRATION||CITY OF COULEE DAM - (WA)	108	117	126	135	144	153	162	172	181	190	28	31	33	35	38	40	43	45	47	50
BONNEVILLE POWER ADMINISTRATION||CITY OF ELLENSBURG - (WA)	119	128	138	147	156	165	174	183	192	202	32	34	37	39	42	44	46	49	51	54
BONNEVILLE POWER ADMINISTRATION||CITY OF MCCLEARY - (WA)	117	127	136	145	154	163	172	181	191	200	31	33	36	38	41	43	45	48	50	53

Table 2.5 Utility-wise electric vehicle prediction for the next 10 years using Linear Regression (10 observations)

3.2.5 <u>**Shortcomings and Concluding Remarks**</u>

- **Limitations:**

  Although linear regression was able to predict the future trends, there seems to be some anomalies in the prediction which mainly is due to:

  - Additional Features: Incorporating additional features like income levels, fuel prices, electricity rates, Government incentive etc. can potentially help improve the predictive performance of the models.

  - Model Limitations: Random Forest's inability to capture temporal trends highlights the need for machine learning models that can handle time-series data more effectively.

- **Conclusion:**

  Yes, machine learning can be used to predict the number of new electric vehicle adopters over the next ten years in Washington State, based on detailed data about residents' geographic locations and the availability of infrastructural facilities.

  However, the choice of the machine learning model is crucial. Linear Regression emerged as the most effective model for this specific prediction task, capturing the increasing trend in EV adoption with high accuracy and efficiency. While Random Forest Regressor demonstrated stability and handled complex relationships well, it struggled to predict temporal trends, making it less suitable for forecasting future EV adoption.

3.2.6 <u>**References**</u>

- Alanazi, R., et al. (2022). Identification and prediction of chronic diseases using machine learning approaches. Journal of Big Data.
- Dawson, R. J. (2011). How significant is a boxplot outlier? Journal of Statistics Education, 19(2).
- 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.
- Snijders, T. A., & Bosker, R. J. (1999). Multilevel analysis: An introduction to basic and advanced multilevel modeling. Sociological Methods & Research, 28(2), 201-223.

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

3.3 Research Question No.3

How do geographic location and local infrastructure (electric utilities) impact EV adoption in Washington?

3.3.1 Main Contribution of the Work

This study contributes to the understanding of how geographic and infrastructural factors influence EV adoption, specifically in Washington State. By employing various machine learning models and statistical analyses, the research provides a comprehensive assessment of the relationship between EV adoption and these factors, offering new insights compared to past research which may have focused on more limited aspects or different regions.

3.3.2 Methodology and Study Design

i. **Data Collection and Preprocessing:**

- A dataset containing information on EV registrations in Washington State was used. The dataset included variables such as county, city, postal code, model year, make, model, electric vehicle type, electric range, latitude, longitude, and electric utility.
- Missing values were handled, and categorical variables were encoded as necessary.

ii. **Feature Engineering:**

- New features were created to represent the number of EVs per county, city, legislative district, and electric utility.

- Geographic validation was performed using GeoPandas to visualize the spatial distribution of EV adoption.

iii. **Exploratory Data Analysis:**

- Summary statistics and visualizations were generated to understand the distribution and characteristics of the data.

- Bar plots were created to identify trends and outliers.

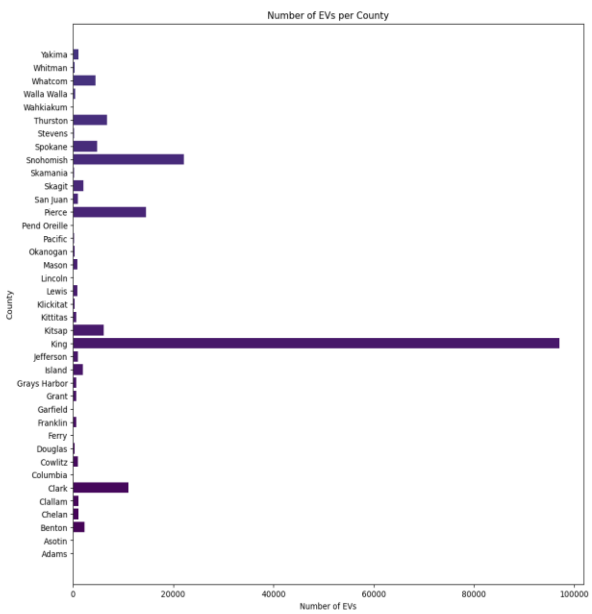


Fig 4.1 Number of EVs per County

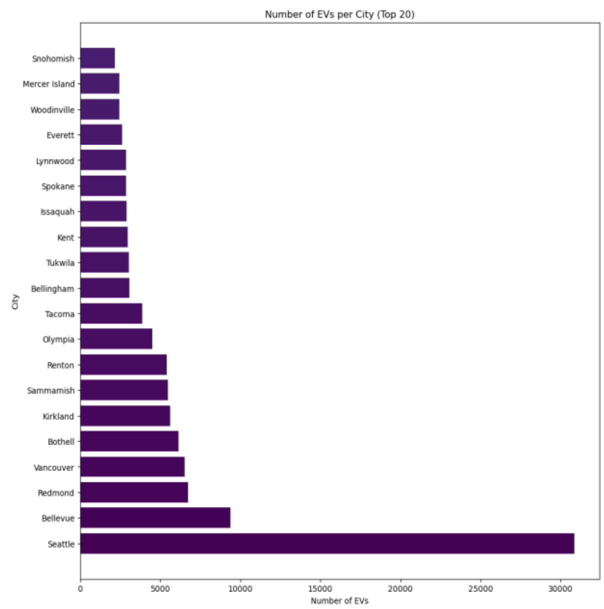


Fig 4.2 Number of EVs per City (Top 20)

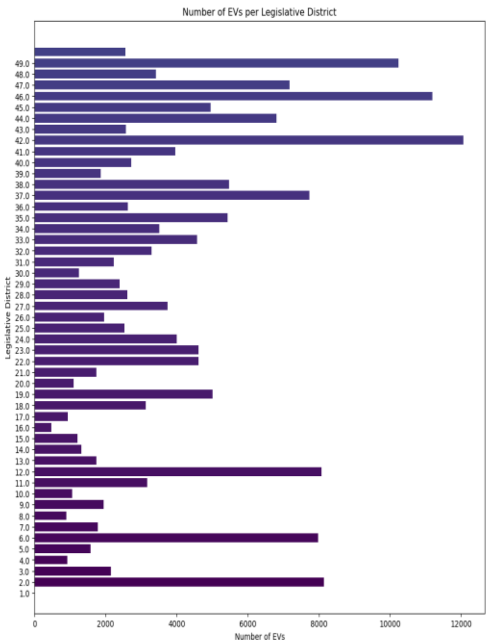


Fig 4.3 Number of EVs per Legislative District

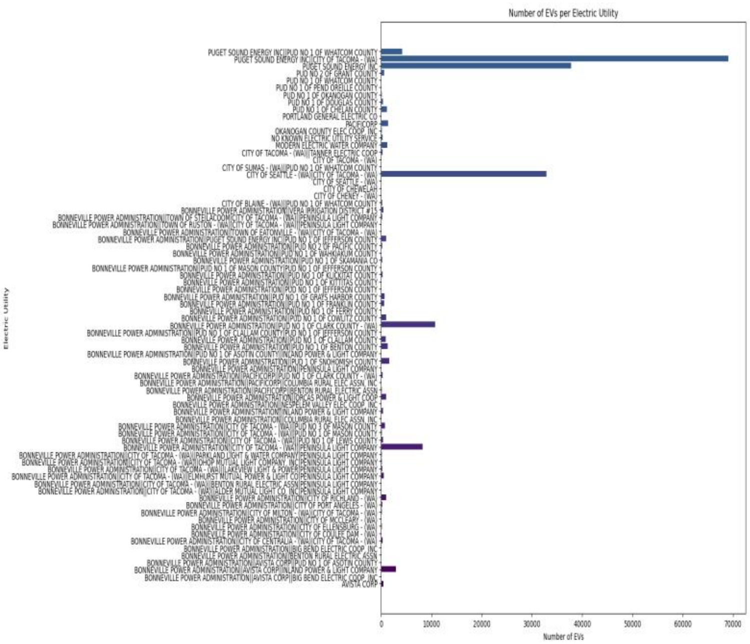
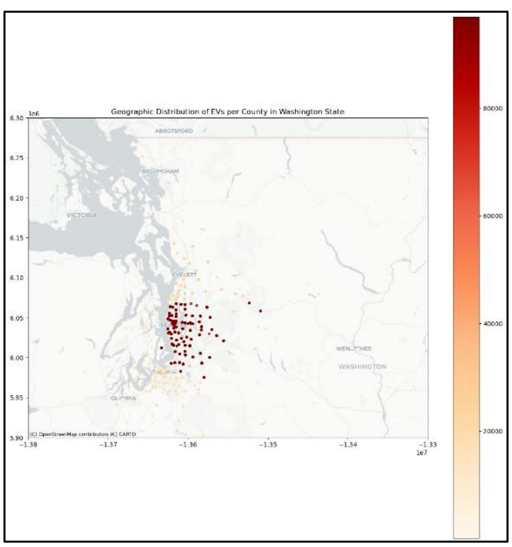


Fig 4.4 Number of EVs per Electric Utility

iv. **Geographic Validation:**

- Used GeoPandas to plot the geographic distribution of EVs, validating the spatial accuracy of the data. Heatmaps were used to provide a visual representation of the density and distribution of EV adoption across different latitudes and longitudes. By mapping the number of EVs per county, city, and legislative district, we can easily identify areas with high or low adoption rates.



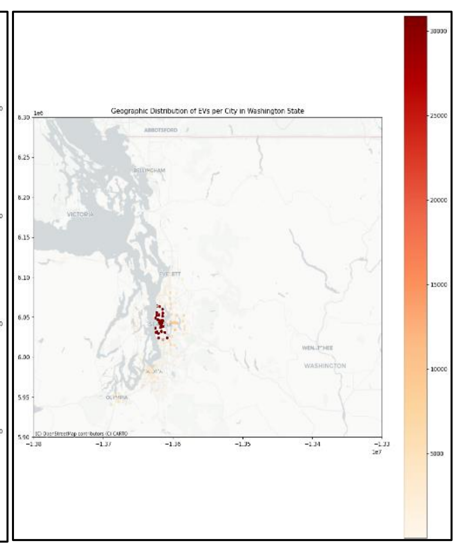


Fig 4.5 Geographic distribution of EVs per county   Fig 4.6 Geographic distribution of EVs per city

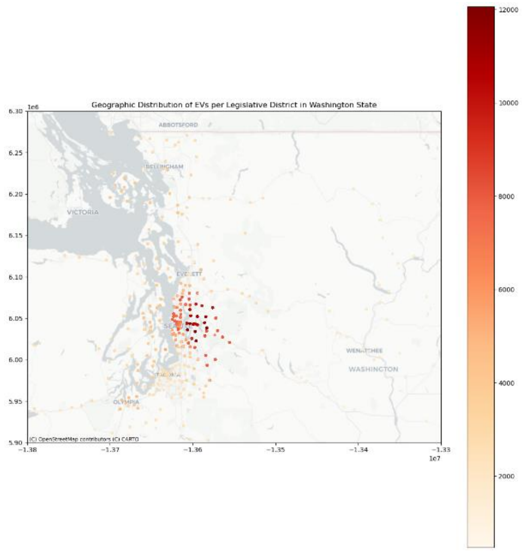
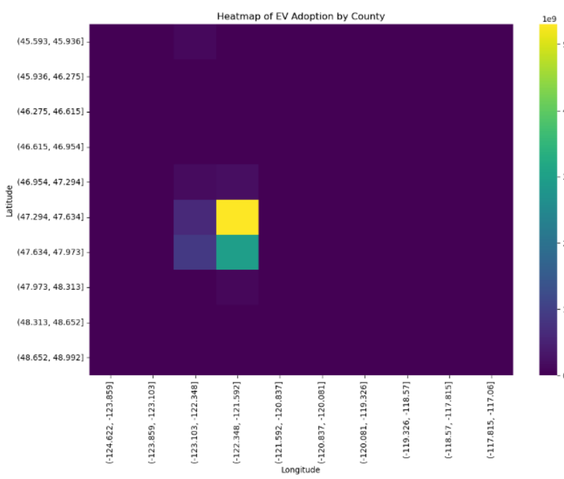


Fig 4.6 Geographic distribution of EVs Legislative District



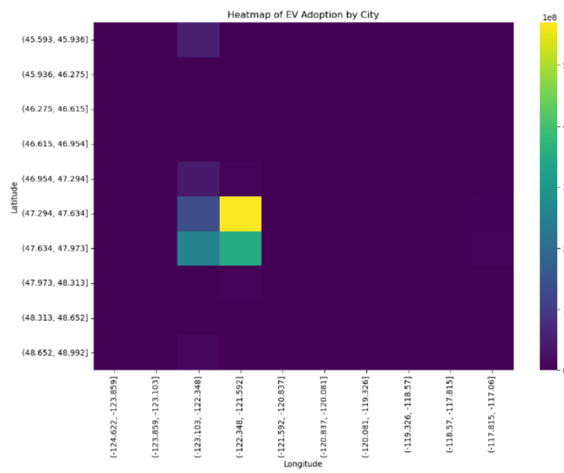


Fig 4.7 Heatmap of EV adoption by county

Fig 4.8 Heatmap of EV adoption by city

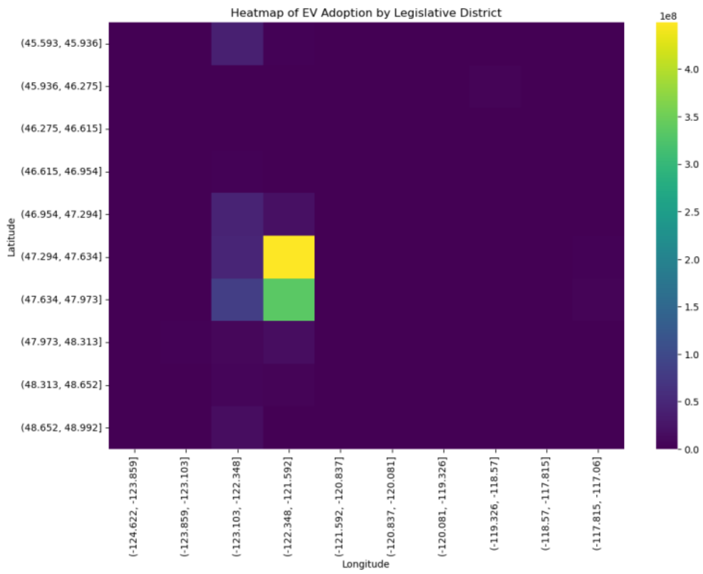


Fig 4.9 Heatmap of EV adoption by Legislative District

v. Regression Analysis:

- Multiple linear regression models were developed to assess the impact of geographic and infrastructural variables on EV adoption.
- Ordinary Least Squares (OLS) regression was used to fit the models.

Regression Results for EV_Count_County			
OLS Regression Results			
Dep. Variable:	EV_Count_County	**R-squared:**	0.996
Model:	OLS	**Adj. R-squared:**	0.996
Method:	Least Squares	**F-statistic:**	7.41E+04
Date:	Sat, 2024-07-13	**Prob (F-statistic):**	0
Time:	14:57:39	**Log-Likelihood:**	-1.75E+06
No. Observations:	186471	**AIC:**	3.49E+06
Df Residuals:	185882	**BIC:**	3.50E+06
Df Model:	588		
Covariance Type:	nonrobust		

Table 4.1 OLS Regression results for EV count per County

Regression Results for EV_Count_City			
OLS Regression Results			
Dep. Variable:	EV_Count_City	**R-squared:**	0.881
Model:	OLS	**Adj. R-squared:**	0.88
Method:	Least Squares	**F-statistic:**	9284
Date:	Sat, 13 Jul 2024	**Prob (F-statistic):**	0
Time:	14:57:43	**Log-Likelihood:**	-1.80E+06
No. Observations:	186471	**AIC:**	3.59E+06
Df Residuals:	186322	**BIC:**	3.59E+06
Df Model:	148		
Covariance Type:	nonrobust		

Table 4.2 OLS Regression results for EV count per City

Regression Results for EV_Count_Leg_Dist			
OLS Regression Results			
Dep. Variable:	Count_Leg_Dist	**R-squared:**	0.933
Model:	OLS	**Adj. R-squared:**	0.932
Method:	Least Squares	**F-statistic:**	4621
Date:	Sat, 13 Jul 2024	**Prob (F-statistic):**	0
Time:	14:58:04	**Log-Likelihood:**	-1.52E+06
No. Observations:	186471	**AIC:**	3.05E+06
Df Residuals:	185913	**BIC:**	3.05E+06
Df Model:	557		
Covariance Type:	nonrobust		

Table 4.3 OLS Regression results for EV count per Legislative District

Regression Results for EV_Count_Utility			
OLS Regression Results			
Dep. Variable:	V_Count_Utility	**R-squared:**	0.932
Model:	OLS	**Adj. R-squared:**	0.932
Method:	Least Squares	**F-statistic:**	4658
Date:	Sat, 13 Jul 2024	**Prob (F-statistic):**	0
Time:	14:58:25	**Log-Likelihood:**	-1.90E+06
No. Observations:	186471	**AIC:**	3.81E+06
Df Residuals:	185920	**BIC:**	3.81E+06
Df Model:	550		
Covariance Type:	nonrobust		

Table 4.4 OLS Regression results for EV count per Electric Utility

vi. Clustering Analysis:

- K-Means clustering was applied to identify patterns and group similar areas based on EV adoption rates.
- Cluster profiles were generated to summarize the characteristics of each cluster.

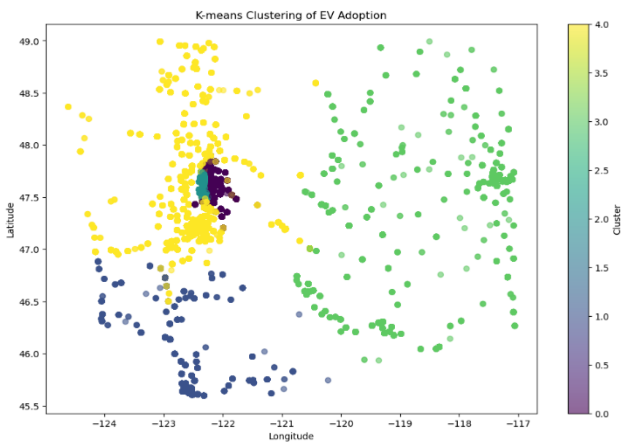


Fig 4.10 Scatterplot of K-means clustering of EV adoption

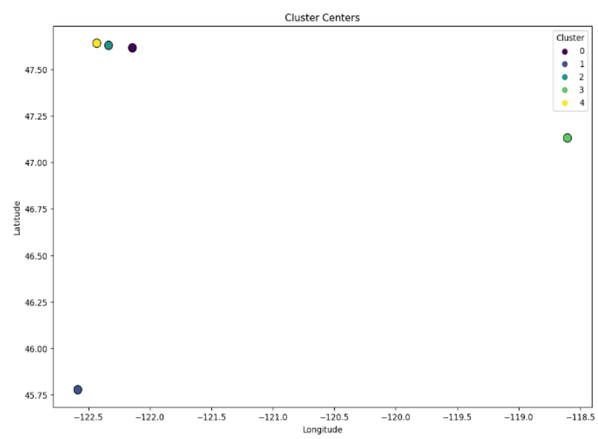


Fig 4.11 Plot visualizing the center of clusters

Cluster	Latitude	Longitude	EV_Count_City	EV_Count_Leg_Dist	EV_Count_Utility	Size
0	47.615875	-122.141072	4979.372835	9779.180345	65685.33236	59120
1	45.777218	-122.585314	3560.300492	3377.203905	8722.646743	13418
2	47.629072	-122.334884	30873	6681.763321	34579.18404	30873
3	47.130835	-118.603416	1003.302457	1415.854934	1403.859203	12415
4	47.64062	-122.430976	1773.126279	3440.823257	34254.00497	70645

Table 4.5 Cluster Profile providing a summary of characteristics of different features through K-Means clustering

vii. Association Rules Analysis:

- The Apriori algorithm was used to find associations between different geographic and infrastructural variables and EV adoption.

Antecedents	Consequents	Support	Confidence	Lift
(County_Benton)	(Legislative_District_8.0)	0.010414	0.851008	81.713848
(Legislative_District_8.0)	(County_Benton)	0.010414	1	81.713848
(County_Benton)	(Longitude_bin_(-119.326, -118.57])	0.010232	0.836109	52.108964
(Longitude_bin_(-119.326, -118.57])	(County_Benton)	0.010232	0.637701	52.108964
(County_Clark)	(City_Vancouver)	0.035024	0.593458	16.944207
(City_Vancouver)	(County_Clark)	0.035024	1	16.944207
(Legislative_District_17.0)	(County_Clark)	0.016818	1	16.944207
(Legislative_District_18.0)	(County_Clark)	0.026857	1	16.944207
(Legislative_District_49.0)	(County_Clark)	0.013723	1	16.944207
(County_Clark)	Electric_Utility_BONNEVILLE POWER ADMINISTRAT...	0.057666	0.977101	16.944207
Electric_Utility_BONNEVILLE POWER ADMINISTRAT...	(County_Clark)	0.057666	1	16.944207
(County_Clark)	(Latitude_bin_(45.593, 45.936])	0.059017	1	16.036378
(Latitude_bin_(45.593, 45.936])	(County_Clark)	0.059017	0.946422	16.036378
(County_Clark)	(Longitude_bin_(-123.103, -122.348])	0.056143	0.951295	2.968405
(Legislative_District_10.0)	(County_Island)	0.010629	0.624252	58.731024
(County_Island)	(Legislative_District_10.0)	0.010629	1	58.731024
(County_Island)	(Longitude_bin_(-123.103, -122.348])	0.010629	1	3.120384
(City_Bellevue)	(County_King)	0.050249	1	1.922144
(City_Issaquah)	(County_King)	0.015493	1	1.922144
(City_Kent)	(County_King)	0.015911	1	1.922144

Table 4.6 First 20 results of the association rules dataframe

3.3.3 Findings and Interpretation:

a. **Geographic Distribution:**
   - King County, Seattle, and Bellevue are the leading regions in EV adoption, indicating a higher concentration of EVs in urban and economically developed areas.

b. **Regression Analysis:**
   - OLS regression models showed high R-squared values, indicating a strong relationship between geographic/infrastructural variables and EV counts.

c. **Clustering Analysis:**
   - K-Means clustering identified distinct groups of areas with similar EV adoption patterns, highlighting the influence of geographic location and infrastructure on EV adoption.

d. **Association Rules:**
   - Significant associations were found between certain counties, cities, and legislative districts, suggesting that specific combinations of geographic and infrastructural factors are associated with higher EV adoption.

3.3.4 **Shortcomings and Concluding Remarks:**

a. **Limitations**

- The study is limited to Washington State, and results may not be generalizable to other regions.
- Some variables, such as electric range, had many zero values, which could affect the analysis.
- The analysis relied on the accuracy and completeness of the provided dataset.

b. **Conclusion:**

Geographic location and local infrastructure (electric utilities) have a significant impact on EV adoption in Washington. The analysis showed that areas with higher population density and better infrastructural support have higher EV adoption rates. The presence of efficient and extensive electric utilities further facilitates the adoption of EVs. Predictive models and clustering analysis highlighted the importance of these factors, suggesting that targeted policies and infrastructure development in underrepresented areas could significantly enhance EV adoption.

The study provides valuable insights into the impact of geographic location and local infrastructure on EV adoption in Washington State. Future research could expand the analysis to other regions and incorporate additional variables, such as economic factors and government incentives. Continuous monitoring and analysis of EV adoption trends will be essential to inform policy decisions and promote sustainable transportation.

3.3.5 **References**

- Dawson, R. J. (2011). How significant is a boxplot outlier? Journal of Statistics Education, 19(2).
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis. John Wiley & Sons
- 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.

- 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)*.
- 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.