Sean McLean

ALY 6140

Module 6

Capstone Report

Introduction

The focus of the capstone project is to look at how to determine where to add more charging stations in the state of Washington based off current data of electric vehicle owners. This will be achieved by studying the demographics of current owners, what kind of electric vehicles they are driving and how many miles each vehicle can drive. Other factors crucial to this question are where vehicles are located per legislative district and if the vehicle has Clean Alternative Fuel Vehicle (CAFV) eligibility. By understanding where drivers reside and how far their cars can drive per charge, it can be determined where the appropriate places are for installing new charging stations across the state.

Exploratory Data Analysis

After the dataset is imported into Python, descriptive statistics are provided to look at the structure and what stands out. Initial observations of the dataset show that most of the electric vehicles are recent, with the majority being from 2019 to present. The average vehicle range for all the vehicles registered is only 52 miles, with the max range topping out at 337. The base MSRP which is the manufacturer's suggested retail price, has a low mean but upon further evidence there is limited data in the column for all electric cars.

The most common location for an electric vehicle owner is in Seattle, Washington and residing in King County. There are 42 makes and 152 models of electric cars registered in the state, with two types of electric batteries. The most popular electric vehicle is the Tesla Model Y, and over half the cars registered are a Tesla brand. Regarding the CAFV eligibility, over half of the cars registered have unknown eligibility as the battery range has not been researched. In the state, there are 74 electric utilities, the most common being Puget Sound Energy Incorporated, part of the city of Tacoma. These statistics are essential to figuring out how to determine where to add more charging stations in the state of Washington based off current data of electric vehicle owners.

Some of the key questions about the dataset imported into Python include what cities and counties have the most EV owners, how are legislative districts impacting EV ownership, and what are the most popular makes and models of electric vehicles. Other variables that can add substance to answering the main question of adding charging stations pertain to whether CAFV impacts sales of the EV models and what are the most common EV's by electric range.

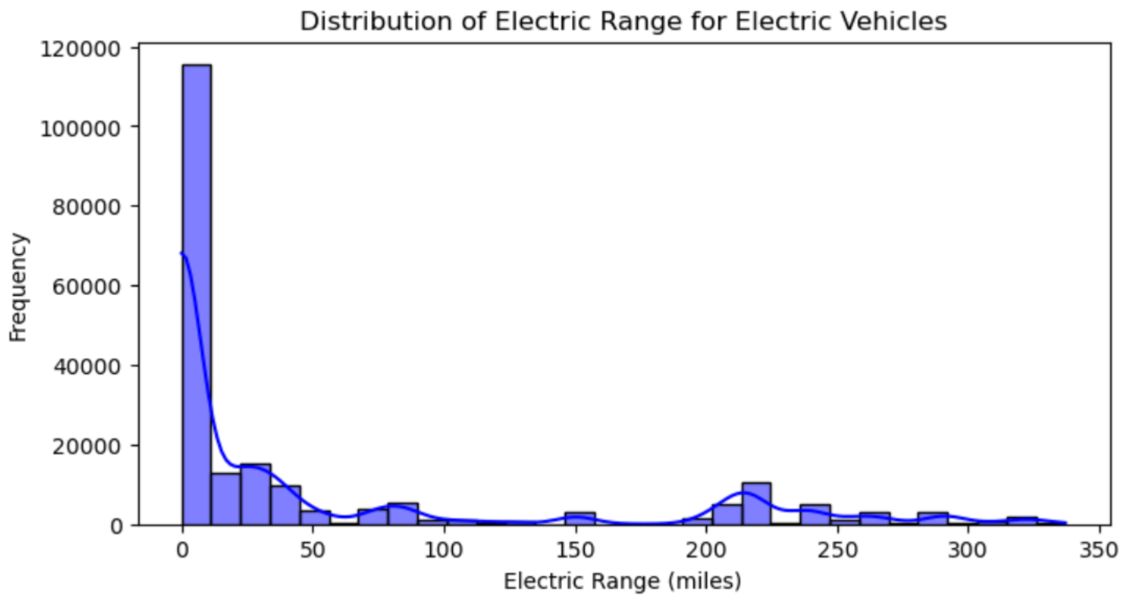
The first task in data cleaning involved removing several columns in the dataset that are deemed irrelevant or redundant. The two that were excluded were the 'VIN (1-10)' and 'Vehicle Location' columns since there are other columns that are similar to them. When checking for missing values in the dataset, most of the variables had at least a few which is a great revelation since there are over 205,000 rows. The one column that had more than just a few missing values was the 'Legislative District' with 442 total.

```
# Number of missing values per column
print(df.isnull().sum())
```

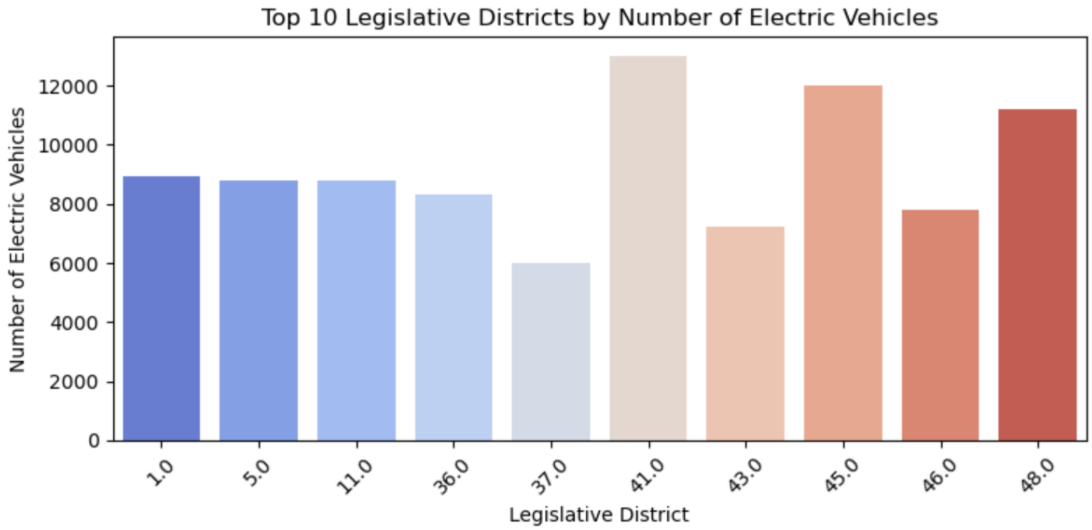
```
VIN (1-10)                                            0
County                                                3
City                                                  3
State                                                 0
Postal Code                                           3
Model Year                                            0
Make                                                  0
Model                                                 1
Electric Vehicle Type                                 0
Clean Alternative Fuel Vehicle (CAFV) Eligibility     0
Electric Range                                        8
Base MSRP                                             8
Legislative District                                442
DOL Vehicle ID                                        0
Vehicle Location                                      8
Electric Utility                                      3
2020 Census Tract                                     3
dtype: int64
```

Overall, since there were few values that were omitted, it was decided to fill the numerical columns with a '0' and the categorical columns with an 'NA'.

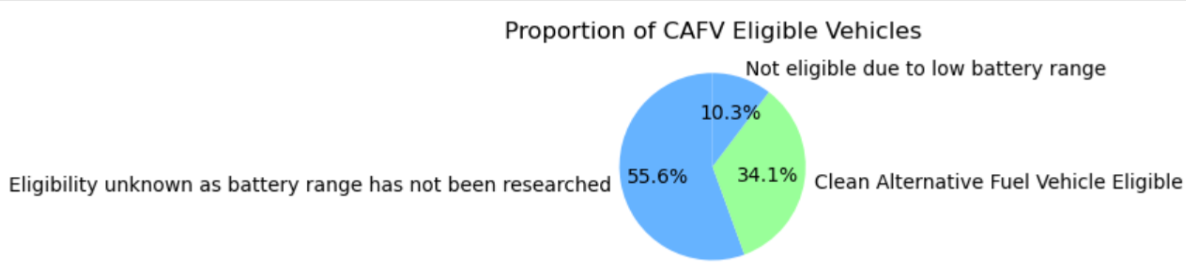
A bar chart was built to analyze the ranges of all the electric vehicles that are contained in the dataset. A large portion of the cars have very short ranges with a few cars intermittently in the chart show larger ranges that exceed 100 miles. This trend suggests that most electric vehicles at this juncture are made to only go short distances with some makes and models having the ability to go around 100 miles. This would also indicate that the vehicles will need charging stations nearby so installing them in places where these makes and models are pivotal. A geographic assessment would need to be required for this process and to also look at what makes and models are shorter ranges in mileage.



An evaluation of the top 10 legislative districts by number of electric vehicles is provided below in a bar chart. The top district in the state of Washington is 41 followed by districts 45 and 48, and all three districts have over 10,000 owners. A further look at these legislative districts might be interesting to see how this could impact sales of electric vehicles and increase the chances of people having vehicles in these districts. From a political standpoint this could also affect how many charging stations can be added to certain districts, so perhaps another dataset that shows this could be analyzed.



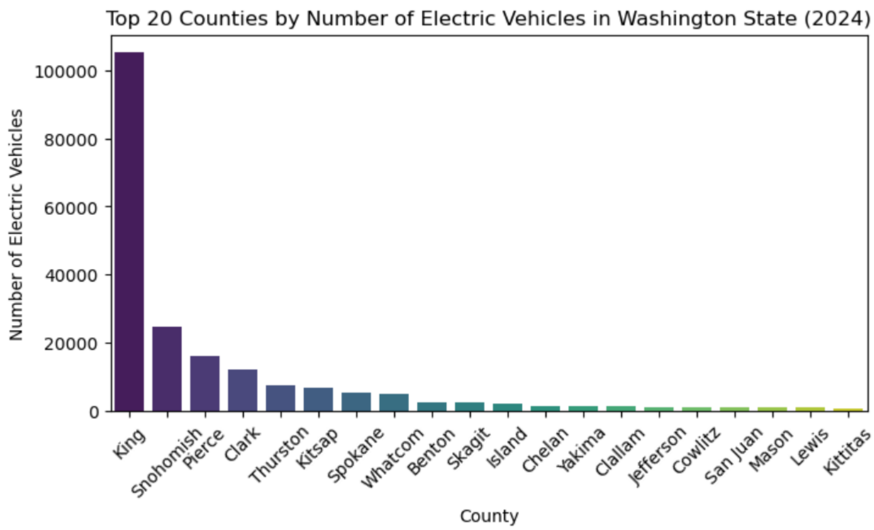
A pie chart was created to assess the proportion of CAFV eligible vehicles in Washington. Over half of the cars have unknown eligibility because their battery range has not been researched, while 10 percent are not eligible due to low battery range. The eligibility guidelines are in place and need to be met so the vehicles are considered clean and potentially receive certain benefits and incentives by government (Hari, 2023). These kinds of standards are akin to some state's annual inspections for all vehicles, and this could potentially affect whether more charging stations are in place. The chart also shows that there is still research that needs to be done on electric batteries and if they meet the guidelines of clean alternative fuel vehicle eligibility. A question to consider is if the electric vehicles do not meet clean standards, then that could be used against them for improving electric vehicle infrastructure.



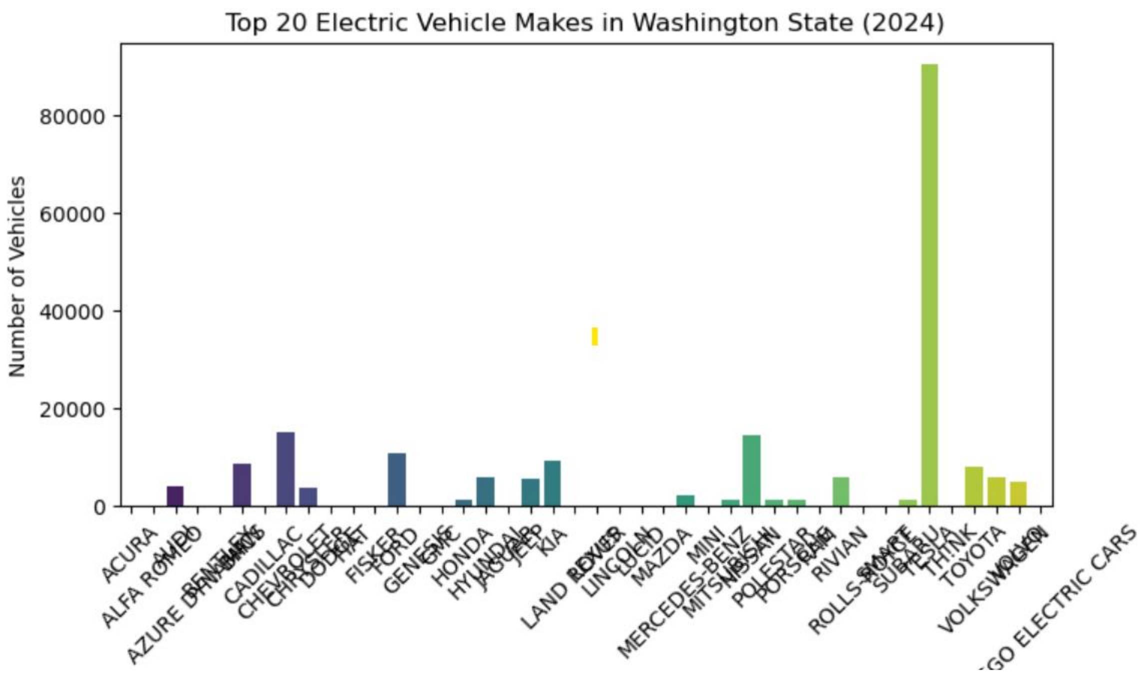
Predictive Models (Statistical/Predictive Analysis)

Descriptive Statistics & Data Visualization: Descriptive statistics and visualizations like bar charts can provide a summary of EV distribution across regions and show the popularity of various makes and models. This approach is crucial for identifying key trends and offering an overall view of the dataset. A bar chart below shows the top 20 counties by electric vehicles in the state. The highest county by far is King County, followed by Snohomish, Pierce, and Clark counties. The bar chart following that shows the top 10 electric makes with Tesla having by far the most followed by Chevrolet, Nissan, and Ford.

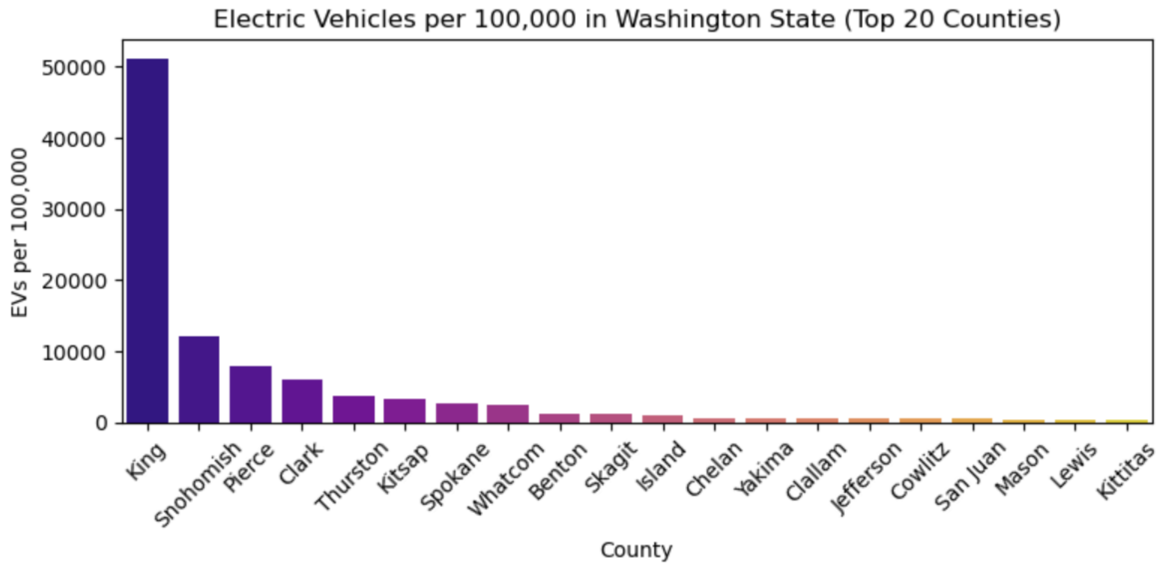
With the amount of people in some counties compared to others, a bar chart was created to show the number of electric cars per capita in each country with the adjustment being set at 100,000 people. The results show that the top 20 counties after this per capita revision did not change the rankings when compared to the total number overall per county. This indicates that people in densely populated areas are still purchasing and driving electric vehicles at higher rates than areas that are more sparsely populated. It also provides ample evidence that Tesla is still the dominant model for electric cars and that other car companies are behind in popularity and standards that Tesla offers.



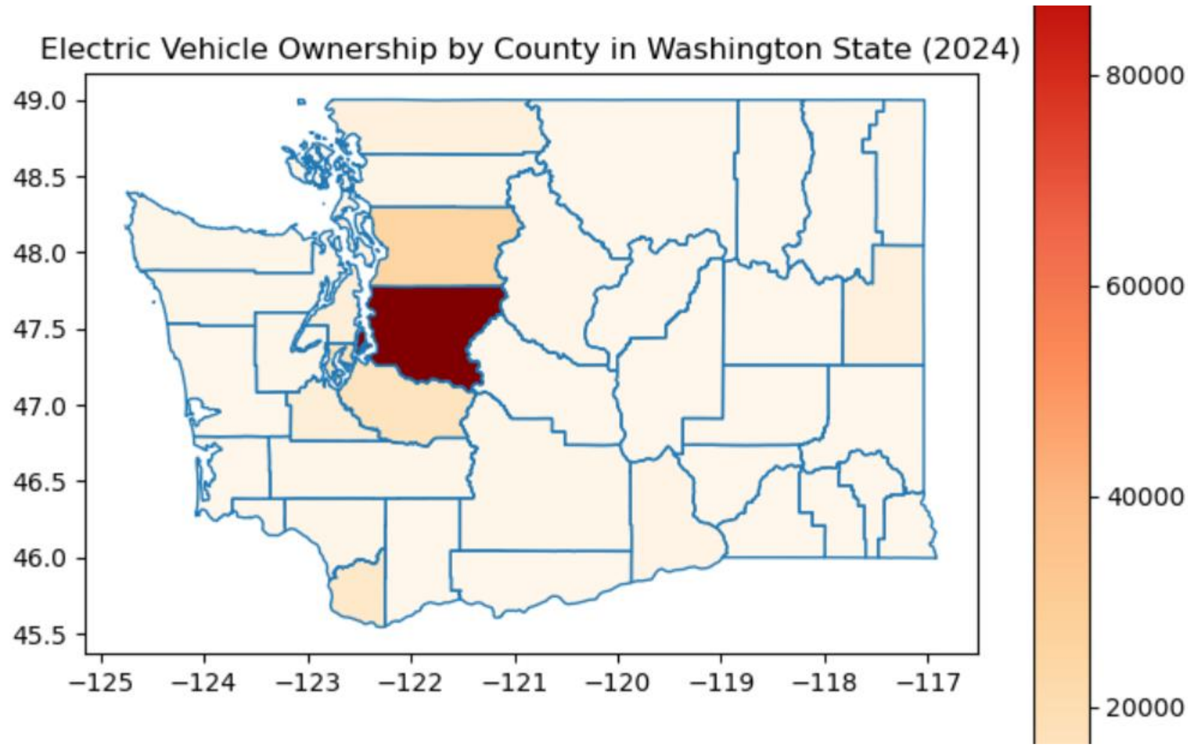
Top 20 Counties by Number of Electric Vehicles in Washington State (2024)
Number of Electric Vehicles
0
20000
40000
60000
80000
100000
King
Snohomish
Pierce
Clark
Thurston
Kitsap
Spokane
Whatcom
Benton
Skagit
Island
Chelan
Yakima
Clallam
Jefferson
Cowlitz
San Juan
Mason
Lewis
Kittitas
County



Top 20 Electric Vehicle Makes in Washington State (2024)
Number of Vehicles
0
20000
40000
60000
80000
ACURA
ALFA ROMEO
CADILLAC
FISKER
FORD
HONDA
KIA
LINCOLN
MAZDA
MINI
MERCEDES-BENZ
RIVIAN
TOYOTA
VOLKSWAGEN



Geospatial Analysis: Geospatial analysis can determine optimal locations for charging stations by mapping EV ownership alongside existing infrastructure. It is essential for identifying areas with growing demand and can guide decisions on resource allocation and infrastructure development. The map below shows high electric vehicle marks in populated counties which are on par with other data from certain cities, legislative districts, and counties. This could indicate that because it is in the Seattle area that it is not surprising that there are a lot of electric vehicle owners. It also means that more analysis of population per county could also help show whether there are any outliers or if ownership is just consistent with highly populated areas. With evidence that despite per capita calculations in population relative to EV owners, rural areas still lag behind urban areas in total ownership.

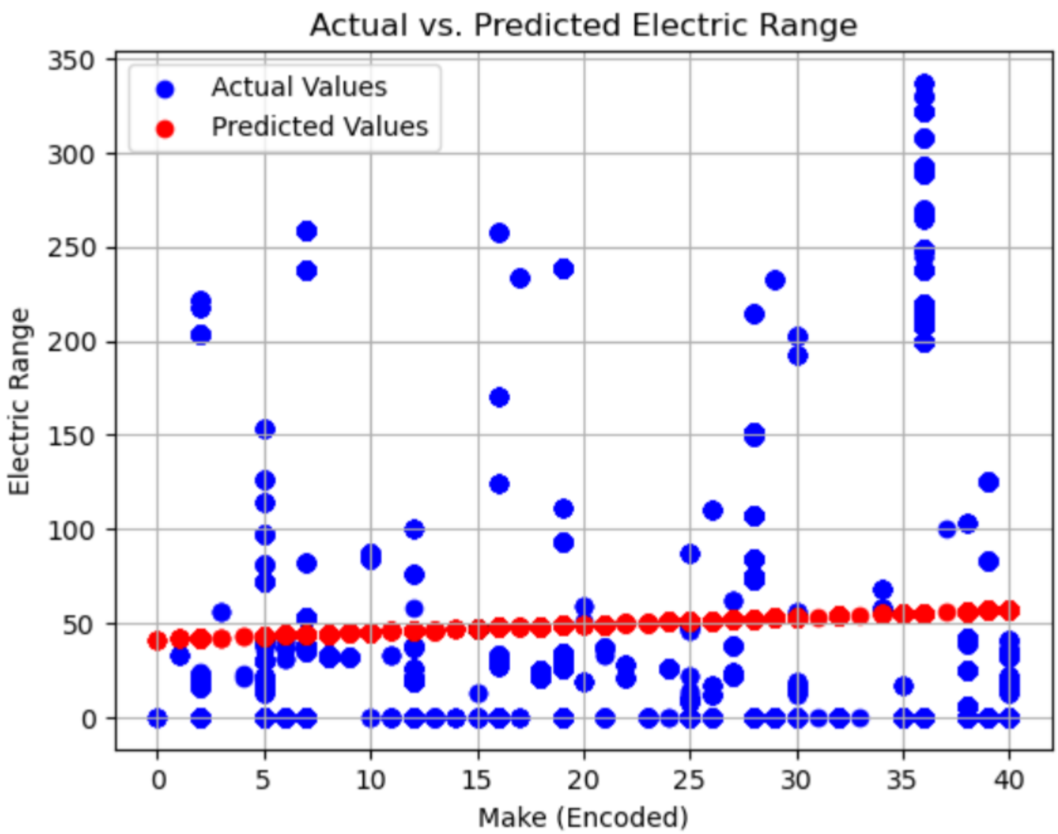


Generalized Linear Models: A Generalized Linear Model (GLM) regression helps explore the relationships between the dependent variable (e.g., Electric Range) and independent variables (e.g., Make, Model, or Legislative Districts) in a dataset. It was chosen for its ability to adapt to all sorts of data types and structures while being able to get past certain roadblocks while doing data analysis (Ahmed, 2024). Applying GLM to the dataset allows for identifying which factors significantly impact the electric vehicle's range. The model estimates the strength and direction of these influences through coefficients and evaluates their importance using p-values. The GLM model shows that 29 of the 46 makes and 54 of the 151 models have high coefficients, suggesting that they have a strong influence on the dependent variables used which is electric range and the independent variables are make and model.

When assessing the p-values of the model, every single make and all but nine models had p-values below 0.05, indicating that the electric car you drive will heavily affect the electric vehicle mile range it provides. This indicates that only a few makes and models have longer electric ranges and most likely contribute to higher sales of their vehicles. At this point this could also mean that a few brands are ahead of the game in having better battery life than other car companies that are still trying to improve their models.

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:         Electric Range   No. Observations:               205439
Model:                            GLM   Df Residuals:                   205286
Model Family:                Gaussian   Df Model:                          152
Link Function:               Identity   Scale:                          4996.6
Method:                          IRLS   Log-Likelihood:            -1.1663e+06
Date:                Sat, 26 Oct 2024   Deviance:                   1.0259e+09
Time:                        18:54:09   Pearson chi2:                 1.03e+09
No. Iterations:                     3   Pseudo R-squ. (CS):             0.4248
Covariance Type:            nonrobust
====================================================================================================
                                     coef    std err          z      P>|z|      [0.025      0.975]
```

GLM Scatterplot: To understand the actual versus electric ranges of the vehicle, a scatterplot was made to show this relationship. By encoding the categorical variable "make" as numeric values, the model estimates how changes in the make influence the electric range, indicated by the regression coefficients. The scatter plot visualizes this relationship, with blue points representing actual electric range values and red points representing predicted values, with closer alignment between these points suggesting a better predictive performance. The predicted values only slightly increases, indicating that regardless of the make that the range of the vehicle will not change drastically.



Interpretations & Conclusions

An analysis of the dataset by incorporating descriptive statistics, modeling, and other data visualizations offered a glimpse into where people in the state of Washington are driving electric vehicles and what makes and models they are purchasing. When answering the question of what factors can help determine where more charging stations should be installed, several columns like electric range, CAFV eligibility, and other demographic data can be effective. It provides information on where owners reside, what kind of makes and models they own, and how far their vehicles can typically go in mileage.

Another key component is over time how CAFV eligibility will evolve, meaning that if more vehicles become eligible in the future, then it can change the geography of where owners will need charging stations. Recommendations are to map out where charging stations are needed based off the data and compare that with a current charging station map in the state. A plan to do more research for CAFV eligibility and add more charging stations in rural areas is also recommended. The plan can then be put into place where it shows what parts of the state can benefit from additional charging stations.

References

U.S. Department of Energy. (2024, Sept. 20). *Electric vehicle population data*. Data.gov. https://catalog.data.gov/dataset/electric-vehicle-population-data

Hari, A. S. (2023, Dec. 4th). *Electric vehicle data analysis*. RPubs. Retrieved October 15, 2024, from https://rpubs.com/Anoop-S-Hari/1124971

Ahmed, S. (2024, Mar. 20th). *A Comprehensive Introduction to Generalized Linear Models.* Medium. https://medium.com/@sahin.samia/a-comprehensive-introduction-to-generalized-linear-models-fd773d460c1d#:~:text=Flexibility%20in%20Analysis%3A&text=GLMs%20provide%20a%20more%20versatile,of%20classical%20linear%20regression%20models.