Impact of CAFV Eligibility on EV Popularity in the State of Washington

MATH 4339

Marwan Aridi

Introduction

- We chose the Electric Vehicle (EV) Population dataset from the State of Washington due to the significant and growing trend of EV usage. This dataset provides valuable insights into the evolving EV market and offers an opportunity to explore its dynamics and influencing factors.
- This dataset contains 205,439 observations and 17 variables, providing a comprehensive overview of registered EVs in Washington. It encompasses a wide range of attributes related to EV characteristics, locations, and eligibility for clean energy programs. Each variable contributes uniquely to understanding the market:
    - **Column 1: *VIN (1-10)***: *A truncated vehicle identification number.*
    - **Column 2: *County***: *The county where the vehicle is registered.*
    - **Column 3: *City***: *The city of vehicle registration.*
    - **Column 4: *State***: *The state of vehicle registration.*
    - **Column 5: *Postal Code***: *ZIP code of the registration area.*
    - **Column 6: *Model Year***: *Year of the vehicle model.*
    - **Column 7: *Make***: *Manufacturer of the vehicle.*
    - **Column 8: *Model***: *Specific model name of the vehicle.*
    - **Column 9: *Electric Vehicle Type***: *Type of EV (e.g., Battery Electric Vehicle, Plug-in Hybrid).*
    - **Column 10: *Clean Alternative Fuel Vehicle (CAFV) Eligibility***: *Whether the vehicle qualifies for clean energy incentives.*
    - **Column 11: *Electric Range***: *The estimated range of the EV in miles on a full charge.*
    - **Column 12: *Base MSRP***: *Manufacturer's suggested retail price for the base model.*
    - **Column 13: *Legislative Distric***t: *Legislative district associated with the registration location.*
    - **Column 14: *DOL Vehicle ID***: *Unique identifier for the vehicle by the Department of Licensing.*
    - **Column 15: *Vehicle Location***: *Location of the vehicle registration.*
    - **Column 16: *Electric Utility***: *The utility company that provides electricity for EVs in the area.*
    - **Column 17: *2020 Census Tract***: *Geographic census tract information for demographic and planning analysis*
- **Response Variable:**

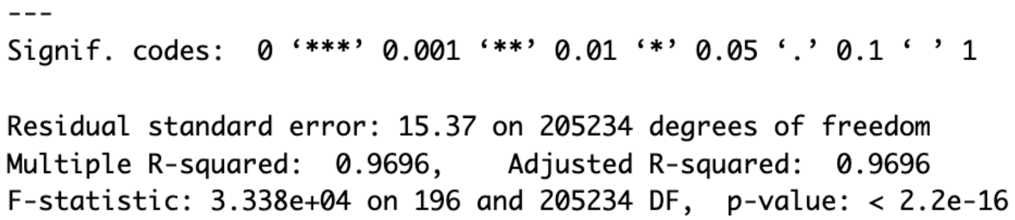
  - Electric Range: This can be treated as the response variable, as it is a key metric that can be influenced by other factors such as vehicle make, model, and type.
- **Potential Predictor Variables:**
  - Make: The vehicle manufacturer may significantly influence the electric range.
  - Model: Different models have different specifications affecting the electric range.
  - Model Year: Older or newer models may have different technological advancements influencing the range.
  - Electric Vehicle Type: Indicates whether the vehicle is a Battery Electric Vehicle (BEV) or a Plug-in Hybrid Electric Vehicle (PHEV), which directly impacts range.
  - Base MSRP: The cost might correlate with battery size or technology, hence affecting the electric range.
  - Clean Alternative Fuel Vehicle (CAFV) Eligibility: This eligibility status might be associated with incentives or technology that can affect vehicle range.
- **Data question: Does the CAFV Eligibility contribute to the popularity of EV make?** Through this dataset, we aim to analyze how factors like Clean Alternative Fuel Vehicle Eligibility (CAFV) influence the popularity of various EV makes and models. Using methods such as ANOVA and Multivariable Linear Regression, we will uncover patterns and relationships to better understand the market dynamics of EVs.

Methods & Results

- **Multivariable Linear Regression:**
  - Using multivariate linear regression (MLR) on this dataset is ideal because it allows you to analyze the relationship between multiple dependent variables and their predictors simultaneously. For instance, there are at least two dependent variables:
    - Electric Range: The range of electric vehicles.
    - Base MSRP: The manufacturer’s suggested retail price.
  - These two variables are likely correlated, meaning they share some pattern influenced by the same predictors (e.g. Model Year, Make). Multivariate linear regression is specifically designed to model such dependencies.

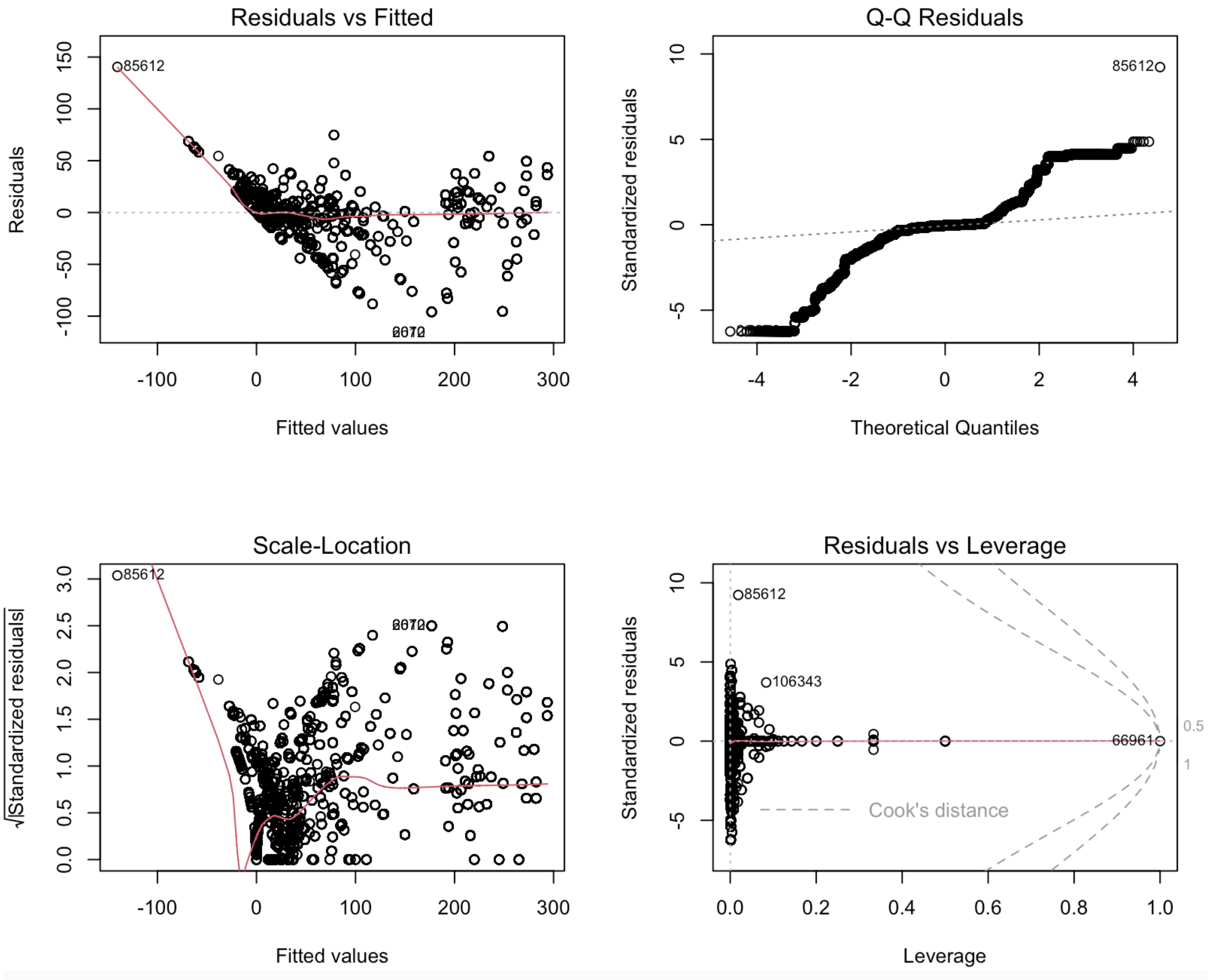
- One of the advantages of using Multivariable Linear Regression is that it gives a simple explanation of the strength of each predictor while ignoring the other factors. This serves us in finding out the one factor among several, which has the greatest impact on the popularity of EVs. In addition to that, Multivariable Linear Regression uses a minimum of computation that is why it is appropriate for large datasets like ours, which has more than 200,000 records. With its flexibility, the model lets us introduce numerous input variables and connections, which provide more information on the data courses and drive data-driven suggestions for EV market research.
- Now, we will start with the Multivariable Linear Regression for our database:

```
# Load necessary libraries
library(tidyverse)

# Load the dataset
ev_data <- read.csv("/mnt/data/Electric_Vehicle_Population_Data.csv")

# Cleaning the data
# Remove rows with missing Electric Range
ev_data_clean <- ev_data %>%
  filter(!is.na(Electric.Range))

# Select relevant columns and handle missing values
# Replace NA in categorical variables with 'Unknown' and numeric columns with
median
predictor_vars <- c("Make", "Model", "Model.Year", "Electric.Vehicle.Type",
"Base.MSRP", "Clean.Alternative.Fuel.Vehicle..CAFV..Eligibility")

ev_data_clean <- ev_data_clean %>%
  mutate(across(all_of(predictor_vars), ~ ifelse(is.na(.), 'Unknown', as.character(.)))) %>%
  mutate(Base.MSRP = ifelse(is.na(Base.MSRP), median(ev_data_clean$Base.MSRP, na.rm = TRUE), Base.MSRP),
         Model.Year = ifelse(is.na(Model.Year), median(ev_data_clean$Model.Year, na.rm = TRUE), Model.Year))

# Convert categorical variables to factors
ev_data_clean <- ev_data_clean %>%
  mutate(across(c(Make, Model, Electric.Vehicle.Type, Clean.Alternative.Fuel.Vehicle..CAFV..Eligibility), as.factor))

# Fit the multivariable linear regression model
model <- lm(Electric.Range ~ Make + Model + Model.Year + Electric.Vehicle.Type + Base.MSRP + Clean.Alternative.Fuel.Vehicle..CAFV..Eligibility, data = ev_data_clean)
# Summarize the model to see the results
summary(model)

# Plot diagnostics to check assumptions
par(mfrow = c(2, 2))
plot(model) # Residuals, QQ plot, Scale-location

# Reset plotting layout
par(mfrow = c(1, 1))
```



```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.37 on 205234 degrees of freedom
Multiple R-squared:  0.9696,    Adjusted R-squared:  0.9696
F-statistic: 3.338e+04 on 196 and 205234 DF,  p-value: < 2.2e-16
```

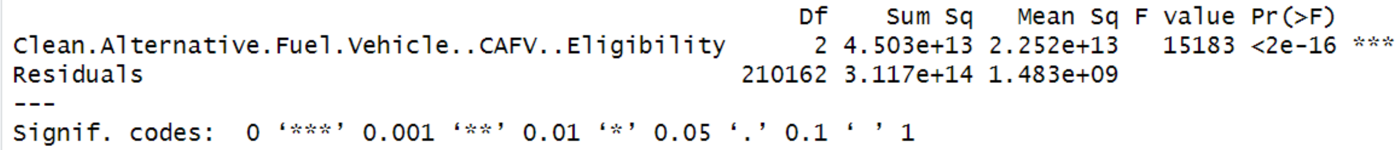
- The output from the Multivariable Linear Regression model shows that there are several major insights that can be drawn. The model gives a high R-squared value (0.9696) that means that about 97% of the change in the electric range is caused by the predictors: Make, Model, Model Year, Electric Vehicle Type, Base MSRP, and CAFV Eligibility. This is a strong indication that the data is appropriately fitted, which means the selected variables are the best in predicting the electric range of the vehicle.
- The codes of significance indicate that many predictors have very low probability values (with p-values less than 0.001), making them highly statistically significant with respect to the dependent variable. The importance of CAFV Eligibility and Base MSRP lies in them being the reasons for differences in electric range for various EV models. The graphs that are used to diagnose the model further allow for a visual evaluation of the model's assumptions, which include residuals and other factors that are required to validate the model. This model is very effective in integrating the electric vehicle range of different elements to gain the most knowledge about the current EV market.



- The Residuals vs Fitted plot shows a bit of a curve, which might mean the model isn't fully capturing all patterns in the data, possibly missing some relationships or showing uneven error spread (heteroscedasticity). The Q-Q plot has points that stray from the line at the ends, suggesting that the residuals aren’t perfectly normal, likely due to a few outliers. In the Scale-Location plot, the spread of errors isn’t consistent, hinting at more issues with uneven error spread. Lastly, the Residuals vs Leverage plot points out some data points with high influence, meaning a few values might have too much impact on the model.

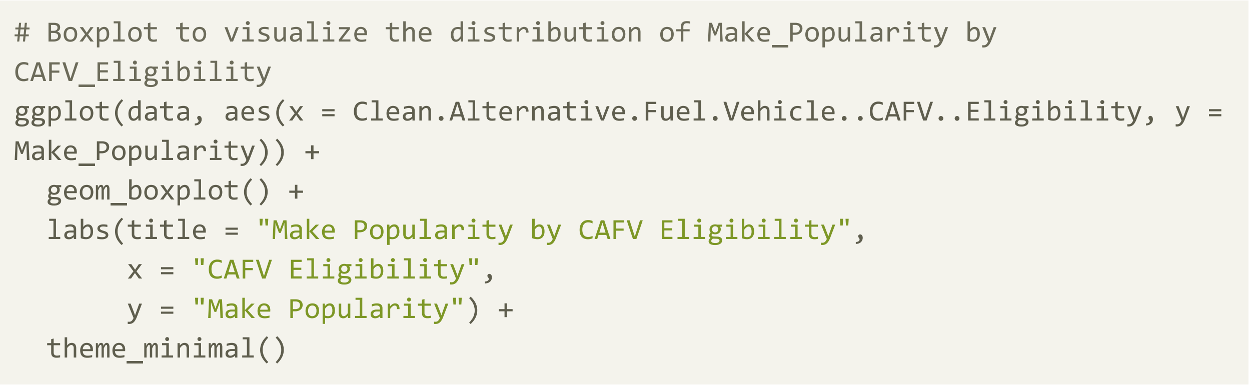
- **ANOVA**
  - We choose the ANOVA model because it is designed to compare the means of a continuous dependent variable, Make_Popularity, across different categories of a categorical independent variable, Clean.Alternative.Fuel.Vehicle..CAFV..Eligibility. In this case, ANOVA helps test whether there are significant differences in the popularity of vehicle makes between different groups. Since we are analyzing a single dependent variable and categorical groups, ANOVA is the appropriate method to determine if CAFV eligibility affects popularity.
  - The advantage of using ANOVA for this specific question is that it allows you to determine if there is a statistically significant difference in the mean popularity of vehicle makes across different groups of the categorical variable, Clean Alternative Fuel Vehicle (CAFV) Eligibility. Since ANOVA tests the differences in means, it helps to assess if the CAFV eligibility status (e.g., "Eligible" vs "Not Eligible") has a meaningful impact on the average popularity of different vehicle makes. This approach is straightforward and efficient when you have one dependent variable (popularity) and one categorical factor (CAFV eligibility), making it ideal for this analysis. Additionally, ANOVA can handle multiple groups (e.g., more than two eligibility statuses), providing a clear method to test for differences across them.
  - Since the p-value for CAFV_Eligibility is less than 0.05, we reject the null hypothesis and conclude that CAFV Eligibility contributes to the popularity of EV make.

```r
# Load necessary libraries
library(tidyverse)

# Load dataset
data <-
read.csv("C:/Users/haotr/Downloads/Electric_Vehicle_Population_Data.csv")
data$Make_Popularity <- as.numeric(table(data$Make)[data$Make])

# Fit the MANOVA model
anova_model <- aov(Make_Popularity ~
Clean.Alternative.Fuel.Vehicle..CAFV..Eligibility, data = data)

# Summary of the MANOVA model
summary(anova_model)
```

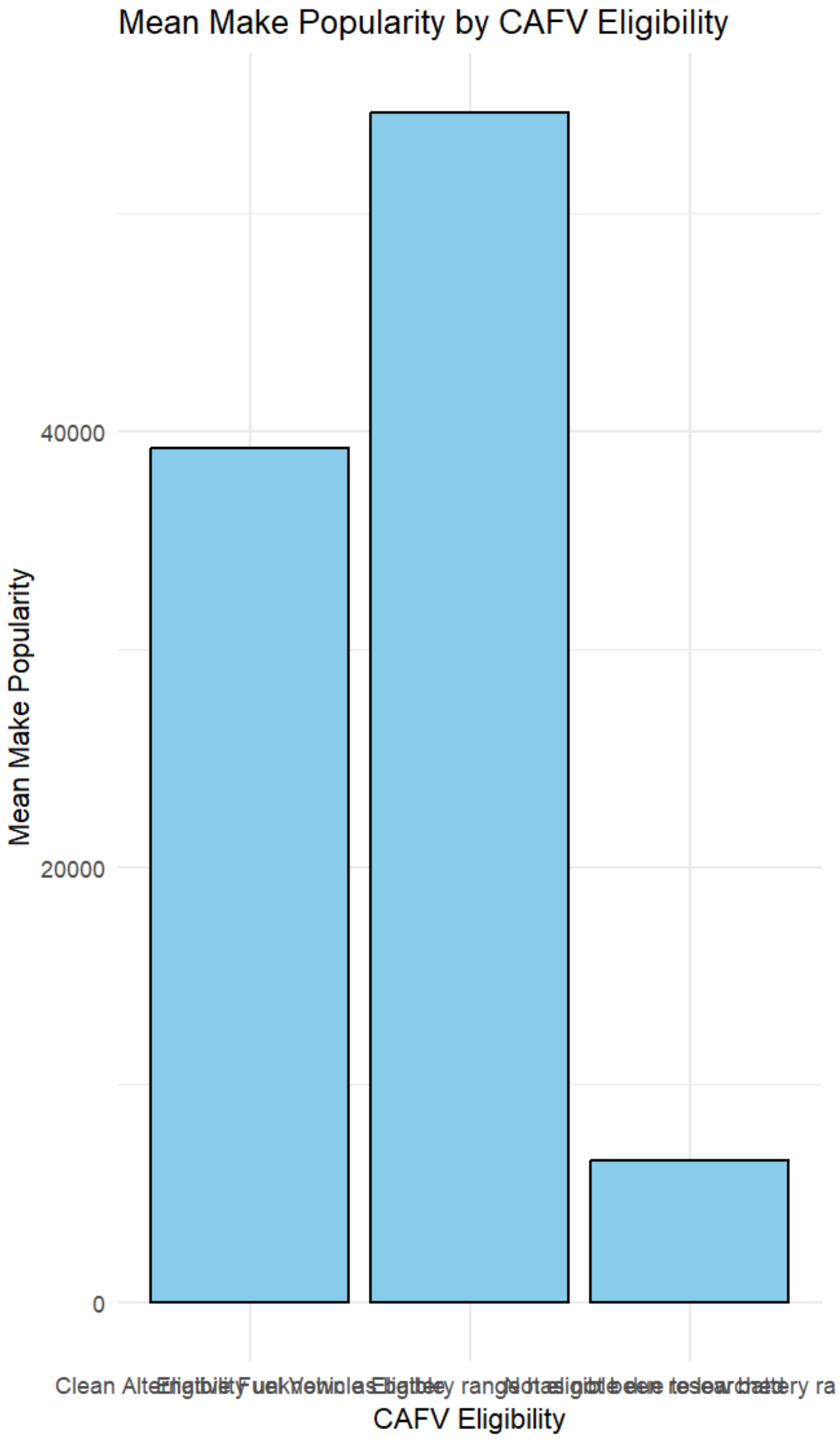


```
                                               Df    Sum Sq   Mean Sq F value Pr(>F)    
Clean.Alternative.Fuel.Vehicle..CAFV..Eligibility      2 4.503e+13 2.252e+13   15183 <2e-16 ***
Residuals                                         210162 3.117e+14 1.483e+09                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The F-value for CAFV Eligibility is 15183, with a very low p-value (< 2e-16), which is much smaller than the usual significance level of 0.05. This means we reject the null hypothesis, suggesting that there are significant differences in Make Popularity across different levels of CAFV Eligibility.
- The high F-value implies the CAFV Eligibility significantly affects the popularity of vehicle makes which means that the vehicle eligibility status (e.g., "Eligible," "Not Eligible") is one of the influential factors in the popularization of electric vehicle makes. According to the model, differences in CAFV Eligibility also have a substantial impact on the popularity of certain vehicle makes within the dataset. This serves as a pro as to the fact that the state of CAFV Eligibility is a key variable in the discovery of the trends in the popularity among electric vehicle brands
- To support the result of the ANOVA test, we have provided Box Plot to visualize the distribution of Make_Popularity by CAFV_Eligibility. According to the plot, the popularity of vehicles that are eligible for CAFV is way more than the ones that are not eligible.



```
# Boxplot to visualize the distribution of Make_Popularity by
CAFV_Eligibility
ggplot(data, aes(x = Clean.Alternative.Fuel.Vehicle..CAFV..Eligibility, y =
Make_Popularity)) +
  geom_boxplot() +
  labs(title = "Make Popularity by CAFV Eligibility",
       x = "CAFV Eligibility",
       y = "Make Popularity") +
  theme_minimal()
```



- The bar plot shows the average popularity of different car makes based on their CAFV Eligibility status. We can see that vehicles marked as "Eligible" for clean alternative fuel have the highest average popularity, with "Unknown" status coming next. In comparison, vehicles marked as "Not Eligible" have a much lower average popularity. This matches what we found with ANOVA, suggesting that CAFV Eligibility plays a big role in how popular a car is. It seems like people prefer electric vehicles that qualify for clean fuel options.

Conclusion

Compare: Given the dataset, which includes a mix of categorical and continuous variables, the multivariable regression model is more versatile and informative. It allows you to: Include both categorical and continuous variables. Quantify the effect of each predictor, whether categorical or continuous. Capture complex relationships, such as interactions between predictors. While ANOVA can be useful for specific questions, such as determining whether there is a significant difference in **Electric Range** between different **Makes**. However, since ANOVA is more limited in scope and cannot handle continuous predictors, it provides only partial insights compared to what a regression model can offer.

Endnotes

All members of this group have worked on each part together to ensure the correctness of each method used in this project.

References

State of California - Number of Cancer Surgeries (Volume) Performed in California Hospitals, Publisher Department of Health Care Access and Information, 2 Dec. 2023, catalog.data.gov/dataset/number-of-cancer-surgeries-volume-performed-in-california-hospitals-a3f18.