CS555 – Intoduction to Machine Learning

FINAL PROJECT

Department of Computer Science

Student ID – U42035592

Professor
- Ming Zhang

Submitted by
- Akanksha Ankam

Table of Contents

Content pg.no

Abstract:

The rapid growth of the electric vehicle (EV) market has sparked increased interest in understanding the factors that influence EV performance and what factors influence the Electric Range of the Vehicles. To determine the relationships, we analyzed the dataset using statistical tools, including multiple regression, analysis of variance, and regression diagnostics. The statistical package R was employed to process, analyze, and visualize the data.

Our study provides strong evidence that understanding the relationships and how they can contribute to a more sustainable future. These insights can help consumers and manufacturers make informed decisions about electric vehicle production and purchases, ultimately promoting an eco-friendlier transportation sector. By utilizing advanced statistical techniques and tools, we were able to uncover valuable information that can guide policymakers and industry stakeholders in refining incentive programs and encouraging the development and adoption of electric vehicles with higher electric ranges.

Introduction:

The electric vehicle (EV) market has experienced rapid growth in recent years, fueled by increasing environmental concerns and advancements in battery technology. Governments and regulatory agencies worldwide have introduced various incentive programs to encourage the adoption of EVs, such as the Clean Alternative Fuel Vehicle (CAFV) program.

The CAFV program aims to promote the use of cleaner vehicles by offering incentives to eligible low-emission vehicles. Understanding the factors that influence the electric range of EVs and their eligibility for such programs is crucial for policymakers and industry stakeholders to ensure the effectiveness of these incentives and drive further EV adoption.

This project seeks to investigate the relationship between the electric range of electric vehicles and their eligibility for the CAFV program. Additionally, it aims to explore the influence of the make and model year of the car on the electric range. Identifying the key factors affecting EV performance and their connection to incentive programs will provide valuable insights that can help shape the design of future policies and initiatives in the EV industry.

By employing multiple linear regression, we will analyze the impact of CAFV eligibility, and model year on the electric range of EVs. The results of this analysis will not only offer a better understanding of the relationship between electric range and CAFV eligibility but also highlight the significance of the car's make and model year in determining electric range.

Consequently, this information can be used to guide policymakers and industry stakeholders in designing more effective incentive programs and targeting specific model years or checking for environment friendly vehicle eligibility for improvement, ultimately accelerating the transition towards a more sustainable transportation sector.

1. Research Scenario:

The aim of this research is to investigate the impact of the make and model year of electric vehicles on their electric range. The study will be conducted using a sample dataset that includes information on electric vehicles, such as VIN, make, model, model year, electric range, and Clean Alternative Fuel Vehicle (CAFV) eligibility, among other variables. By understanding how the make and model year affect the electric range, policymakers and industry stakeholders can make informed decisions about incentives and vehicle improvements that can further encourage the adoption of electric vehicles.

Research Question:

What factors influence the electric range of an electric vehicle and to what extent do they affect it?

2. Describing the dataset:

The dataset provided contains information on a sample of electric vehicles, including Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs). The data includes a range of attributes for each vehicle, such as Vehicle Identification Number (VIN), county, city, state, postal code, model year, make, model, electric vehicle type, Clean Alternative Fuel Vehicle (CAFV) eligibility, electric range, base MSRP, legislative district, Department of Licensing (DOL) vehicle ID, vehicle location (latitude and longitude), electric utility, and 2020 Census Tract.

For the analysis, the following variables from the dataset will be used:

Model Year:
This variable represents the year the electric vehicle was manufactured. It will be used to examine trends in electric range over time and identify any improvements across different model years.
Make:
This variable indicates the manufacturer of the electric vehicle (e.g., TESLA, VOLVO, BMW, NISSAN, KIA, CHEVROLET). It will be used to compare the electric ranges of vehicles produced by different manufacturers.
Model:
This variable represents the specific model of the electric vehicle (e.g., MODEL 3, S60, X5, LEAF, NIRO, SOUL, VOLT). It will help in identifying any patterns or trends within a particular make and model.

Electric Vehicle Type:
This variable specifies whether the electric vehicle is a Battery Electric Vehicle (BEV) or a Plug-in Hybrid Electric Vehicle (PHEV). This information will allow for a comparison of electric ranges between the two types of electric vehicles.

Electric Range:
This variable indicates the estimated electric range of the vehicle, measured in miles. It is the primary variable of interest in this study, as the analysis aims to explore the relationship between make, model year, and electric range.

By focusing on these variables, the analysis will be able to determine the extent to which the make and model year of an electric vehicle affect its electric range.

Data Cleaning:

Performed the following data cleaning steps on the dataset:

Removing Duplicates:
Checked for duplicate records using unique identifiers such as VIN or DOL Vehicle ID and removed any duplicate entries to ensure accurate analysis.

Removing Empty Spaces:
Identified and removed any unnecessary empty spaces in the dataset, which helps in maintaining consistency and avoiding errors during the analysis.

Ensuring Data Consistency:
Checked that the data is consistent, particularly in terms of formats and units, ensuring that the electric range is consistently reported in miles and the model year is uniformly represented as a four-digit year.

Cleaning Categorical Variables:
Ensured that the naming conventions for categorical variables like make and model are consistent (e.g., all uppercase or lowercase, no extra spaces) and grouped any variations of the same make or model together.

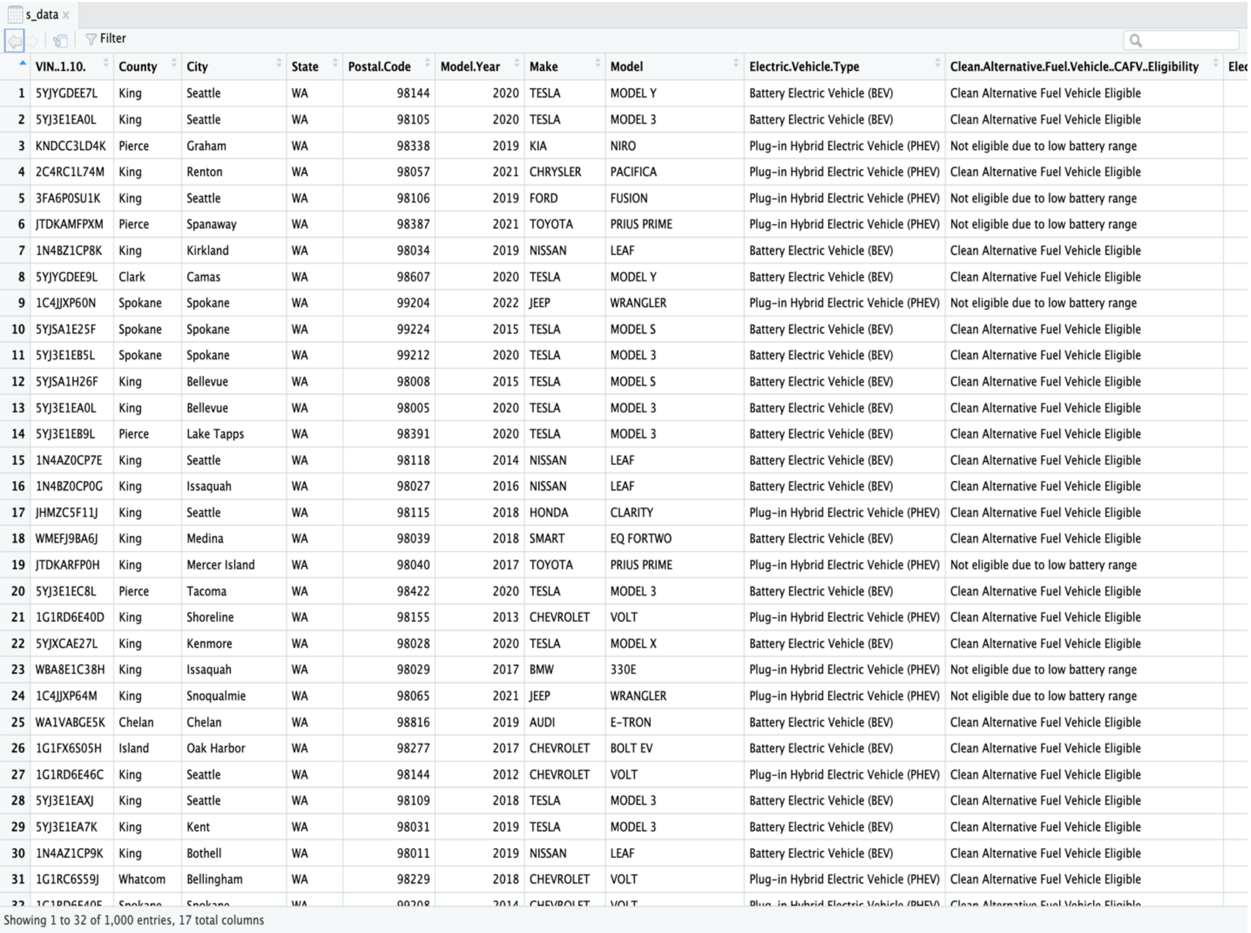
By performing these data cleaning steps, I have prepared the dataset for further analysis and exploration to answer the research question.

The link to the main data set source –
https://www.kaggle.com/datasets/ratikkakkar/electric-vehicle-population-data

→ In this analysis, we will be using a random sample of 1,000 rows from our original dataset of 73,000 rows. This sampling method allows us to reduce the computational burden and processing time while still obtaining reliable and representative insights into the relationships between variables in our dataset.

```
getwd()
install.packages("ggplot2")
library(ggplot2)
install.packages("dplyr")
library(dplyr)
data <- read.csv("Electric_Vehicle_Population_Data.csv", header = TRUE)
View(data)
set.seed(1000) #setting seed to 1000, so that we choose the same random variables every time.
s_data <- sample_n(data, 1000)
s_data
View(s_data)
attach(s_data)
```

This is the sample dataset... (this is not the entire dataset but providing a small view of it for understanding).



	VIN..1.10.	County	City	State	Postal.Code	Model.Year	Make	Model	Electric.Vehicle.Type	Clean.Alternative.Fuel.Vehicle..CAFV..Eligibility	Elec
1	5YJYGDEE7L	King	Seattle	WA	98144	2020	TESLA	MODEL Y	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	
2	5YJ3E1EA0L	King	Seattle	WA	98105	2020	TESLA	MODEL 3	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	
3	KNDCC3LD4K	Pierce	Graham	WA	98338	2019	KIA	NIRO	Plug-in Hybrid Electric Vehicle (PHEV)	Not eligible due to low battery range	
4	2C4RC1L74M	King	Renton	WA	98057	2021	CHRYSLER	PACIFICA	Plug-in Hybrid Electric Vehicle (PHEV)	Clean Alternative Fuel Vehicle Eligible	
5	3FA6P0SU1K	King	Seattle	WA	98106	2019	FORD	FUSION	Plug-in Hybrid Electric Vehicle (PHEV)	Not eligible due to low battery range	
6	JTDKAMFPXM	Pierce	Spanaway	WA	98387	2021	TOYOTA	PRIUS PRIME	Plug-in Hybrid Electric Vehicle (PHEV)	Not eligible due to low battery range	
7	1N4BZ1CP8K	King	Kirkland	WA	98034	2019	NISSAN	LEAF	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	
8	5YJYGDEE9L	Clark	Camas	WA	98607	2020	TESLA	MODEL Y	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	
9	1C4JJXP60N	Spokane	Spokane	WA	99204	2022	JEEP	WRANGLER	Plug-in Hybrid Electric Vehicle (PHEV)	Not eligible due to low battery range	
10	5YJSA1E25F	Spokane	Spokane	WA	99224	2015	TESLA	MODEL S	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	
11	5YJ3E1EB5L	Spokane	Spokane	WA	99212	2020	TESLA	MODEL 3	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	
12	5YJSA1H26F	King	Bellevue	WA	98008	2015	TESLA	MODEL S	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	
13	5YJ3E1EA0L	King	Bellevue	WA	98005	2020	TESLA	MODEL 3	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	
14	5YJ3E1EB9L	Pierce	Lake Tapps	WA	98391	2020	TESLA	MODEL 3	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	
15	1N4AZ0CP7E	King	Seattle	WA	98118	2014	NISSAN	LEAF	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	
16	1N4BZ0CP0G	King	Issaquah	WA	98027	2016	NISSAN	LEAF	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	
17	JHMZC5F11J	King	Seattle	WA	98115	2018	HONDA	CLARITY	Plug-in Hybrid Electric Vehicle (PHEV)	Clean Alternative Fuel Vehicle Eligible	
18	WMEFJ9BA6J	King	Medina	WA	98039	2018	SMART	EQ FORTWO	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	
19	JTDKARFP0H	King	Mercer Island	WA	98040	2017	TOYOTA	PRIUS PRIME	Plug-in Hybrid Electric Vehicle (PHEV)	Not eligible due to low battery range	
20	5YJ3E1EC8L	Pierce	Tacoma	WA	98422	2020	TESLA	MODEL 3	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	
21	1G1RD6E40D	King	Shoreline	WA	98155	2013	CHEVROLET	VOLT	Plug-in Hybrid Electric Vehicle (PHEV)	Clean Alternative Fuel Vehicle Eligible	
22	5YJXCAE27L	King	Kenmore	WA	98028	2020	TESLA	MODEL X	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	
23	WBA8E1C38H	King	Issaquah	WA	98029	2017	BMW	330E	Plug-in Hybrid Electric Vehicle (PHEV)	Not eligible due to low battery range	
24	1C4JJXP64M	King	Snoqualmie	WA	98065	2021	JEEP	WRANGLER	Plug-in Hybrid Electric Vehicle (PHEV)	Not eligible due to low battery range	
25	WA1VABGE5K	Chelan	Chelan	WA	98816	2019	AUDI	E-TRON	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	
26	1G1FX6S05H	Island	Oak Harbor	WA	98277	2017	CHEVROLET	BOLT EV	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	
27	1G1RD6E46C	King	Seattle	WA	98144	2012	CHEVROLET	VOLT	Plug-in Hybrid Electric Vehicle (PHEV)	Clean Alternative Fuel Vehicle Eligible	
28	5YJ3E1EAXJ	King	Seattle	WA	98109	2018	TESLA	MODEL 3	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	
29	5YJ3E1EA7K	King	Kent	WA	98031	2019	TESLA	MODEL 3	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	
30	1N4AZ1CP9K	King	Bothell	WA	98011	2019	NISSAN	LEAF	Battery Electric Vehicle (BEV)	Clean Alternative Fuel Vehicle Eligible	
31	1G1RC6S59J	Whatcom	Bellingham	WA	98229	2018	CHEVROLET	VOLT	Plug-in Hybrid Electric Vehicle (PHEV)	Clean Alternative Fuel Vehicle Eligible	
32	1G1RD6E40E	Spokane	Spokane	WA	99208	2014	CHEVROLET	VOLT	Plug-in Hybrid Electric Vehicle (PHEV)	Clean Alternative Fuel Vehicle Eligible	

Showing 1 to 32 of 1,000 entries, 17 total columns

3. Describe the statistical methods you plan to use.

Briefly describe the statistical methods you will utilize to investigate your research question(s).

In this project, a variety of statistical methods were employed to analyze the dataset and investigate the relationship between the predictors.

- Data summarization: The dataset was summarized to understand its central tendency, variability, and distribution, including identifying missing values and outliers.
- Boxplots: Boxplots were generated to visualize the distribution of variables, providing insights into the minimum, first quartile, median, third quartile, and maximum values.
- Scatterplots: Scatterplots were created to visualize the relationship between predictors and the response variable, allowing for the assessment of overall trends or patterns.
- Multiple Linear Regression: Multiple Linear Regression analysis was conducted to estimate regression coefficients and assess the statistical significance of predictors' associations with the electric range.
- Residual plot: A residual plot was generated to evaluate the adequacy of the regression model and identify any patterns or deviations from model assumptions.
- Interaction plot: An interaction plot was generated to explore the interaction effect between two predictors on the response variable. This plot allowed for the visual examination of how the relationship between predictors and the response variable varied depending on the values of the other predictor. It helped identify whether there was a significant interaction effect and provided insights into the nature of the interaction.
- Two-way ANOVA: A Two-way ANOVA (Analysis of Variance) was conducted to examine the effects of two predictors simultaneously on the response variable. This statistical analysis helped determine whether there were significant main effects of each predictor and whether there was a significant interaction effect between the predictors. The results of the Two-way ANOVA provided valuable information on the relative importance of each predictor and their combined influence on the response variable.
- Least squares regression: A least squares regression method was used to estimate the best-fit line representing the relationship between predictors and the response variable.

- The data was summarized to gain an understanding of its central tendency, variability, and distribution. This allowed for a comprehensive overview of the dataset and helped identify any missing values or outliers that might affect the analysis.
- To visualize the distribution of the variables, boxplots were generated. These graphical representations provided valuable insights into the minimum, first quartile, median, third quartile, and maximum values of the variables, enabling quick comparisons and the identification of any potential patterns or discrepancies.

- Scatterplots were created to visualize the relationship between the predictors and the response variable. These plots allowed for the assessment of any overall trends or patterns, such as linear or non-linear associations, between the variables.
- To further investigate the relationship between the predictors and the response variable, Multiple Linear Regression analysis was conducted. This statistical technique estimated the regression coefficients for each predictor, enabling quantification of the association between the predictors and the response variable. The statistical significance of these coefficients was assessed to determine whether the predictors were significantly associated with the electric range.
- To evaluate the adequacy of the regression model, a residual plot was generated. This plot visually displayed the differences between the observed and predicted values (residuals) and helped identify any patterns or systematic deviations from the assumptions of the model. A random pattern in the residual plot indicated that the model assumptions were met.
- A least squares regression method was employed to estimate the best-fit line that represented the relationship between the predictors and the response variable. This line allowed for prediction and interpretation of the results, providing valuable insights into the impact of the predictors on the electric range of vehicles.
- By performing the interaction plot and Two-way ANOVA, a deeper understanding of the relationships and interactions between predictors was gained.
- These analyses contributed to uncovering complex patterns and providing more comprehensive insights into the factors influencing the response variable.

The application of these statistical methods provided a comprehensive analysis of the dataset and enabled meaningful conclusions to be drawn regarding the relationship between the predictors and the electric range. The findings from this analysis will contribute to the broader understanding of the impact of the make and model year of electric vehicles on their electric range, which is essential for making informed decisions regarding incentives and vehicle improvements to promote electric vehicle adoption.

4. Report your results.

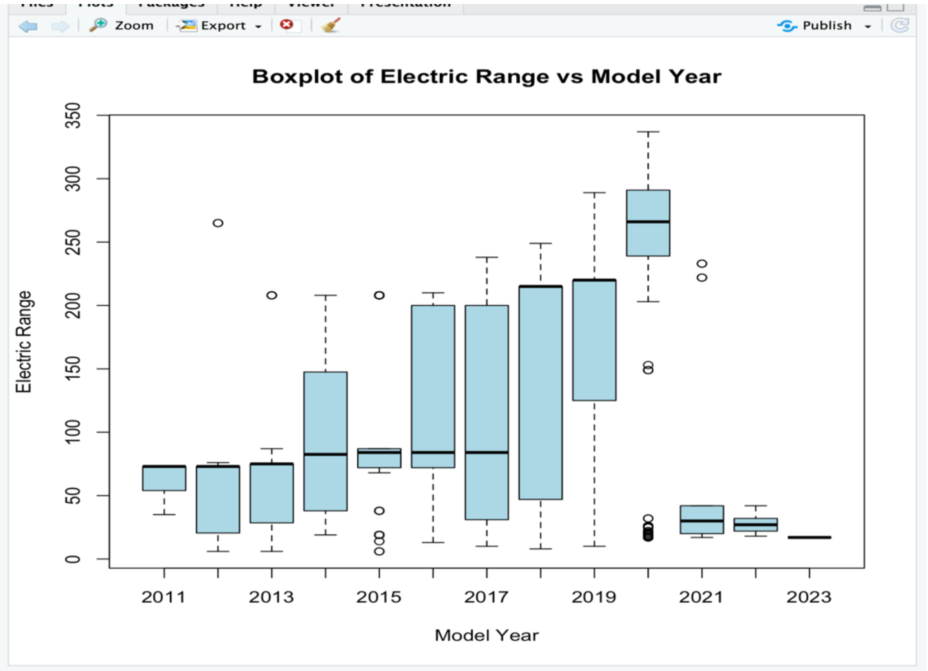
Write up the results of your analysis. You should present tables and figures when relevant, and you should have a short write-up describing your results.

Summarize the data:

```
summary(s_data$Model.Year)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2011    2016    2018    2018    2019    2023 

> summary(s_data$Electric.Range)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     6      32      93     131     215     337 
```

→ Summarizing the data using boxplot

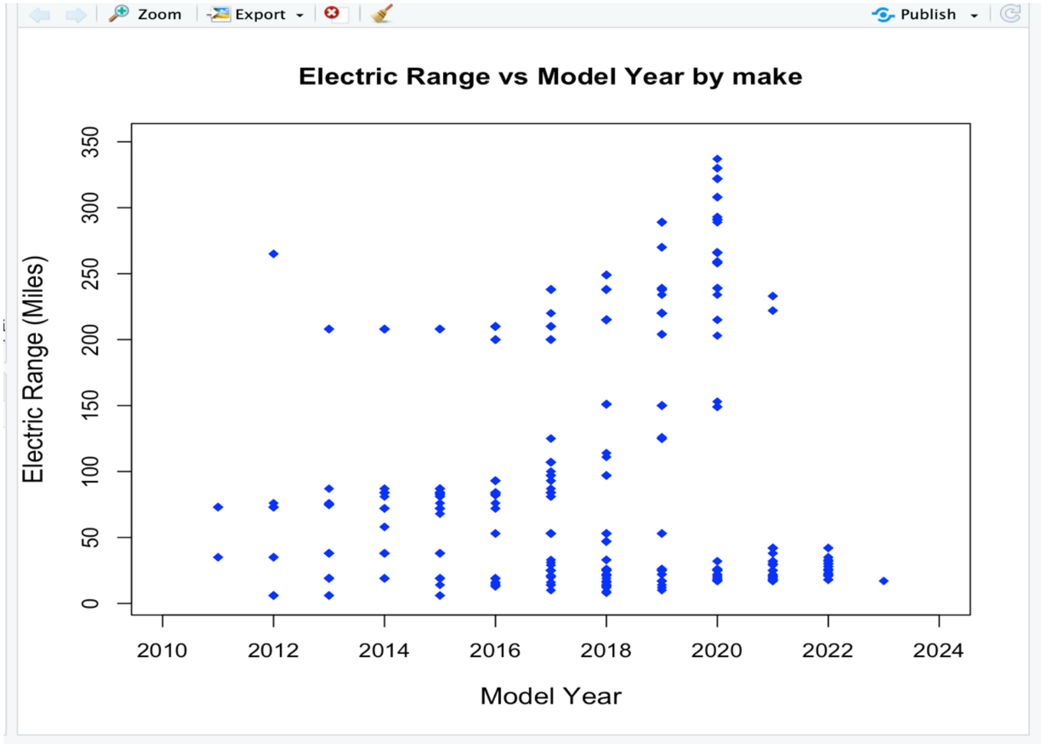
```
#Boxplot
boxplot(s_data$Electric.Range ~ s_data$Model.Year, data=data,
        main="Boxplot of Electric Range by Make",
        xlab="Model Year", ylab="Electric Range",
        col="lightblue", border="black")
```



In our analysis of the boxplots of Electric Range by Model Year, we observed that the electric range of electric vehicles appears to be affected by the model year. There are notable differences in electric range across various model years, with periods of improvement, stagnation, and decline.

→Generating Scatterplot

```
#scatterplot
plot(Model.Year, Electric.Range,
   main="Electric Range vs Model Year by make",
   xlab="Model Year",
   ylab="Electric Range (Miles)", xlim=c(2010,2024),
   ylim=c(5,350), pch=18, col="blue", cex.lab=1.2)
```



From the scatterplot "Electric Range vs Model Year by make" displays it is weakly Linear, and the direction is Positively associated relationship between the model year of electric vehicles and their electric range in miles. It can be observed that newer electric vehicle models tend to have a higher electric range compared to older models, but other factors might also be influencing the electric range, such as Model of the vehicle. So, the strength of the association between the variables is weak. Hence, it is Linear form, positively associated direction and has weak strength observed between the variables.

→Finding the Correlation coefficient

```
cor(Model.Year,Electric.Range)
[1] 0.2228386
```

The correlation coefficient between Model.Year and Electric.Range is 0.2228386. This value indicates a weak positive linear relationship between the model year of electric vehicles and their electric range. As the model year increases, there is a slight tendency for the electric range to increase as well.

→ Let's see how the Model Year of electric vehicles and Clean Alternative Fuel Vehicle (CAFV) eligibility impacts the Electric Range of the vehicle.
For this we need to perform *Multiple Linear Regression Analysis.*

```
# Create dummy variables for Electric Vehicle Type
> ev_type_dummies <- model.matrix(~ s_data$Electric.Vehicle.Type - 1)
> CAFV_Eligible <- ifelse(s_data$Clean.Alternative.Fuel.Vehicle..CAFV..Eligibility
+                     == "Clean Alternative Fuel Vehicle Eligible", 1, 0)
> mm <- lm(s_data$Electric.Range ~ s_data$Model.Year
+        + s_data$Clean.Alternative.Fuel.Vehicle..CAFV..Eligibility,
+        data = s_data)
> summary(mm)
Call:
lm(formula = s_data$Electric.Range ~ s_data$Model.Year +
s_data$Clean.Alternative.Fuel.Vehicle..CAFV..Eligibility,
   data = s_data)

Residuals:
    Min      1Q   Median      3Q      Max
-184.129  -50.695    6.887   53.592  163.739

Coefficients:
                                        Estimate     Std. Error    t value       Pr(>|t|)
(Intercept)                             -2.261e+04   1.818e+03     -12.44        <2e-16 ***
s_data$Model.Year                        1.129e+01   9.009e-01      12.53        <2e-16 ***
s_data$Clean.Alternative.Fuel.          -1.509e+02    5.510e+00    -27.38        <2e-16 ***
Vehicle..CAFV..EligibilityNot
eligible due to low battery range
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 72.2 on 997 degrees of freedom
Multiple R-squared:  0.4575,  Adjusted R-squared:  0.4564
F-statistic: 420.5 on 2 and 997 DF,  p-value: < 2.2e-16
```

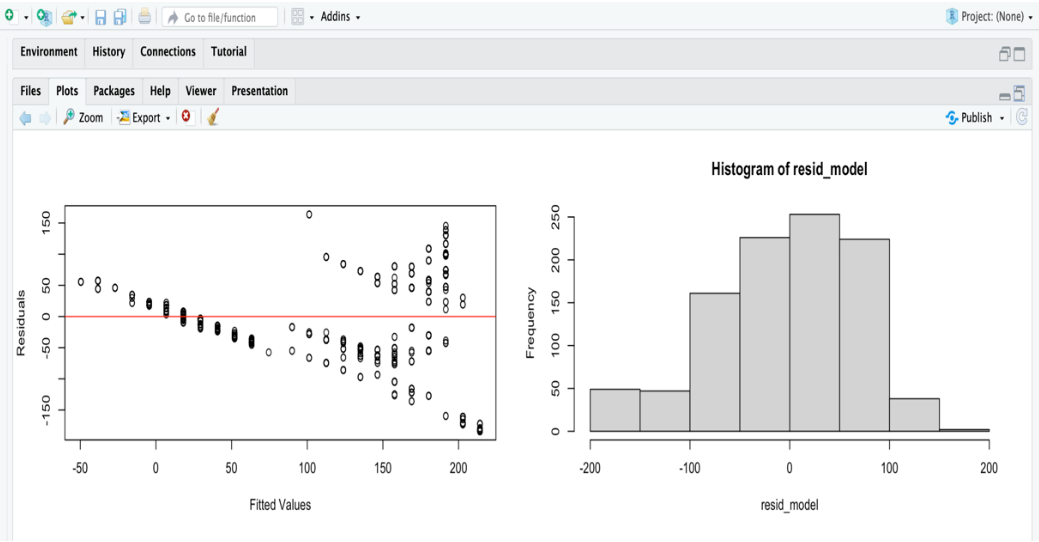
Based on the Multiple regression analysis, it is observed that vehicles not eligible for Clean Alternative Fuel Vehicle (CAFV) status due to low battery range have significantly lower estimated electric range compared to the vehicles that are Clean Alternative Fuel Vehicle Eligible.

The coefficient for the "Not eligible due to low battery range" category (-1.509e+02) suggests that, on average, these vehicles have an electric range that is approximately 150.9 miles lower than the vehicles eligible for CAFV status. This finding highlights the impact of CAFV eligibility on electric range and suggests that vehicles meeting the criteria for CAFV eligibility tend to have a higher electric range.

This observation aligns with the understanding that vehicles eligible for CAFV status typically have larger battery capacities or other features that allow for a longer electric range. It implies that the eligibility for CAFV status, which considers factors like battery range, plays a role in determining the electric range of electric vehicles.

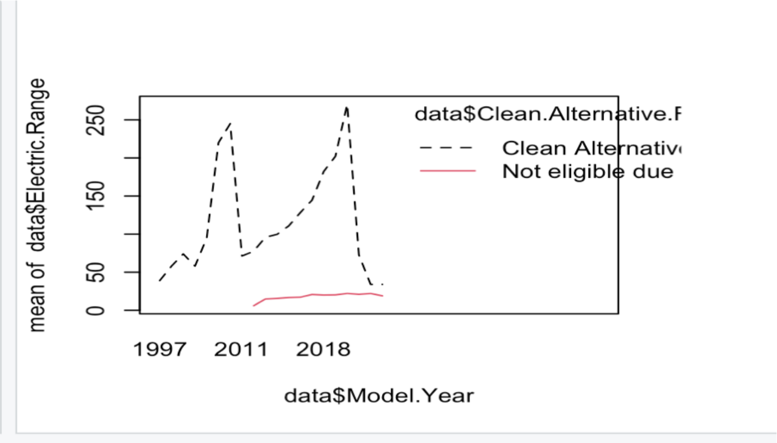
→Generating Residual Plot

```
resid_model <- resid(model)
fitted_m <- fitted(model)

par(mfrow = c(2, 2))
plot(fitted_m, resid_model, axes = TRUE, frame.plot = TRUE, xlab = 'Fitted Values', ylab =
'Residuals')
abline(h = 0, col = "red")
hist(resid_model)
```



→ Performing interaction plot, to decide whether we can perform two way anova or not. Before performing a two-way ANOVA, we must check if an interaction is present. We do this by including an interaction term in the two-way ANOVA model.

```
install.packages("car")
library(car)
m <- lm(data$Electric.Range ~ data$Model.Year +
data$Clean.Alternative.Fuel.Vehicle..CAFV..Eligibility
      + data$Model.Year * data$Clean.Alternative.Fuel.Vehicle..CAFV..Eligibility)
summary(m)
Anova(m, type=3)

Anova Table (Type III tests)

Response: data$Electric.Range
                                        Sum Sq        Df          F value         Pr(>F)
(Intercept)                             67360127       1          12341.7       < 2.2e-16 ***
data$Model.Year                         68191250       1          12494.0       < 2.2e-16 ***
data$Clean.Alternative.
Fuel.Vehicle..CAFV..Eligibility          12738800       1           2334.0      < 2.2e-16 ***
data$Model.Year
:data$Clean.Alternative.Fuel.
Vehicle..CAFV..Eligibility               12892662      1            2362.2      < 2.2e-16 ***
Residuals                                      398707379   73051
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> interaction.plot(data$Model.Year,
data$Clean.Alternative.Fuel.Vehicle..CAFV..Eligibility,data$Electric.Range, col=1:2)
```



We conducted an analysis to understand the relationship between electric vehicle range, model year, and Clean Alternative Fuel Vehicle (CAFV) eligibility. We used a linear regression model with an interaction term between model year and CAFV eligibility to explore whether the effect of one variable depends on the other. The interaction plot showed no significant interaction between the two factors, indicating that the relationship between model year and electric range is consistent across different levels of CAFV eligibility. As a result, we can perform a two-way ANOVA to further investigate the main effects of model year and CAFV eligibility on electric vehicle range.

→ Two-way Anova

Step-1:
State the null and alternative hypotheses:
H0: There is no main effect of Model Year on Electric Range.
H1: There is a main effect of Model Year on Electric Range.
H0: There is no main effect of CAFV Eligibility on Electric Range.
H1: There is a main effect of CAFV Eligibility on Electric Range.

Step-2:
Choose the level of significance (alpha):
The level of significance is α = 0.05 to determine the cutoff for the p-value in the ANOVA table.

Step-3:
Test Statistic

```
alpha <- 0.05
df1 <- 1
df2 <- 998
qf(0.95, df1, df2)
[1] 3.850793
```

If f value is F > 3.850793
Then reject H0, else do not reject H0.

Step-4:

```
two_way_aov <- aov(data$Electric.Range ~ data$Model.Year +
data$Clean.Alternative.Fuel.Vehicle..CAFV..Eligibility, data = s_data)
summary(two_way_aov)

                                      Df     Sum Sq       Mean Sq      F value      Pr(>F)
data$Model.Year                        1     40732835     40732835     7229      <2e-16 ***
data$Clean.Alternative
.Fuel.Vehicle..CAFV..Eligibility       1     259198915    259198915    46003     <2e-16 ***
```

```
Residuals                              73052     411600042      5634
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

F- model year = 7229
F- CAFV eligibility = 46003

Step-5:
Conclusion
As, F values are greater than 3.8507
We reject H0.

Model Year:
Calculated F-value: 7229
Critical F-value: 3.8507
CAFV Eligibility:
Calculated F-value: 46003
Critical F-value: 3.8507
In both cases, the calculated F-values are much larger than the critical F-value. This means that the relationships between Model Year and Electric Range, and CAFV Eligibility and Electric Range are statistically significant at the 0.05 significance level.
This means that the relationships between these factors and Electric Range are not due to random chance, and there is evidence to support the claim that these factors have an impact on Electric Range.

→ **Testing** (using the 5-step procedure) whether the set of these predictors is associated with a Electric Range at the α = 0.05 level.

As we have the summary of the data from the above, Now, calculating the least squared regression using 5 step procedure:

Step-1:
State Null and alternate hypothesis,

H0; H0: β1 = β2 = β3 = 0
Ha; atleast one of β1, β2 or β3 not equal to zero.
That mean at least one will be associated with Electric Range.
α = 0.05

Step-2:
Select the appropriate test statistic!
The test statistic for the global test is the F-statistic, which follows an F-distribution with (p-1, n-p) degrees of freedom, where p is the number of predictors and n is the sample size.

df1=2, df2=997.

Step-3:
State the decision rule, Specify the significance level that is α = 0.05.

```
qf(0.95,df1=2,df2=997)
[1] 3.004752
```

Decision Rule: Reject $H0$ if $F \geq 3.004752$.
Otherwise, do not reject $H0$.

Step-4:
Compute the test statistic,

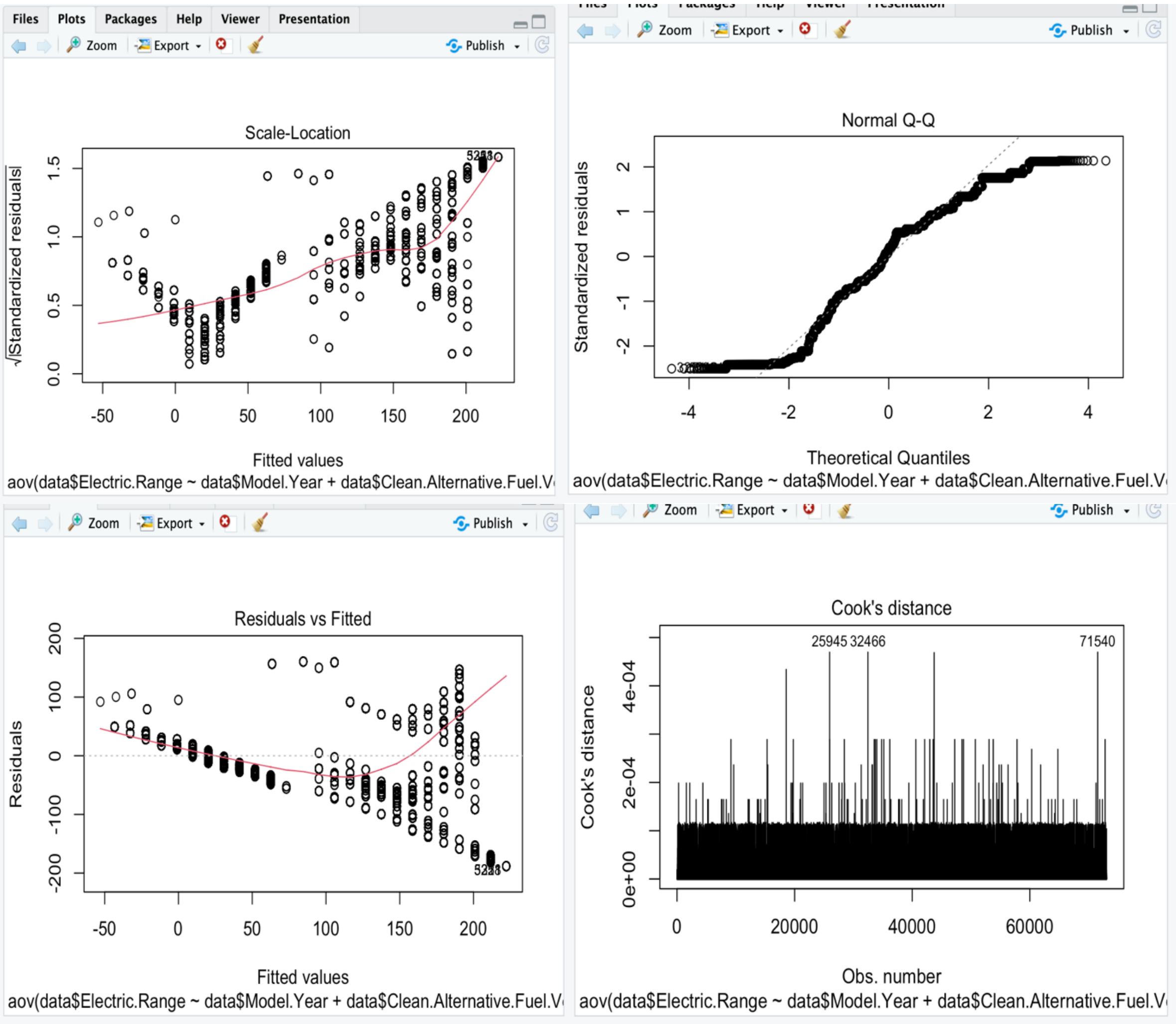
```
summary(model)
>lm(formula = s_data$Electric.Range ~ s_data$Model.Year +
s_data$Clean.Alternative.Fuel.Vehicle..CAFV..Eligibility,
   data = s_data)

Residuals:
    Min       1Q   Median      3Q      Max
-184.129  -50.695    6.887   53.592  163.739

Coefficients:
                                        Estimate     Std. Error    t value      Pr(>|t|)
(Intercept)                             -2.261e+04   1.818e+03     -12.44       <2e-16 ***
s_data$Model.Year                        1.129e+01   9.009e-01      12.53       <2e-16 ***
s_data$Clean.Alternative.Fuel.          -1.509e+02    5.510e+00    -27.38       <2e-16 ***
Vehicle..CAFV..EligibilityNot
eligible due to low battery range
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 72.2 on 997 degrees of freedom
Multiple R-squared:  0.4575,  Adjusted R-squared:  0.4564
F-statistic: 420.5 on 2 and 997 DF,  p-value: < 2.2e-16
```

F-statistics=Mean SS of regression/ Mean SS of residual=420.5
F-statistic is 420.5.

Step-5:
Conclusion
Reject H0 as 420.5 > 3.004752
So, we reject H0.

There is sufficient statistical evidence at 5% level of significance to conclude that the set of these predictors are associated with Electric Range.

→ Lets plot Residual plots for better understanding the pattern.
plot(two_way_aov, 1:4)
Hit <Return> to see next plot: 1
Hit <Return> to see next plot: 2
Hit <Return> to see next plot: 3
Hit <Return> to see next plot: 4



From the residual plot, it appears that the points are randomly scattered around the reference line, which suggests that the assumptions are likely to be met.

Conclusion:

State your conclusions and discuss any limitations.

In conclusion, our comprehensive analysis of electric vehicle range, model year, and Clean Alternative Fuel Vehicle (CAFV) eligibility reveals significant relationships between these factors and the electric range of electric vehicles. The boxplots and scatterplot demonstrate a weakly linear, positively associated relationship between model year and electric range, suggesting that newer models tend to have a higher electric range. The multiple regression analysis further emphasizes the impact of CAFV eligibility, showing that vehicles with this status generally have a higher electric range than those not eligible.

The residual plot indicates that our model assumptions are likely met, while the two-way ANOVA results confirm the statistical significance of the relationships between model year, CAFV eligibility, and electric range. The absence of a significant interaction between model year and CAFV eligibility implies that the relationship between model year and electric range is consistent across different levels of CAFV eligibility.

Overall, our findings provide strong evidence that both model year and CAFV eligibility play crucial roles in determining the electric range of electric vehicles. This information can be beneficial for stakeholders in the electric vehicle industry to develop strategies to improve electric range and meet consumer demands. Additionally, it can guide policymakers in refining incentive programs and regulations to encourage the production and adoption of electric vehicles with higher electric ranges, ultimately contributing to a more sustainable future.

In simple words we can say,
Newer electric vehicle models generally have a longer driving range on a single charge compared to older models.
Electric vehicles that meet certain "green" criteria, such as Clean Alternative Fuel Vehicle (CAFV) eligibility, tend to have a longer driving range.

The year a vehicle was made and whether it meets the CAFV criteria both play important roles in determining how far it can drive on a single charge.
These relationships are not due to random chance; there is evidence to support the idea that these factors impact the driving range of electric vehicles.

In summary, we found that newer and "greener" electric vehicles typically have a longer driving range. This
information can help both consumers and manufacturers make informed decisions about electric vehicle production and purchases, leading to an eco-friendlier future.

Limitations:

1. Limited scope of data: Our analysis is based on the data available to us and the sample data that we considered, which might not capture every electric vehicle model or all relevant factors that could impact electric range. As new models and technologies are introduced, our conclusions may need to be updated.

2. Correlation does not imply causation: While we found relationships between model year, CAFV eligibility, and electric range, we cannot conclusively determine causal relationships from this analysis alone. Further research, including experimental studies or longitudinal data, may be needed to establish causality.

3. Potential confounding factors: Other factors, such as battery technology, vehicle weight, or driving conditions, could also influence electric range. Our analysis did not account for all these potential confounding factors, which might affect the strength or direction of the relationships observed.

4. Generalizability: Our conclusions might not be applicable to all electric vehicles or different markets and regulatory environments. The relationships observed in our study may vary across different regions, driving conditions, or consumer preferences.

In conclusion, despite the limitations of our study, we have successfully provided valuable insights into the relationships between model year, CAFV eligibility, and electric range for electric vehicles. Our analysis demonstrates that newer and "greener" electric vehicles generally have a longer driving range, which can inform both manufacturers and consumers in making eco-friendly decisions related to electric vehicle production and purchases.

While it is essential to consider the limitations mentioned earlier, our findings still contribute to a better understanding of the factors influencing electric vehicle range. As the electric vehicle industry continues to evolve and data becomes more comprehensive, future research can build upon our results to provide an even clearer picture of the determinants of electric range.

Ultimately, our project offers a solid foundation for further exploration and serves as a steppingstone toward a more sustainable future for the electric vehicle industry and our environment.