Electric Vehicle Adoption

GROUP-A
RISHAV MONDAL
AASHIKASHRESTHA
SANDHYA POUDEL
JOSEPH DIRAVIAM LINDON RITHI VERONICA

Professor: Yue Gao
Date : 12/18/2023

Abstract:

This paper is motivated by the imperative to understand the intricate factors influencing electric vehicle (EV) range and eligibility, crucial for advancing sustainable transportation. Leveraging data from Washington State's EV registry, we employ χ^2-square, one-way ANOVA and linear regression models such as OLS to analyze and to scrutinize the relationships among **CAFV eligibility, model year, vehicle type, model, make,** and **electric range**. Results indicate significant associations between these variables, revealing that their collective impact enhances predictive power. The study underscores the importance of considering these factors collectively rather than in isolation. In conclusion, unraveling these complexities is essential for guiding policies, technologies, and consumer confidence in fostering widespread EV adoption and environmental stewardship.

Introduction

The adoption of electric vehicles (EVs) represents a critical and transformative shift in the automotive industry, driven by environmental concerns, technological advancements, and government initiatives. As our society continues to search for sustainable alternatives, understanding the factors influencing EV range and eligibility becomes paramount. This report delves into a comprehensive analysis of EV data from Washington State, focusing on key variables such as model year, vehicle type, model, make, and their impact on electric range and eligibility.

Motivation of the Paper

The motivation behind this paper lies in the growing significance of electric vehicles as a sustainable mode of transportation. With a global emphasis on reducing carbon footprints and mitigating climate change, EVs have emerged as a viable solution. However, challenges persist in understanding the intricate factors that influence the electric range of these vehicles and their eligibility for Clean Alternative Fuel Vehicle (CAFV) programs. By unraveling these complexities, we aim to contribute to the broader understanding of EV dynamics and support informed decision-making in the transition to sustainable transportation.

Specific Problem Under Study

The specific problem under study revolves around identifying and analyzing the multifaceted factors that shape the electric range and eligibility of EVs. These factors include the model year, vehicle type, model, and make, each playing a unique role in determining the performance and suitability of electric vehicles. By dissecting the relationships among these variables, we seek to uncover patterns, correlations, and insights that can guide stakeholders, policymakers, and researchers in enhancing the adoption and efficiency of electric vehicles.

Why Studying the Problem is Important

Studying the factors influencing EV range and eligibility holds paramount importance for several reasons:

- **Environmental Impact:** Understanding the determinants of EV range aids in assessing their environmental impact. This knowledge is crucial for gauging the effectiveness of EVs in reducing greenhouse gas emissions and achieving sustainability goals.

- **Policy Implications:** Policymakers can use insights from this study to tailor regulations and incentives that encourage the production and adoption of EVs. CAFV eligibility, in particular, is a key criterion for policy considerations.

- **Consumer Confidence:** For EVs to gain widespread acceptance, consumers must have confidence in the reliability and range of these vehicles. By deciphering the variables influencing electric range, manufacturers can design vehicles that meet consumer expectations.

- **Technological Advancements:** As technology evolves, understanding how improvements in model years and vehicle types impact electric range is crucial. This knowledge informs ongoing research and development efforts in the electric vehicle sector.

Research Questions

Although there can be many potential research question like:

- What is the impact of model year on the electric range of electric vehicles?
- How do different vehicle types (Battery Electric Vehicle vs. Plug-in Hybrid Electric Vehicle) influence electric range and eligibility?
- Is there a correlation between the model and make of electric vehicles and their electric range?
- What are the policy implications of Clean Alternative Fuel Vehicle (CAFV) eligibility on electric vehicle adoption?

This study aims to address the following research questions:

What are the intricate factors influencing the electric range and eligibility of electric vehicles?

Methods and Analysis

Data Description

The dataset has been obtained from the data catalog of the US government's open data source Data.gov whichis a collection of Evs in the state of Washington, registered under Washington State Department of Licensing . There are a total of 4 independent variables and 1 dependent variable. There are a totalof 150482 rows and therefore has a total of 752410 data points.All of the independent variables are mostly categorical, with the output data being continuous ratio-type.

Independent variables:

1. **Model**: This refers to the type of the model of the electric vehicle in our dataset. It is a nominal data type and includes different types of electric vehicles. For example 'Kona', 'Grand Cherokee', 'Model 3', etc. It has been encoded to be used for analysis to the object datatype for appropriate modeling.

2. **Clean Alternative Fuel Vehicle (CAFV) Eligibility**: This refers to the eligibility of the vehicle based on whether or not it satisfies the eligibility criteria. It is a nominal type of data that takes three types of data: 'Eligible', 'Not eligible', and eligibility unknown'.

3. **Electric Vehicle Type**: This is also a nominal type of data. There are two possible options: 'Battery Electric Vehicle(BEV)' or Plug-in Hybrid Electric Vehicle (PHEV). The BEV type exclusively as a battery for its power source whereas the PHEV includes both battery and Internal Combustion Engine for its power source.

4. **Make**: This refers to the company that makes the vehicle. It is also a nominal type of data and has variables like 'Hyundai', 'Jeep', 'Tesla' etc. Dependent variable:

5. **Electric range:** It represents the total range of the vehicle in miles. It is a continuous ratio-type of data that has a true and meaningful 'zero value'.

6. **Model Year:** It represents the year the model of the car was first created by the respective company.

Preparation of the Data

Dropping unnecessary columns

After analyzing all the different types of independent variables which amounted to more than 15, we chose to keep only those specific independent variables that can be quantifiably shown to have an effect on the dependent variables. In this process, the following columns were dropped from the data frame: 'Vehicle location', 'DOL Vehicle ID', 'Postal Code', '2020 Census Tract', 'VIN (1-10', 'Legislative District', 'Vehicle Location', 'State', and 'Base MSRP'.

Renaming of the columns

The initial columns contained names that were long and had spaces in them. Since Regression analysis functions only take titles without spaces, we renamed them with underscores between them. This made the variables fit for regression model analysis. For example, 'Electric Vehicle Type' transformed to 'Electric_Vehicle_Type' etc.

Categorical Data Encoding

Since all the independent variables are categorical, we encoded them into numerical values to make it possible to process them within the program. We used the LabelEncoder() utility from the sklearn machine learning library. This encoding process generated extra columns beside the main dataframe which is a encoded version of the string categorical variables:

1. 'CAFV_encoded' column was generated to hold the encoded information for Clean_Alternative_Fuel_Vehicle_(CAFV)_Eligibility column. It held values from 0 to 2, representing the three possible for the variable.
2. 'Type encoded' column was generated to hold encoded information for 'Electric_Vehicle_Type'. It took two values from 0 to 1 representing whether the vehicle was Battery Electric or Plug-in Hybrid Electric.
3. 'Model_encoded' column was generated to hold encoded information for the 'Model' column. It held values from 0 to the total number of unique values for the nominal variable.
4. 'Make_ncoded' column was generated to hold the encoded information for the 'Make' It held different integer values representing the different makes of the cars in the dataset.

Introduction of Dummy Variables

The Pandas library function 'get_dummies()' was used to generate the dummy variables for the Model year, CAFV and the vehicle type variables.

1. For 'Model Year', dummy variable columns were generated with the naming format: y_<year> representing each year for the variable.
2. For 'CAFV_encode', dummy variable columns were generated with the naming format: 'CAFV_x' where x takes the values from 0 to 2 representing either 'Eligible', 'Eligibility uknown', or 'Not eligible' respectively.
3. For **Type_encoded**, dummy variable columns were with the naming convention, Type_y. Here **y** held values either '0' or '1' representing either Battery Electric Vehicle or Plug-in Hybrid Electric Vehicle.

Removing of Vehicle data prior to 2010

Due to factors of relevance in predicting power of historical data while taking into account the shifting trends in the vehicle market, we chose to remove vehicles prior to 2010. This resulted in a dataset with a total of 150446 rows and 60 columns (a total of 9,026,760 data points).

Checking for missing values

The isnull() function was used to find out of there were any missing values in the related variables. No missing values were found in the dependent variable 'Electric_Range' and the independent variables 'Model Year', 'Make', 'Electric_Vehicle_Type', 'Clean_Alternative_Fuel_Vehicle_ (CAFV)_Eligibility. Therefore no further cleaning was required.

Result

Statistics

We provide a comprehensive analysis of the dataset, focusing on electric vehicles and associated factors. The dataset comprises information on various attributes, including model year, electric range, and encoded categorical variables. The analysis aims to provide insights into the characteristics and trends within the electric vehicle market.

Dataset Overview

Size of the Dataset: 150,446 records
Time Range: 2010 to 2024

Summary Statistics of some important variables

Model Year:

- Mean: 2020
- Standard Deviation: 3.00
- Minimum: 2010
- Maximum: 2024

Electric Range:

- Mean: 67.86
- Standard Deviation: 96.23
- Minimum: 0
- Maximum: 337

Yearly Distribution

- The dataset includes records for each year from 1997 to 2024.
- But for the analysis, we remove any data before the year 2010.
- Notable changes in variables occur from 2018, with a significant increase in CAFV-encoded and Type-encoded values.

	Model Year	Electric_Range	CAFV_encoded	Type_encoded	Model_encoded	Make_encoded	y_2010	y_2011	y_2012	y_2
count	150446.000000	150446.000000	150446.000000	150446.000000	150446.000000	150446.000000	150446.000000	150446.000000	150446.000000	150446.000
mean	2020.009226	67.858707	0.700344	0.223834	0.700344	0.223834	0.000160	0.005291	0.010854	0.030
std	3.004906	96.226194	0.668531	0.416814	0.668531	0.416814	0.012629	0.072546	0.103618	0.171
min	2010.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
25%	2018.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
50%	2021.000000	18.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000
75%	2023.000000	97.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000
max	2024.000000	337.000000	2.000000	1.000000	2.000000	1.000000	1.000000	1.000000	1.000000	1.000

8 rows × 26 columns

Fig.1: Statistics related to the dataset

We will not be able to include the statistics of some of the other variables, since most of them are categorical variables which have been encoded for analysis.

Charts

We plotted multiple charts to represent the distribution of the data across the dataset. As mentioned previously there are a few important variables which will be considered as the independent variables, wit **Electric Range** being the dependent variable. The charts represent the relationship of the variables.

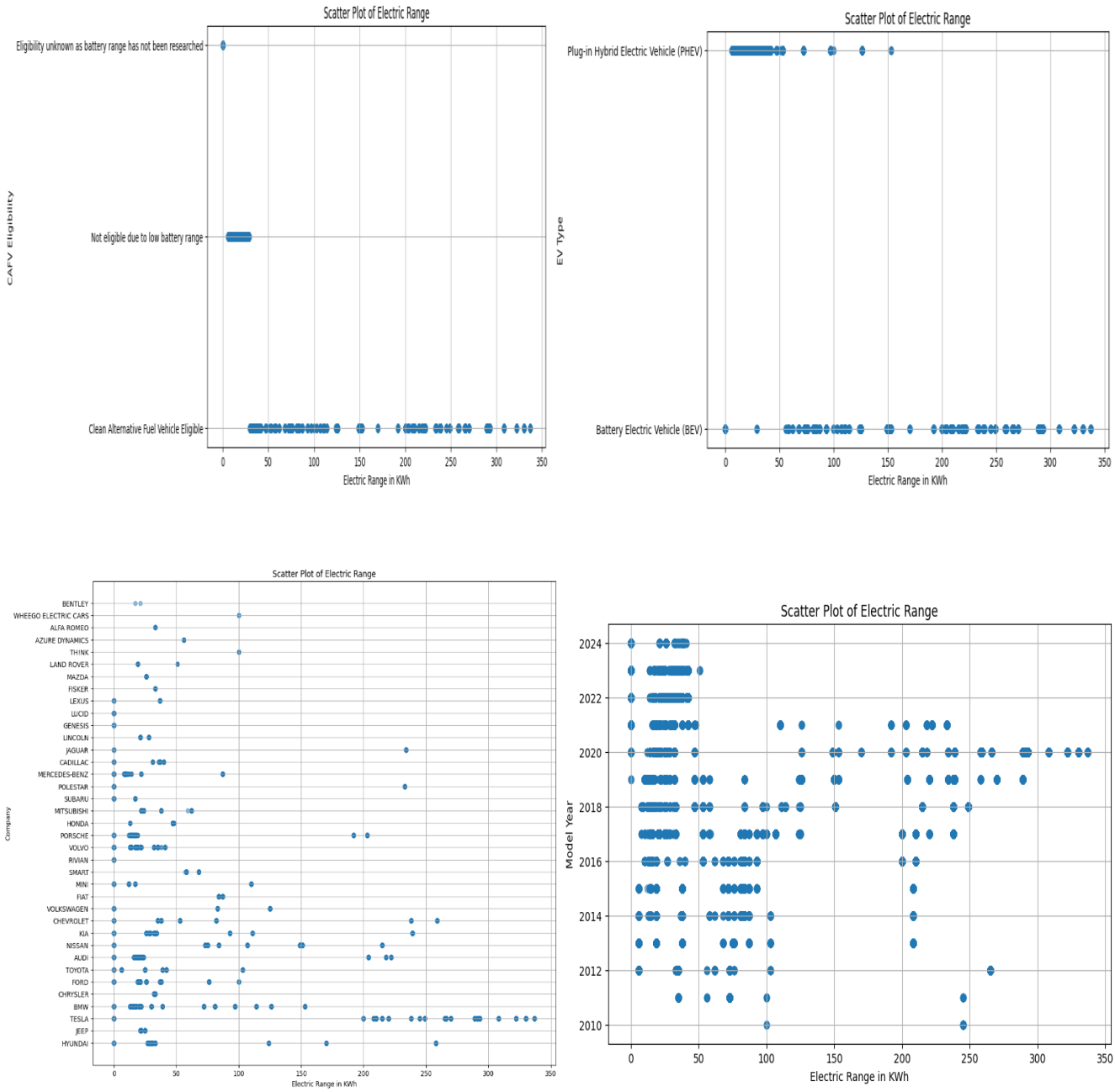
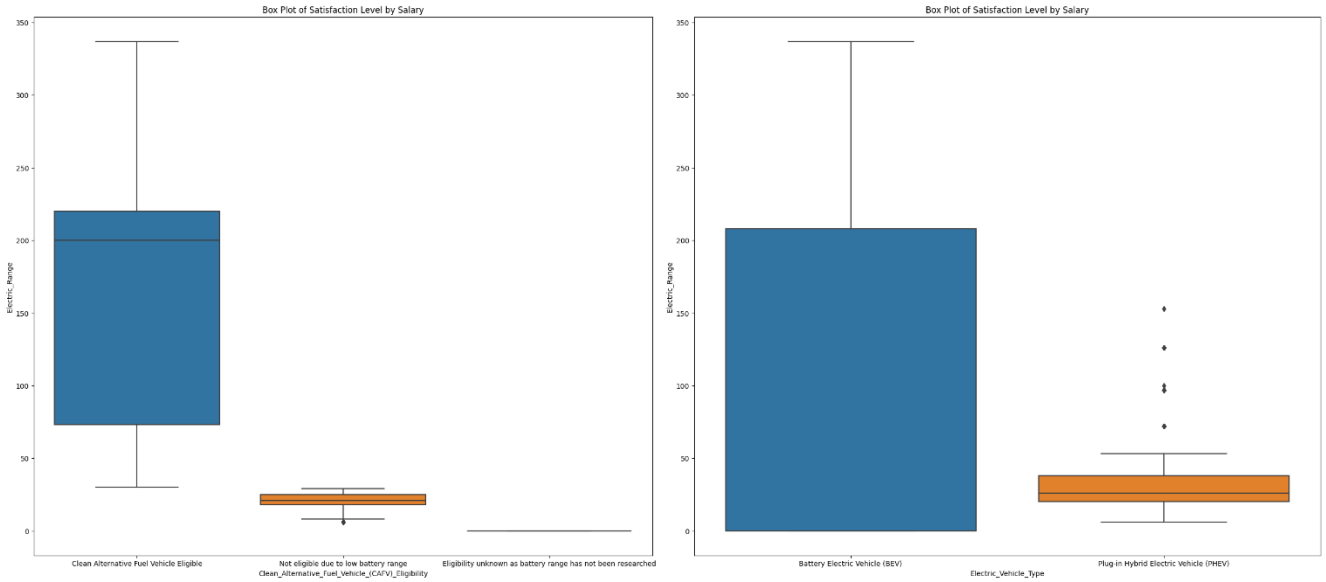


Fig.2: The Scatter plots explain the distribution of the data according to the vriables.

We also plotted a boxplot to see in which **Electric Range** did most of the the vehicles fell in according to the categories of the two most important independent variables **CAFV Eligibility** and **Vehicle Type**.



(a) (b)

Fig.3: **a** and **b** show the box plot of data distribution.

Here, **(a)** shows that the CAFV eligible cars mostly has the electric range between 75-225 KWh, while the unknown one fell only within 25-35 KWh and the ineligible ones are all in the 0 range.(Note. *This plot does not show the which category has the most vehicles, but shows in which electric range did most vehicles lie in their respective category*). Similarly **(b)** shows that most of the cars in the category of battery electric vehicles fall within the electric range of 0-225 Kwh, while for the plug-in hybrid vehicles mostly lie within the electric range of 25-45 Kwh.

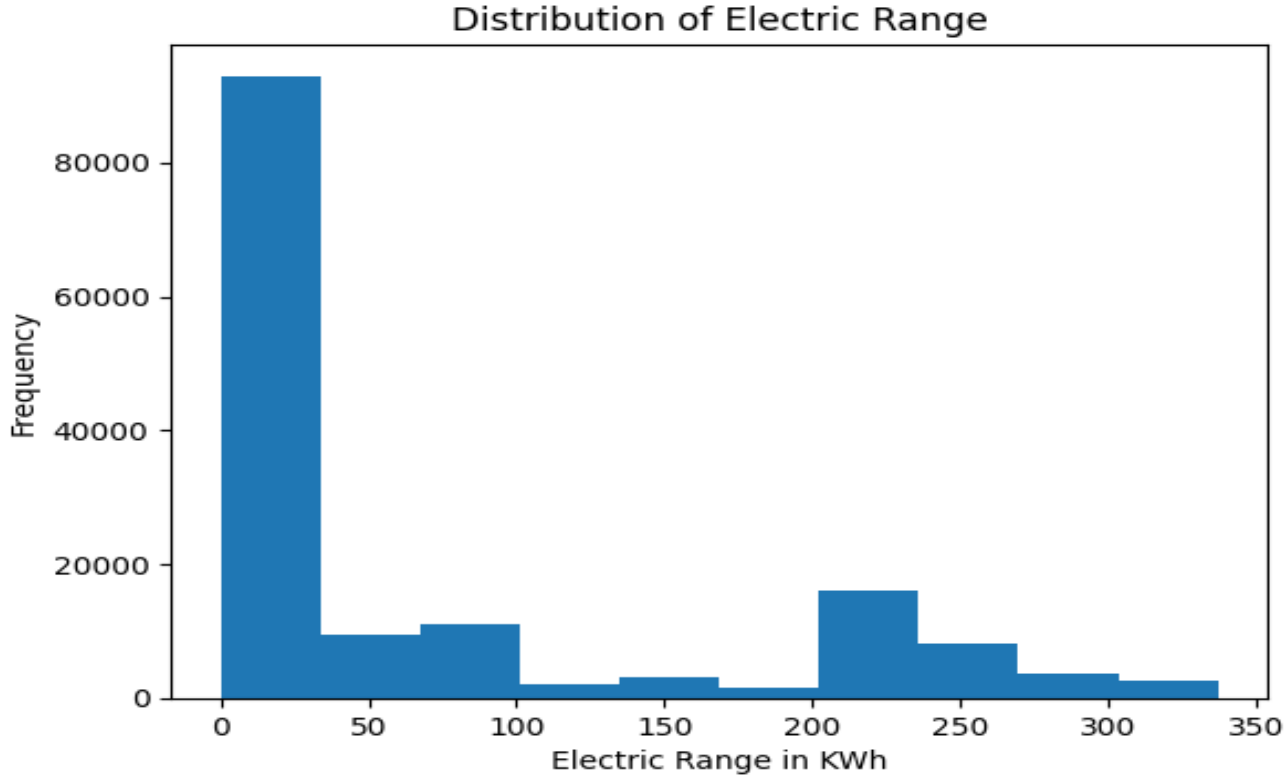


Fig.4: Histogram to show the population distribution of the cars according to electric range.

The above histogram shows the real distribution of the cars according to their electric range, and we can clearly see that most of the cars recorded in the dataset has an electric range of 0 KWh, which mean most of the cars are ineligible to be a CAFV.

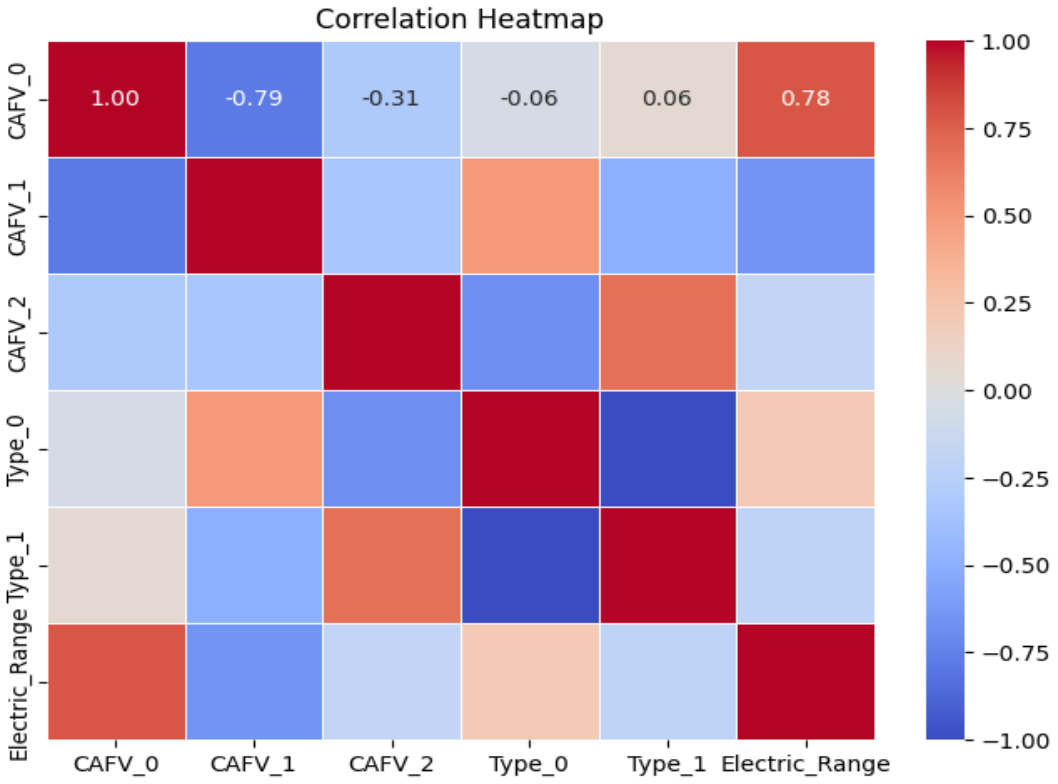


Fig. 5: Heat map showing the relationship between the various variables.

I the above heat map we see the relationship between two of the most important independent variables for the analysis with the dependent variable electric range. Since most of the variables are categorical, we could only plot their dummy variables, while for the other variables, plotting the dummy variable became impractical.

Hypothesis

We performed a few hypotheses like the **χ^2-test** and **One-Way ANOVA** test. We got some very interesting results from the test analysis.

χ^2-test

The chi-square test was conducted to investigate the association between categorical variables with the dependent variable, specifically focusing on the relationship between 'Electric Vehicle Type' and 'CAFV Eligibility' with the dependent variable ‘Electric Range’. The results indicate a significant association ($p < 0.05$), suggesting that the type of electric vehicle is related to its eligibility for Clean Alternative Fuel Vehicle programs. Further analysis of the chi-square results is necessary to interpret the strength and nature of this association.

```
Chi-squared statistic: 150358.15822091844
p-value: 0.0
Since p-value<alpha,we reject the null hypothesis: There is a significant association between CAFV and Electric Range.
```

Fig.6: χ^2-test for CAFV and Electric Range

```
Chi-squared statistic: 150358.15822091844
p-value: 0.0
Since p-value<alpha,we reject the null hypothesis: There is a significant association between Vehicle Type and Electric Range.
```

Fig.7: χ^2-test for Vehicle Type and Electric Range

We, also performed a **χ^2-test** for all the independent variables combined against the dependent variable 'Electric Range'.

```
Chi-squared statistic: 2378772.5363605344
p-value: 0.0
Since p-value<alpha,we reject the null hypothesis: There is a significant association between CAFV, Model, Type, Make and Year
of vehicle and its Electric Range.
```

Fig.8: χ^2-test for all the independent variables and Electric Range

One-Way ANOVA

We applied one-way analysis of variance (ANOVA) to examine the impact on the 'Electric Range' variable by the 'Electric Vehicle Type' and the 'CAFV Eligibility.' The analysis revealed a statistically significant difference in electric range among different types of electric vehicles ($p < 0.05$). Post-hoc tests, such as Tukey's HSD, may be employed to identify specific group differences and understand which electric vehicle types significantly differ in terms of electric range.

```
F-statistic: 120675.095183961
p-value: 0.0
Reject the null hypothesis: The Electric Range varies signifiacntly according to the CAFV eligibility of a vehicle.
```

Fig.9: One-Way ANOVA test between CAFV and Electric Range

```
F-statistic: 6777.380930231579
p-value: 0.0
Reject the null hypothesis: The Electric Range varies signifiacntly according to the Type of a vehicle.
```

Fig.10: One-Way ANOVA test between Vehicle Type and Electric Range

Models

We present the results of a few linear regression analysis(OLS-Ordinary Least Squares) aimed at exploring the factors influencing the 'Electric_Range' of electric vehicles in Washington State. The dependent variable is 'Electric_Range,' and various independent variables, including categorical variables such as 'CAFV' and 'Type,' as well as yearly indicators from 2010 to 2024, have been included in the model.

The first test was performed linear regression analysis using OLS on the variable 'CAFV Eligibility' and 'Electric Range', while the second was performed on the variables 'Vehicle Type' and 'Electric Range'. Since the variables are categorical we used the dummy variables hence produced to perform the analysis. The results thus obtained is shown in the images below.

OLS Regression Results

Dep. Variable:	Electric_Range	R-squared:	0.616
Model:	OLS	Adj. R-squared:	0.616
Method:	Least Squares	F-statistic:	8.045e+04
Date:	Fri, 15 Dec 2023	Prob (F-statistic):	0.00
Time:	19:54:56	Log-Likelihood:	-8.2852e+05
No. Observations:	150446	AIC:	1.657e+06
Df Residuals:	150442	BIC:	1.657e+06
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.918e+11	7.28e+12	-0.054	0.957	-1.47e+13	1.39e+13
CAFV_0	3.918e+11	7.28e+12	0.054	0.957	-1.39e+13	1.47e+13
CAFV_1	3.918e+11	7.28e+12	0.054	0.957	-1.39e+13	1.47e+13
CAFV_2	3.918e+11	7.28e+12	0.054	0.957	-1.39e+13	1.47e+13

Omnibus:	2055.934	Durbin-Watson:	2.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3831.606
Skew:	0.032	Prob(JB):	0.00
Kurtosis:	3.779	Cond. No.	1.13e+14

(1)

OLS Regression Results

Dep. Variable:	Electric_Range	R-squared:	0.043
Model:	OLS	Adj. R-squared:	0.043
Method:	Least Squares	F-statistic:	6777.
Date:	Fri, 15 Dec 2023	Prob (F-statistic):	0.00
Time:	19:55:17	Log-Likelihood:	-8.9720e+05
No. Observations:	150446	AIC:	1.794e+06
Df Residuals:	150444	BIC:	1.794e+06
Df Model:	1		
Covariance Type:	nonrobust		

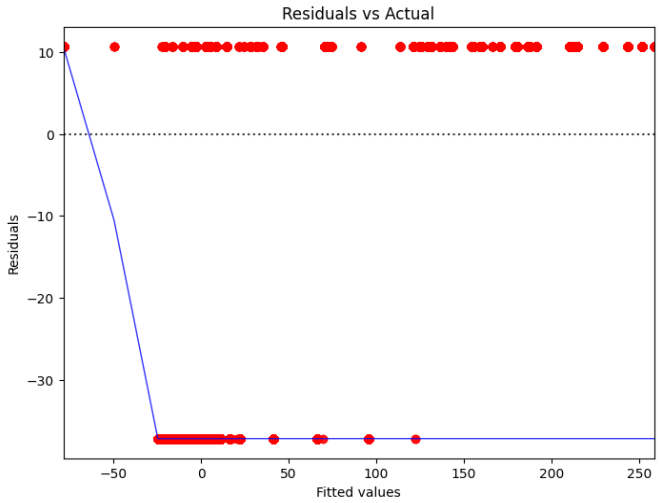
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	36.4143	0.194	187.628	0.000	36.034	36.795
Type_0	42.1732	0.251	168.077	0.000	41.681	42.665
Type_1	-5.7589	0.354	-16.264	0.000	-6.453	-5.065

Omnibus:	18112.374	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	24833.617
Skew:	0.985	Prob(JB):	0.00
Kurtosis:	2.710	Cond. No.	1.28e+17

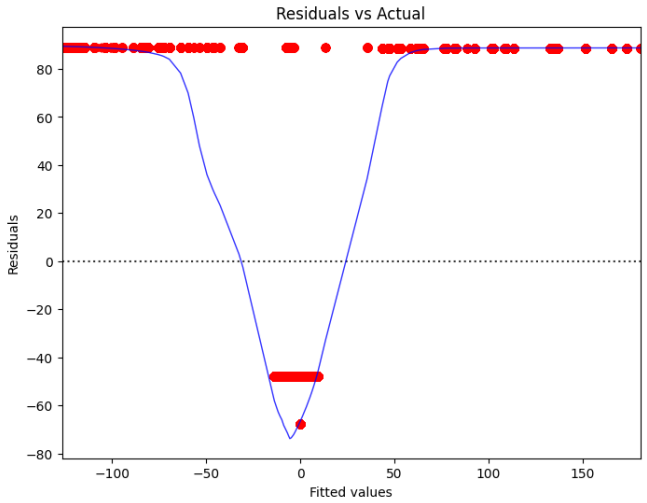
(2)

Fig.11: **1** shows result between CAFV Eligibility and Electric Range, while **2** the same for Vehicle Type and Electric Range

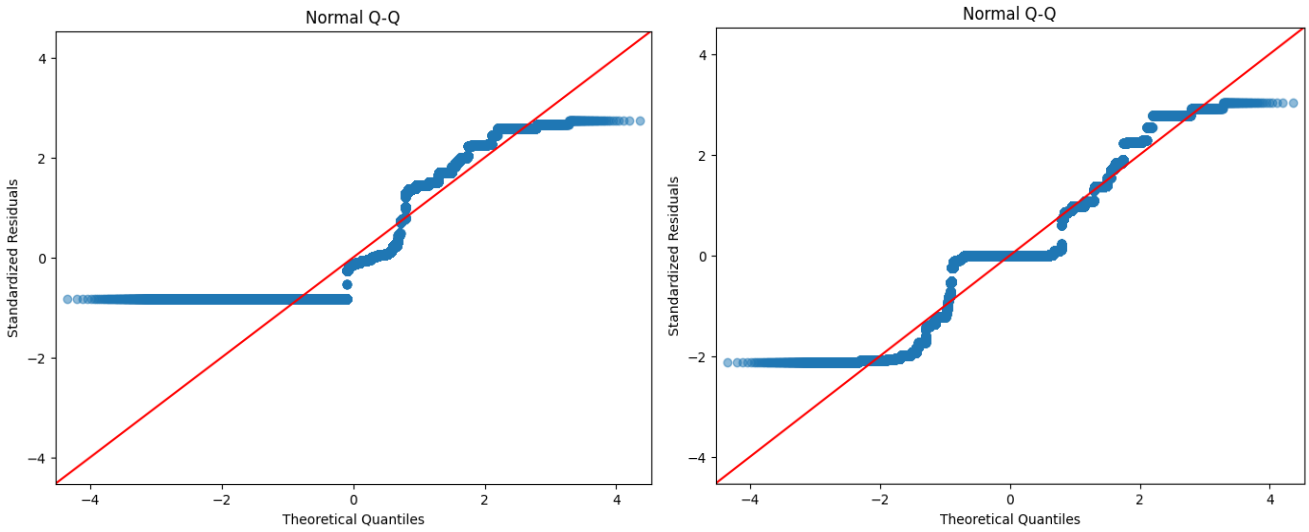
We have also plotted the residual-fitted and the qq-plots for the results obtained from the OLS regression that we performed to check the performance of the model.



(a)



(b)



(c) (d)

Fig.12: residual-fitted and q-q plot for the OLS regression

(a) and (c) are the plots related to OLS regression from Fig.11 image (2) demonstrating the details of the regression performed using 'Vehicle Type' and 'Electric Range', and (b) and (d) demonstrates the results for Fig.11 image (1) of 'CAFV Eligibility' and 'Electric range'.

We then performed multiple linear regression analysis on the two independent variables against the dependent variable 'Electric Range'. The performance of the model increased by a lot. We ca see the result and performance of the regression model below.

OLS Regression Results

Dep. Variable:	Electric_Range	R-squared:	0.813
Model:	OLS	Adj. R-squared:	0.813
Method:	Least Squares	F-statistic:	1.634e+05
Date:	Fri, 15 Dec 2023	Prob (F-statistic):	0.00
Time:	20:01:38	Log-Likelihood:	-7.7444e+05
No. Observations:	150446	AIC:	1.549e+06
Df Residuals:	150441	BIC:	1.549e+06
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.567e+11	3.42e+12	-0.075	0.940	-6.96e+12	6.45e+12
CAFV_0	3.819e+11	5.09e+12	0.075	0.940	-9.6e+12	1.04e+13
CAFV_1	3.819e+11	5.09e+12	0.075	0.940	-9.6e+12	1.04e+13
CAFV_2	3.819e+11	5.09e+12	0.075	0.940	-9.6e+12	1.04e+13
Type_0	-1.252e+11	1.67e+12	-0.075	0.940	-3.4e+12	3.15e+12
Type_1	-1.252e+11	1.67e+12	-0.075	0.940	-3.4e+12	3.15e+12

Omnibus:	19590.924	Durbin-Watson:	2.003
Prob(Omnibus):	0.000	Jarque-Bera (JB):	84606.344
Skew:	-0.585	Prob(JB):	0.00
Kurtosis:	6.482	Cond. No.	1.51e+16

Fig.13: Multiple Linear Regression with two independent variables

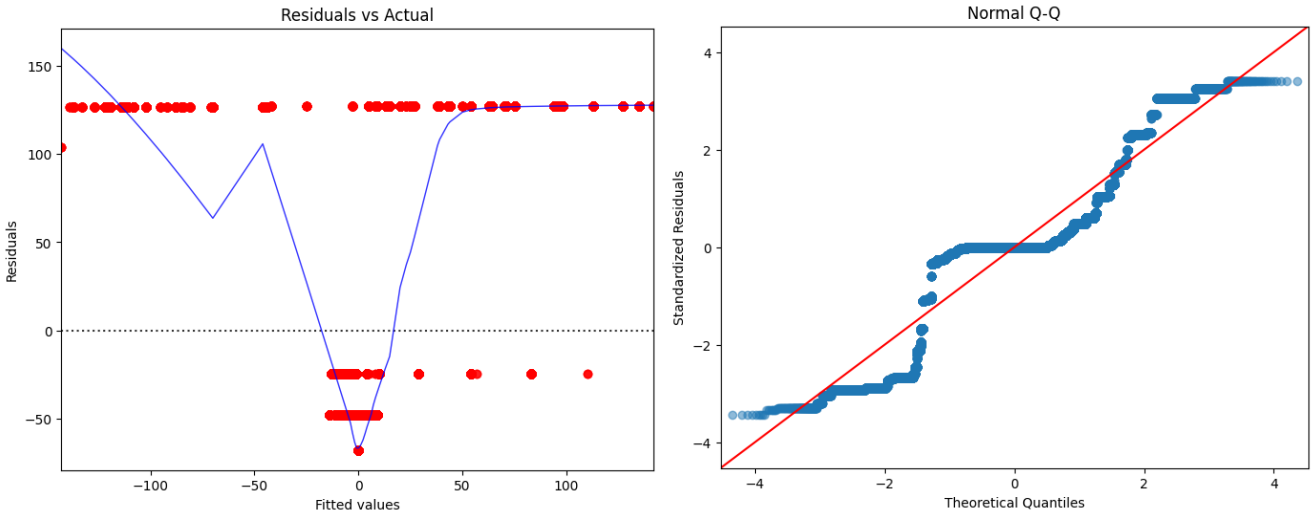


Fig.14: residual-fitted and qq plots for multiple linear regression of 2 variables

Last but not the least , we performed the same multiple linear regression but this time we added a new independent variable 'Model Year' to the formula. The result was even better than the previous multiple linear regression analysis.

OLS Regression Results

Dep. Variable:	Electric_Range	R-squared:	0.899
Model:	OLS	Adj. R-squared:	0.899
Method:	Least Squares	F-statistic:	7.415e+04
Date:	Fri, 15 Dec 2023	Prob (F-statistic):	0.00
Time:	20:01:43	Log-Likelihood:	-7.2827e+05
No. Observations:	150446	AIC:	1.457e+06
Df Residuals:	150427	BIC:	1.457e+06
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	8.251e+11	3.06e+12	0.270	0.787	-5.17e+12	6.82e+12
CAFV_0	1.094e+12	3.75e+12	0.291	0.771	-6.26e+12	8.45e+12
CAFV_1	1.094e+12	3.75e+12	0.291	0.771	-6.26e+12	8.45e+12
CAFV_2	1.094e+12	3.75e+12	0.291	0.771	-6.26e+12	8.45e+12
Type_0	4.137e+11	1.5e+12	0.275	0.783	-2.53e+12	3.36e+12
Type_1	4.137e+11	1.5e+12	0.275	0.783	-2.53e+12	3.36e+12
y_2010	-2.333e+12	2.93e+12	-0.795	0.427	-8.08e+12	3.42e+12
y_2011	-2.333e+12	2.93e+12	-0.795	0.427	-8.08e+12	3.42e+12
y_2012	-2.333e+12	2.93e+12	-0.795	0.427	-8.08e+12	3.42e+12
y_2013	-2.333e+12	2.93e+12	-0.795	0.427	-8.08e+12	3.42e+12
y_2014	-2.333e+12	2.93e+12	-0.795	0.427	-8.08e+12	3.42e+12
y_2015	-2.333e+12	2.93e+12	-0.795	0.427	-8.08e+12	3.42e+12
y_2016	-2.333e+12	2.93e+12	-0.795	0.427	-8.08e+12	3.42e+12
y_2017	-2.333e+12	2.93e+12	-0.795	0.427	-8.08e+12	3.42e+12
y_2018	-2.333e+12	2.93e+12	-0.795	0.427	-8.08e+12	3.42e+12
y_2019	-2.333e+12	2.93e+12	-0.795	0.427	-8.08e+12	3.42e+12
y_2020	-2.333e+12	2.93e+12	-0.795	0.427	-8.08e+12	3.42e+12
y_2021	-2.333e+12	2.93e+12	-0.795	0.427	-8.08e+12	3.42e+12
y_2022	-2.333e+12	2.93e+12	-0.795	0.427	-8.08e+12	3.42e+12
y_2023	-2.333e+12	2.93e+12	-0.795	0.427	-8.08e+12	3.42e+12
y_2024	-2.333e+12	2.93e+12	-0.795	0.427	-8.08e+12	3.42e+12

Omnibus:	16283.752	Durbin-Watson:	1.989
Prob(Omnibus):	0.000	Jarque-Bera (JB):	42737.830
Skew:	-0.619	Prob(JB):	0.00
Kurtosis:	5.299	Cond. No.	1.84e+16

Fig.15: Multiple Linear Regression with three independent variables

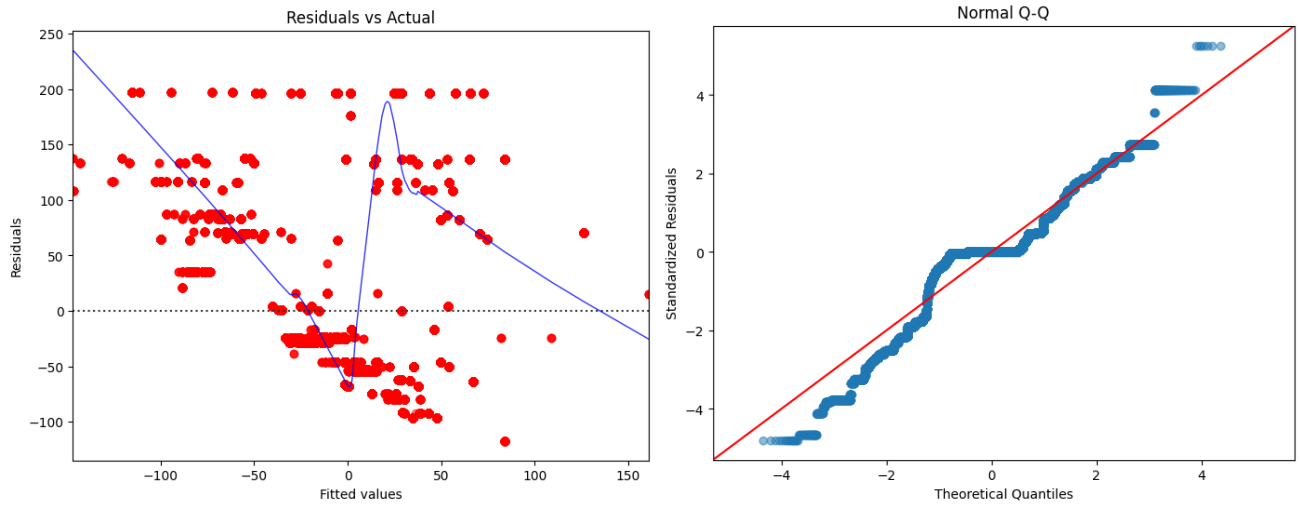


Fig.16: residual-fitted and qq plots for multiple linear regression of 3 variables

Discussion and Conclusion

Interpretation

Correlation

From the correlation heatmap, we see there is a moderate correlation between the 'CAFV_encoded' and the 'Type_encoded' .This means that the eligibility of a vehicle as an alternative fuel vehicle goes hand in hand with the type of vehicle. Specific type of vehicle is more likely to also be Clean alternative fuel eligible.

From the heat map we can also say that the electric range also has a direct relationship with the eligibility of the vehicle as a CAFV, since the heat map shows a significant correlation between the 'Electric_range' and the 'CAFV_0'.

Similarly , the heat map also shows that all the vehicles that belong to the battery electric vehicle(BEV) usually has a significant relationship with electric range of the vehicle.

Hypothesis

We performed multiple χ^2 analysis between Electric, and all the independent variables 'CAFV eligibility', 'Vehicle Type', 'Model', 'Make' , and 'Model Year' and found a p-value close to zero in all of them. This means there is a statistically significant relationship between all the independent variables individually and the Electric range.

We also performed two one-way ANOVA to see if the mean electric range varies across vehicles of different eligibility and make. The p-value obtained for both was very low and close to zero. This means the there is a statistically substantial variation in the range depending on whether or not the vehicle is eligible.

4 types of linear regression analysis were performed

Firstly, between Electric range and CAFV values. It had high $\mathbf{R^2}$ value of 0.616 , which means 61.6% o f the variability in the dependent variable has been successfully defined using the independent variable, and low p-value of 0.0 and high F-stat of 8.045e+04 means that the model is statistically significant. But the individual CAFV were not statistically significant at all. This means the independent variables w ere significant as a whole but not individually.

Secondly, between Electric range and Vehicle type. This had a 0.043 $\mathbf{R^2}$ value, which means that the mo del does not have predictive power with the independent variables as a whole, but the F-stat of 6777 sug gest that the model is statistically significant.

Thirdly, between Electric range with CAFV and Type together. It had a high $\mathbf{R^2}$ valueof 0.813 which m eans 81.3% of the variability in the dependent variable has been successfully defined using the indepen dent variables, and low p-value of 0.0 and high F-stat of 1.634e+05 means that the model is statistically significant. There it had the best predictive value with 80% of the variance in the range explained by th e independent variables of CAFV and Type together. However, they had individual p-values 0.940. Thi s indicates the independent variables were not individually statistically significant.

Fourthly, between Electric range with CAFV, Type and model year. This also had a R-squared value of 0.899 which means 89.9% of the variability in the dependent variable has been successfully defined us ing the independent variables, and low p-value of 0.0 and high F-stat of 7.415e+04 means that the mod el is statistically significant. This means the model had high predictive power taking the independent va riables as a group. However, the independent variables individually had high p-value, so were not statis tically significant individually.

Plots

The Residuals vs Fitted plot for the for the different results shows that the more independent variables we added, the more non-linear the data became. Good Residual vs Fitted plot would involve random scattering of the point, i.e. the plot will not follow any set pattern. From our plots we found that the Residual vs Fitted plot showed that the regression analysis with the worst result had the a plot following a particular pattern and vice versa.

The QQ- probability plot we created showed that for the regression analysis with the worst result the model is underestimating the variance of the errors in the tails of the distribution. In simpler terms, there are more extreme values (outliers) in your data than the normal distribution would predict. This problem continuously rectifies itself as we increase the number of independent variables used for the analysis. With the analysis with the best result produced a QQ-plot that was nearly straight with proper estimation of variables.

Limitations:

We had to churn the data down to a few independent variables of only 5 independent variables from more than 15. This is obviously a huge limitation because there could be more variables that might affect the electric vehicle range. Plus, the residual plots vs actual values for more independent variables showed that that the data is increasing non-linear. This means a non-linear modeling could have been a better fit to explained the observed data.

Future work and conclusions

This analysis can be extended to include data that are more numerical in nature. Currently it holds only categorical data and is there therefore unable to show the true linear relationship that could have been shown with modeling with scalar independent values. The work can further be extended to include to more types of electric vehicles such as hydrogen based, water-based etc. Exploration can be conducted to find the impact of different geographic regions, climate conditions and driving patterns on EV range and eligibility. The relationship between advancements in battery technology and EV range can also be similarly explored in terms of the influence of the battery degradation over time on EV range and eligibility.

We started our report with the question: "What are the intricate factors influencing electric vehicle range and eligibility?". We found that the model year, vehicle type, model and make significantly affects range, however they affect more together than individually.

References

1. https://www.sciencedirect.com/science/article/abs/pii/S0921344918300363
2. https://www.sciencedirect.com/science/article/abs/pii/096585649390062P
3. https://escholarship.org/content/qt2k09h787/qt2k09h787.pdf
4. https://www.sciencedirect.com/science/article/abs/pii/S1361920920306234
5. https://www.sciencedirect.com/science/article/abs/pii/S095965261934781X
6. https://www.tandfonline.com/doi/abs/10.1080/15568318.2020.1818330
7. https://github.com/suneethapatchala/-Exploratory-Data-Analysis-On-Electric-Vehicle-Population/tree/main
8. https://catalog.data.gov/dataset/electric-vehicle-population-data
9. https://ww2.arb.ca.gov/news/california-moves-accelerate-100-new-zero-emission-vehicle-sales-2035
10. https://books.google.com/books?hl=en&lr=&id=0ecUzLRyN5cC&oi=fnd&pg=PP6&dq=how+is+CAFV+eligibility+related+to+electric+range+of+an+EV&ots=hHpazrLucs&sig=fl8PcD6g51mxgofETiSLcEpQbFY#v=onepage&q&f=false