data_loading

April 26, 2023

```
<IPython.core.display.HTML object>
```

Electrical Vehicles Regression Analysis

Group 2, C1

1 Introduction

In recent years, more and more people are buying Electrical Vehicles (EVs) for environmental, aesthetic, and financial reasons. The number of car companies inventing EVs for their brand are increasing. Companies such as Tesla, Ford, and Rivian are taking advantage of this move toward EVs. The goal of our project is to examine just how much the range of EVs changes due to other factors such as battery pack. In this paper we analyze the relation between the range and other variables such as acceleration, top speed, battery pack, efficiency, fast charge and price in order to give us a deeper understanding of what element influences the range of EVs the most. Additionally, we can also see what car manufacturers should do to improve their EVs, and provide useful data for car companies.

2 Background

2.1 Dataset

Divyanshu Gupta, Kaggle (2021), Cars Dataset with Battery Capacity [Data File]. Retrieved from https://www.kaggle.com/datasets/divyanshugupta95/cars-dataset-with-battery-pack-capacity

2.2 Data Collection Method

The data was collected from different companies such as Tesla, Porsche, BMW. The data set also gives us the specific make and model of the cars. The dataset contains 14 explanatory variables with 1 response variable and a total of 102 data points.

3 Preliminary Analysis

3.1 Hypothesized Variables That Impact Electric Vehicle Range

The main covariate that we believe will have the largest impact on the range of an electric vehicle is battery packs in kilowatts per hour. The other variables that we are looking at would be acceleration, top speed, efficiency, how fast the car charges, and price. Qualitative variables would

be the plug style, number of seats, power train (all wheel drive vs. four wheel drive), and type of car. This mix of qualitative and quantitative variables will allow us to give a clearer understanding of how different factors affect the range of EVs.

```
          Brand                              Model  AccelSec  TopSpeed_KmH  \
0         Tesla    Model 3 Long Range Dual Motor       4.6           233
1    Volkswagen                         ID.3 Pure      10.0           160
2      Polestar                                 2       4.7           210
3           BMW                               iX3       6.8           180
4         Honda                                 e       9.5           145
..          …                                 …       …           …
97       Nissan                      Ariya 63kWh       7.5           160
98         Audi    e-tron S Sportback 55 quattro       4.5           210
99       Nissan              Ariya e-4ORCE 63kWh       5.9           200
100      Nissan  Ariya e-4ORCE 87kWh Performance       5.1           200
101       Byton                M-Byte 95 kWh 2WD       7.5           190

     Range_Km  Battery_Pack Kwh  Efficiency_WhKm  FastCharge_KmH RapidCharge  \
0         460              70.0              161             940         Yes
1         270              45.0              167             250         Yes
2         400              75.0              181             620         Yes
3         360              74.0              206             560         Yes
4         170              28.5              168             190         Yes
..        …                …                 …               …         …
97        330              63.0              191             440         Yes
98        335              86.5              258             540         Yes
99        325              63.0              194             440         Yes
100       375              87.0              232             450         Yes
101       400              95.0              238             480         Yes

    PowerTrain    PlugType  BodyStyle Segment  Seats  PriceEuro
0          AWD  Type 2 CCS      Sedan       D      5      55480
1          RWD  Type 2 CCS  Hatchback       C      5      30000
2          AWD  Type 2 CCS   Liftback       D      5      56440
3          RWD  Type 2 CCS        SUV       D      5      68040
4          RWD  Type 2 CCS  Hatchback       B      4      32997
..         …          …        …     …    …        …
97         FWD  Type 2 CCS  Hatchback       C      5      45000
98         AWD  Type 2 CCS        SUV       E      5      96050
99         AWD  Type 2 CCS  Hatchback       C      5      50000
100        AWD  Type 2 CCS  Hatchback       C      5      65000
101        AWD  Type 2 CCS        SUV       E      5      62000

[102 rows x 15 columns]
```

3.2 Exploring Influence of Outliers

We tested different outlier removal methods such as taking away data that was two and three standard deviations away from mean. Figure 1 compares the linear relationship of covariates against Range_Km with and without outliers respectively:

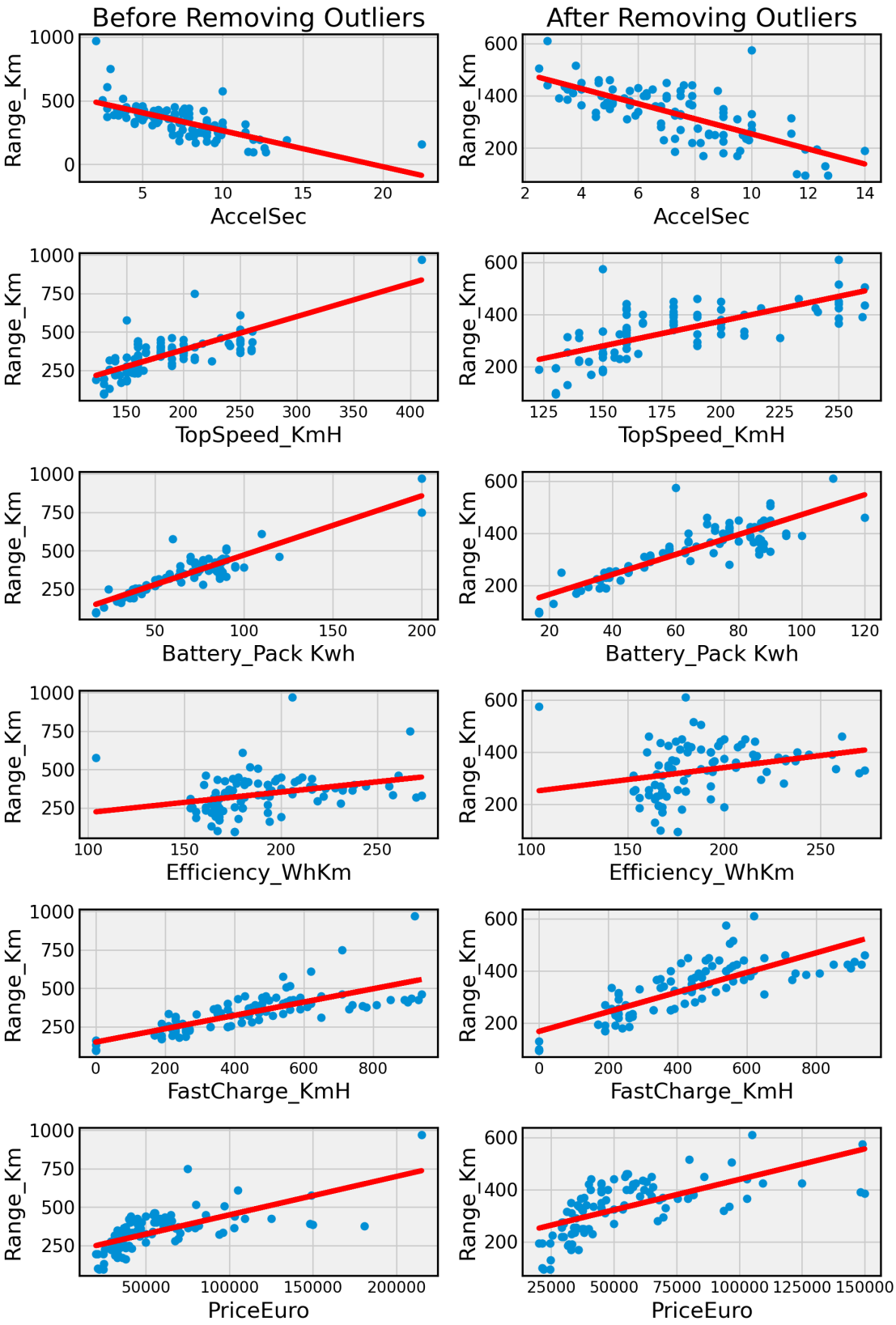
Before Removing Outliers
After Removing Outliers
Range_Km
AccelSec
TopSpeed_KmH
Battery_Pack Kwh
Efficiency_WhKm
FastCharge_KmH
PriceEuro

Figure 1

After analyzing the best-fit lines with and without outliers we went on to analyze the corresponding residual plots. Figure 2 compares the residual plots of covariates against Range_Km with and without outliers respectively:

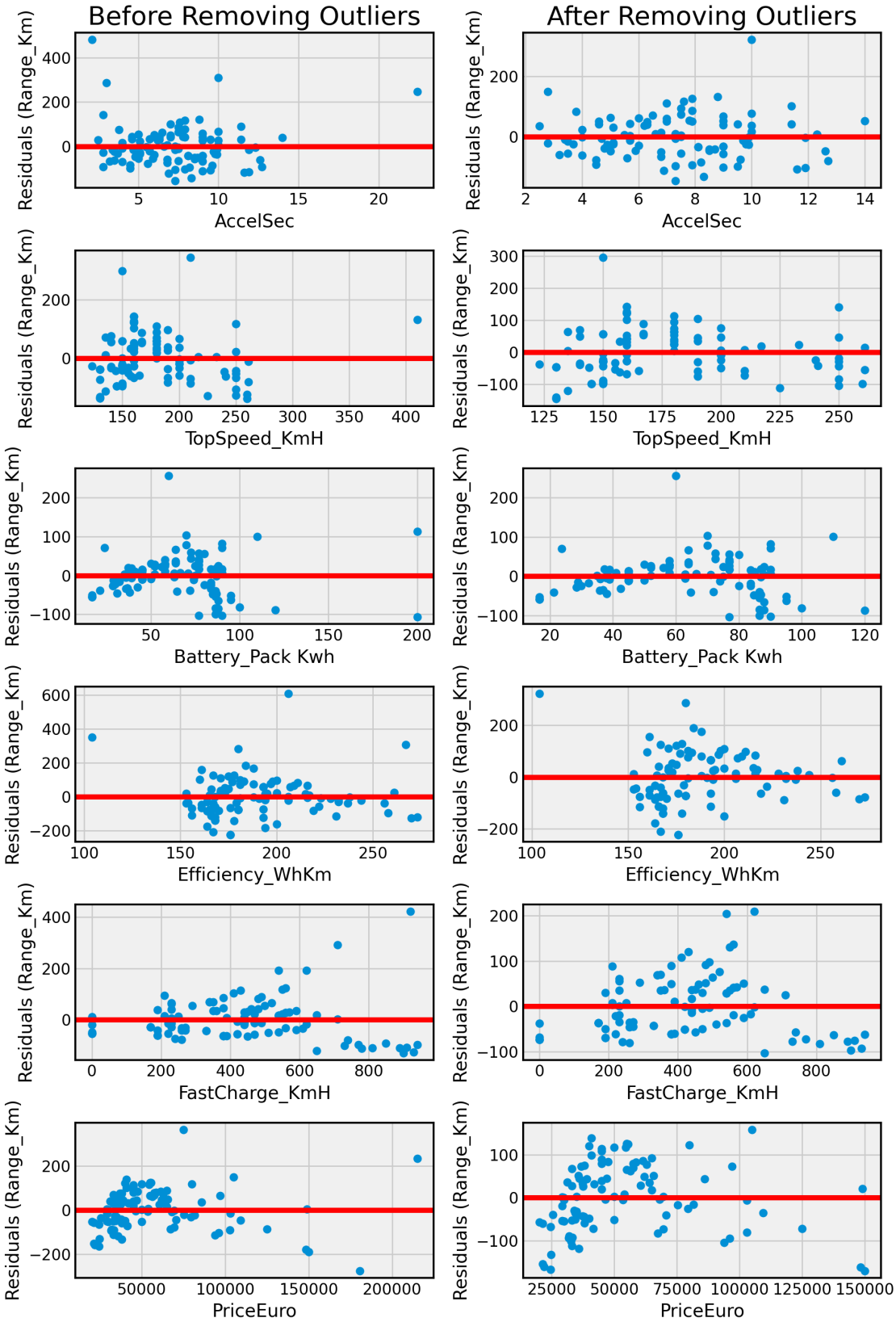
Before Removing Outliers
After Removing Outliers
Residuals (Range_Km)
AccelSec
TopSpeed_KmH
Battery_Pack Kwh
Efficiency_WhKm
FastCharge_KmH
PriceEuro

Figure 2

We found a minimal difference in the strength of the linear relationship (measured with R^2) between the covariates and response variable when filtering outliers. Additionally, our dataset is relatively small with 102 samples. For these reasons, we decided not to remove outliers for further stages of our analysis.

Covariate	R^2 Before	R^2 After
`AccelSec`	0.46	0.522
`TopSpeed_KmH`	0.56	0.463
`Battery_Pack Kwh`	0.829	0.753
`Efficiency_WhKm`	0.098	0.068
`FastCharge_KmH`	0.569	0.6
`PriceEuro`	0.458	0.411

Table 1

3.3 Normal Quantile Plot

In order to verify one of the central assumptions of linear regression, normally distributed residuals, we created a Q-Q plot in Figure 3 for all of our covariates to observe that this assumption holds true for most of our covariates.

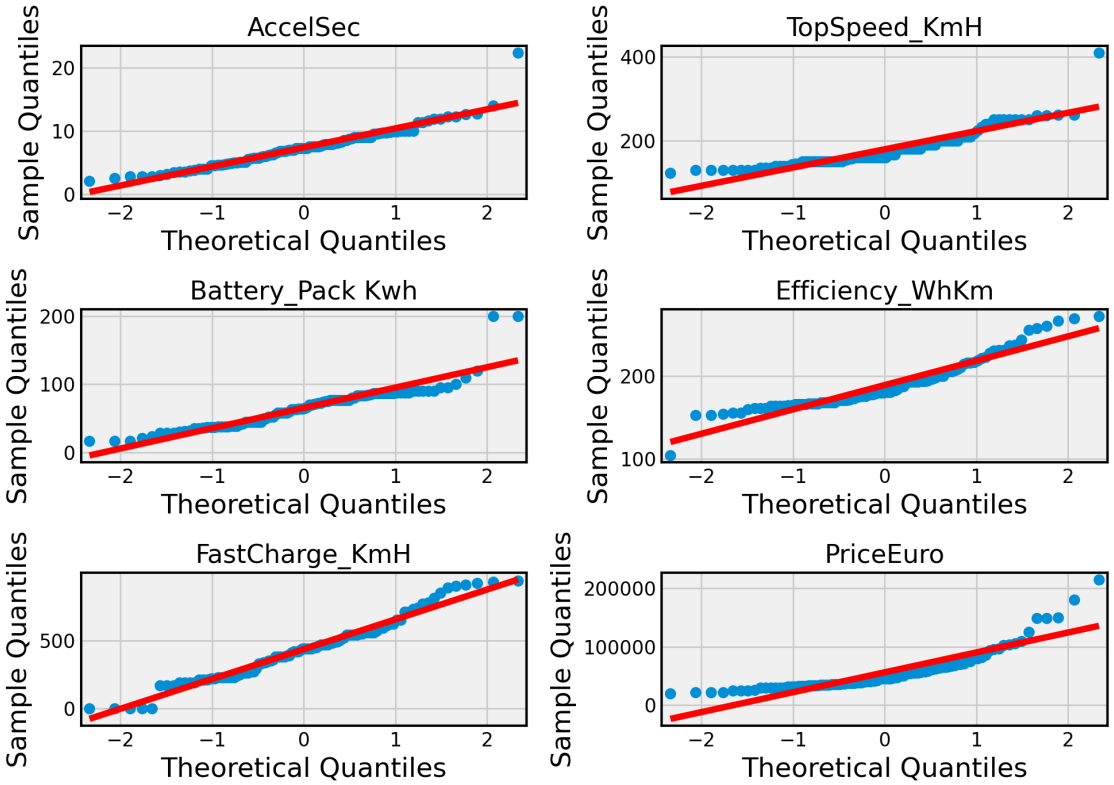

Figure 3

4 Transformations

Analyzing most of the residuals we can see that most of them are random, normally distributed, independent, and show homoscedasticity. However, these conditions are not met for the covariates Battery pack and Price in Euro. For this reason, we decided that it is important to transform Battery pack and Price in Euro.

We tested two conventional transformations. First, we applied square root transformations on Battery pack and Price in Euro. Next, we applied log transformations to the two covariates. We determined that the log transformations led to a stronger linear relationship (R^2) so our final transformation was a log transform on Battery Pack and Price. Figure 4 shows the improved residual plots after the log transformation:

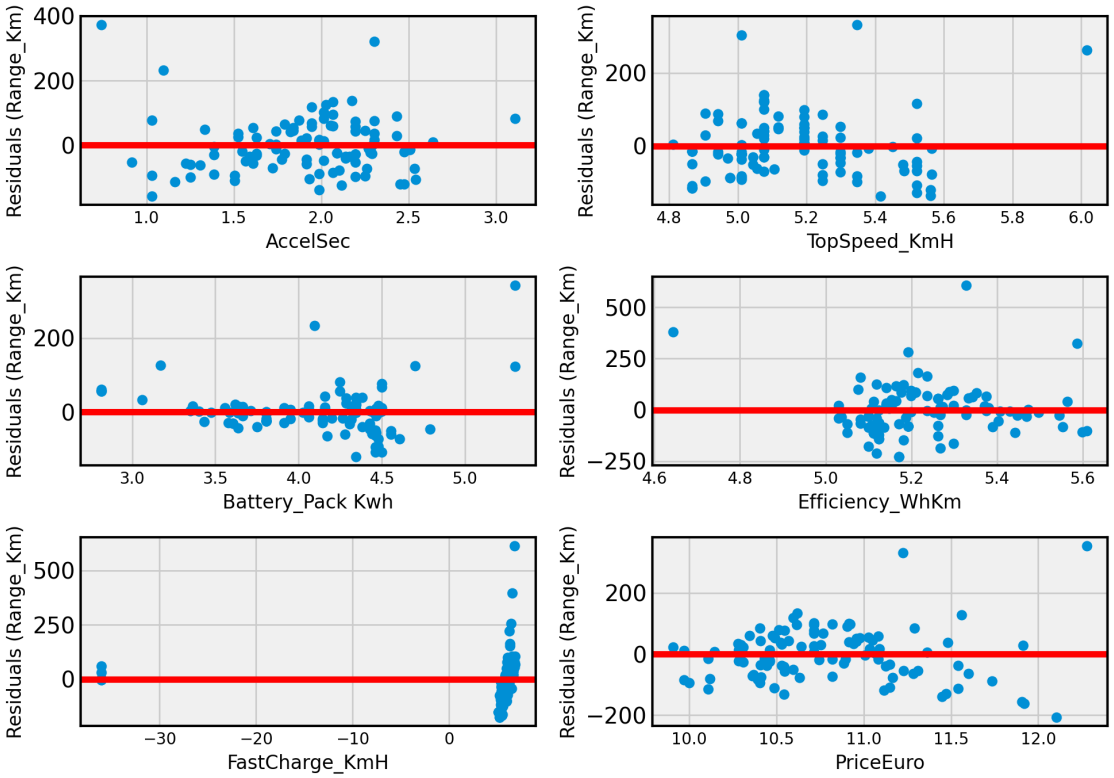

Figure 4

Covariate	R^2 Before	R^2 After
AccelSec	0.46	0.539
TopSpeed_KmH	0.56	0.552
Battery_Pack Kwh	0.829	0.768
Efficiency_WhKm	0.098	0.087
FastCharge_KmH	0.569	0.187
PriceEuro	0.458	0.529

Table 2

5 Main Results

5.1 Multicollinearity

We created a correlation coefficient matrix (Figure 5) in order to test for multicollinearity within our covariates. Range was highly correlated with all of our covariates which is a good sign that our model is predictive of range. We noticed that acceleration has a negative correlation with most of the other variables including range. The efficiency coefficient had the lowest correlation coefficient with range at 0.48. There is high correlation between price and our other covariates, and we believe there is evidence of multicollinearity because of the high correlation coefficient. Removing price may reduce the multicollinearity issues we encounter. It may be best to use a stepwise forward regression, to allow the model to remove covariates that are causing multicollinearity.

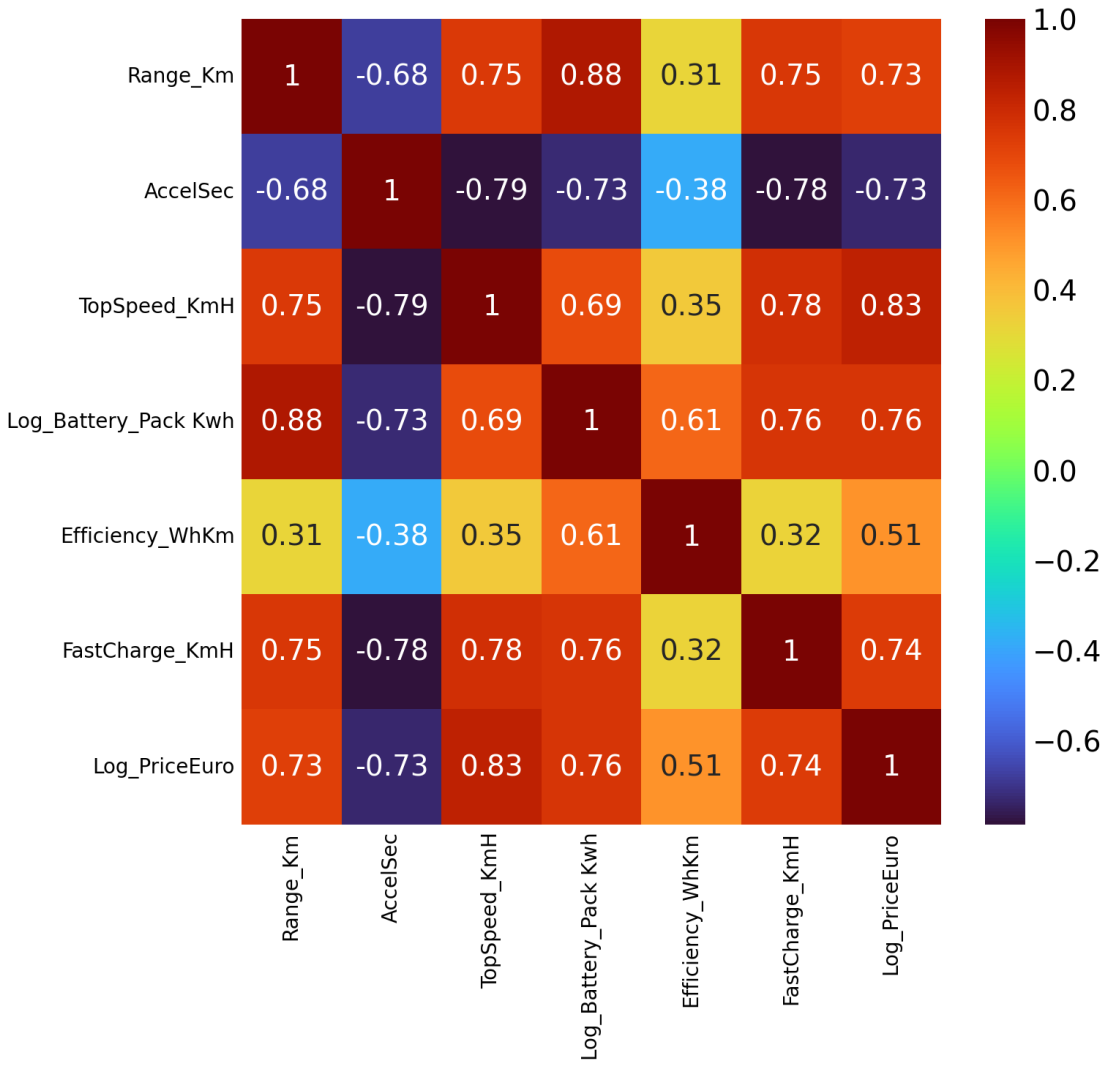

Figure 5

With the covariates that remained after feature selection, we measured multicollinearity using the variance inflation factor (VIF) given by:

$$VIF_i = \frac{1}{1 - {R_i}^2}$$

The result (Table 3) shows that all of the selected covariates are below 5, which means that they will not have multicollinearity. The variable TopSpeed is less significant than the other covariates, but because it is significant at the 5% level and has a VIF level below 10, we believe it should still be included for analysis.

Term

Estimate

Std Error

t Ratio

Prob > |t|

VIF

Intercept

-511.9363

34.95776

-14.64

$ < .0001*$

.

TopSpeed_KmH

0.4494041

0.175015

2.57

0.0120*

1.9033032

Log(Battery_Pack Kwh)

235.09752

14.19828

16.56

$ < .0001< /td >< td > 2.7759986 < /td >< /tr >< tr >< td > Efficiency_W hKm < /td >< td > −1.013424 < /td >< td > 0.184575 < /td >< td > −5.49 < /td >< td > < .0001$

1.8698509

Table 3

5.2 Variable Selection

Method

Prob to enter/leave

Equation

R_a^2

Forward Selection

PE: 0.25 PL: 0.1

$\hat{y} = -909.39 + 0.29x_1 + 39.04x_2 - 2.08x_3 + 288.90x_4 - 28.80x_5 + 8.22x_6 - 45.13x_7 - 52.62x_8 + 70.71x_9 + 43.78x_{10}$

0.941

Backward Elimination

PE: 0.25 PL: 0.1

$\hat{y} = -909.39 + 0.29x_1 + 39.04x_2 - 2.08x_3 + 288.90x_4 - 28.80x_5 + 8.22x_6 - 45.13x_7 - 52.62x_8 + 70.71x_9 + 43.78x_{10}$

0.941

Mixed

PE: 0.25 PL:0.25

$\hat{y} = -909.39 + 0.29x_1 + 39.04x_2 - 2.08x_3 + 288.90x_4 - 28.80x_5 + 8.22x_6 - 45.13x_7 - 52.62x_8 + 70.71x_9 + 43.78x_{10}$

0.941

AICC

N/A

$\hat{y} = -909.39 + 0.29x_1 + 39.04x_2 - 2.08x_3 + 288.90x_4 - 28.80x_5 + 8.22x_6 - 45.13x_7 - 52.62x_8 + 70.71x_9 + 43.78x_{10}$

0.941

BIC

N/A

$\hat{y} = -909.39 + 0.29x_1 + 39.04x_2 - 2.08x_3 + 288.90x_4 - 28.80x_5 + 8.22x_6 - 45.13x_7 - 52.62x_8 + 70.71x_9 + 43.78x_{10}$

0.941

Table 4

5.3 Model Fit

The transformations that we included in our model included the log of price and log of battery pack. We achieved a perfect ${R_a}^2$ value and realized that the car Model and brand covariate had a high cardinality (factor levels were equal to the number of data points in this case), hence, causing overfitting (over-generalization) of the fitted model. For this reason, we removed the a Model and brand covariate and were left with a much more applicable adjusted ${R_a}^2$ and low RMSE that predicts range.

Summary of Fit	
RSquare	0.954078
RSquare Adj	0.949031
Root Mean Square Error	28.60422

Summary of Fit	
Mean of Response	338.6275
Observations(or Sum Wgts)	102

Table 5

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	10	1546901.5	154690	189.0612
Error	91	74456.3	818	**Prob > F**
C. Total	101	1621357.8		<.0001*

Table 6

```
<table style="width: 100%">
    <tr>
        <td>Term</td>
        <td>Estimate</td>
        <td>Std Error</td>
        <td>t Ratio</td>
        <td>Prob $ > |t| $</td>
    </tr>
    <tr>
        <td>Intercept</td>
        <td>-909.3919</td>
        <td>117.3835</td>
        <td>-7.75</td>
        <td>$ < .0001$</td>
    </tr>
    <tr>
        <td>TopSpeed_KmH</td>
        <td>0.2942874</td>
        <td>0.140142</td>
        <td>2.10</td>
        <td>$0.0385$</td>
    </tr>
    <tr>
        <td>RapidCharge[No]</td>
        <td>39.042448</td>
        <td>8.552891</td>
        <td>4.56</td>
        <td>$ < .0001$</td>
    </tr>
    <tr>
        <td>Efficiency_WhKm</td>
        <td>-2.078849</td>
        <td>0.161668</td>
```

```
    <td>-12.86</td>
    <td>$ < .0001$</td>
</tr>
<tr>
    <td>Log(Battery_Pack Kwh)</td>
    <td>288.89749</td>
    <td>14.30833</td>
    <td>20.19</td>
    <td>$ < .0001*$</td>
</tr>
<tr>
    <td>BodyStyle{SPV&amp;amp;Hatchback&amp;amp;MPV&amp;amp;SUV-Station&amp;amp;Sedan&amp;amp;Cabrio&amp;amp;Liftba
    <td>-28.80209</td>
    <td>5.374477</td>
    <td>-5.36</td>
    <td>$ < .0001*$</td>
</tr>
<tr>
    <td>BodyStyle{SPV&amp;amp;Hatchback-MPV&amp;amp;SUV}</td>
    <td>8.2153396</td>
    <td>3.623798</td>
    <td>2.27</td>
    <td>$0.0258$</td>
</tr>
<tr>
    <td>BodyStyle{Station&amp;amp;Sedan&amp;amp;Cabrio&amp;amp;Liftback-Pickup}</td>
    <td>-45.12672</td>
    <td>12.9488</td>
    <td>-3.49</td>
    <td>$0.0008*$</td>
</tr>
<tr>
    <td>BodyStyle{Station&amp;amp;Sedan-Cabrio&amp;amp;Liftback}</td>
    <td>-52.61664</td>
    <td>6.97113</td>
    <td>-7.55</td>
    <td>$ < .0001*$</td>
</tr>
<tr>
    <td>BodyStyle{Cabrio-Liftback}</td>
    <td>70.712853</td>
    <td>11.46589</td>
    <td>6.17</td>
    <td>$ < .0001*$</td>
</tr>
<tr>
    <td>Log(PriceEuro)</td>
    <td>43.782414</td>
```

```
        <td>14.52427</td>
        <td>3.01</td>
        <td>$0.0033^*$</td>
    </tr>
</table>
```

Table 7

After we used forward selection as our variable selection method and set p-value enter is 0.25 and p-value leave as 0.1, we determined our final model to be as follows:

$\hat{y} = -909.39 + 0.29x_1 + 39.04x_2 - 2.08x_3 + 288.90x_4 - 28.80x_5 + 8.22x_6 - 45.13x_7 - 52.62x_8 + 70.71x_9 + 43.78x_{10}$ where x

x_1	Topspeed_KmH
x_2	Rapid Charge
x_3	Efficiency_WhKm
x_4	log(Battery_Pack Kwh)
x_5	BodyStyle{SPV&Hatchback&MPV&SUV-Station&Sedan&Cabriolet&Lift back&Pickup}
x_6	BodyStyle{SPV&Hatchback-MPV&SUV}
x_7	BodyStyle{Station&Sedan&Cabriolet&Lift back-Pickup}
x_8	BodyStyle{Station&Sedan-Cabriolet&Lift back}
x_9	BodyStyle{Cabriolet-Lift back}
x_{10}	log(PriceEuro)

Table 8

We found our final model to be adequate as it had a high ${R_a}^2$ and all of the covariates we found were significant at the $\alpha = 0.05$ level. The reason we used forward selection is to remove the issue of multicollinearity and it was the simplest and easiest model. Figure 9 shows the performance of our model. Some points to highlight are that our model's ${R_a}^2$ is 0.949, has a statistically significant F statistic of 189.061 in the ANOVA test.

5.4 Model Interpretation

There is a natural interpretation of the covariates in that each one shows us how much they affect the range of EVs. The data we got shows that the battery pack does affect the range the most. Other than the powertrain covariate, all of the data was significant at the 5% level of significance. We found how the different body styles of cars can affect the range as well.

6 Conclusion

Based on the results of our model we can see that there is a strong positive correlation between battery pack and range. This can help consumers and car manufacturers when designing new EVs to create larger battery packs and understand that with the combination of Bodystyle we can see

which are the best cars for range. For example, Cabrio or convertibles and hatchbacks have the greatest impact on range with their coefficient being 42. SUVs and pickup trucks for example have a negative effect on range.

Another interesting covariate we found was how rapid charge effects range, with it being the third largest β that we found. This makes sense as charging cars will lead to more range for the car, it could make the car more efficient. The price was also another factor that impacts range, one that makes sense as price increases, the technology and range of the EV increases as well. Our model allows car manufacturers, consumers, and investors to understand the factors of EVs that most affect range allowing for better improvement of EVs for the future, improving the environment and lowering the market for gas powered vehicles.

7 Contributions

Group met on Zoom a couple times to work on the project together. Each member had worked on some parts of the project as listed below:

Yuhang Du: Presentation

Matthew George: JMP Analysis, Introduction, Background, Conclusion

Xiangru He: JMP Analysis, Introduction, Background, Contributions

Alex Lavaee: Data Cleaning, Data Preparation, Visual Generation, Report Preparation

Yujie Yang: Presentation