Final Paper

Kotomi Oda, Kaleb Cervantes, Nikhil Taringonda

2022-05-02

Which characteristics of Electric Vehicles have a significant impact on increasing their price?

The dataset we used was the "Cheapest Electric Cars" from Kaggle user KOUSTUBHK. This user scraped data from https://ev-database.org/ in August 2021. The dataset contains 180 rows and 11 columns. Some of the columns were stored as strings, but clearly contained numeric substrings that were useful predictors. Some of these strings have the value "-" which indicates a null value. As such, these will be converted to NA when the numeric parts are parsed out. After the data was cleaned, the following figure was able to be shown.

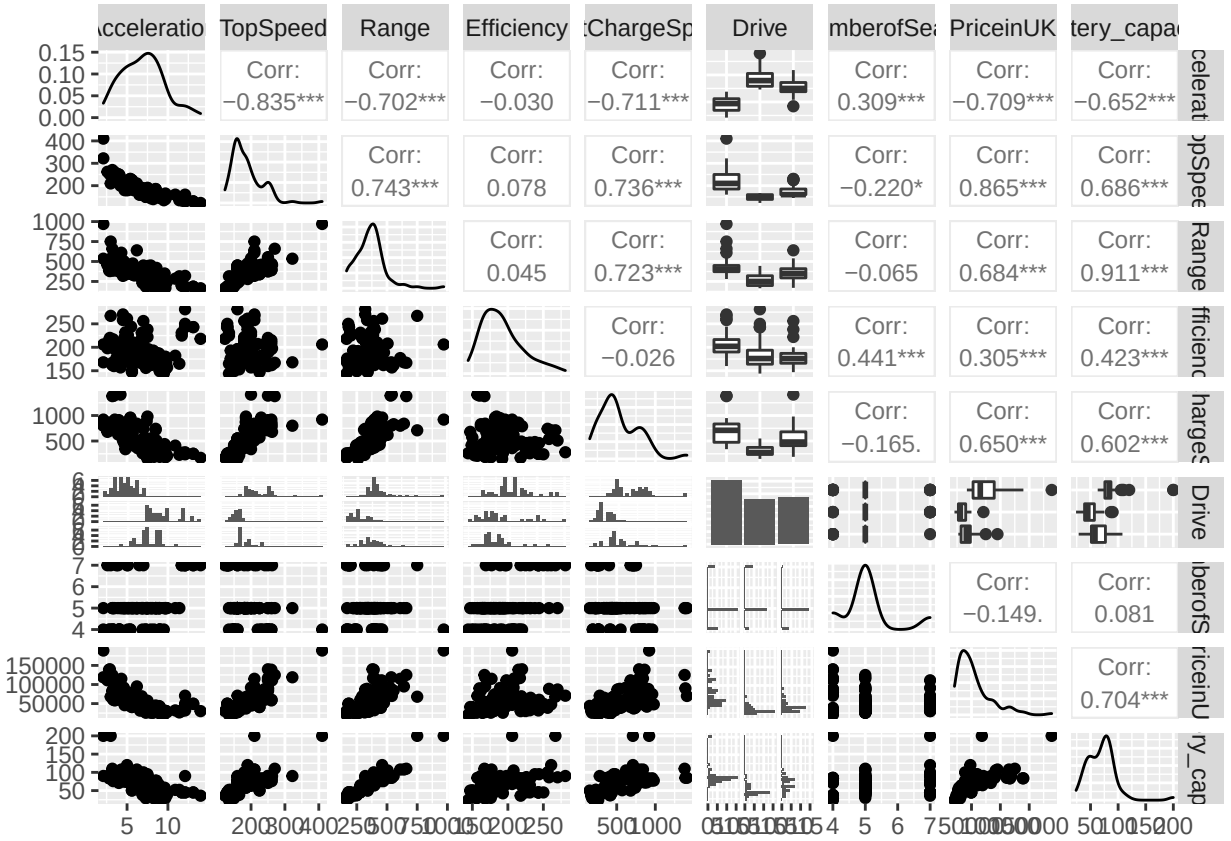

Figure 1: Correlation/Scatterplot/Density Matrix

In order to select the model, we began with a full — and untransformed — additive model. The purpose of this were to recognize which variables introduced a lot of multicolinearity. From this, the scatterplot matrix, and the correlation matrix, we were able to recognize two pairs of variables that had a lot of multicolinearity: `batery_capacity` and `Range`, `Acceleration` and `TopSpeed`. We ended up dropping the variables `battery_capacity` and `Acceleration`. This was because these predictors were less accurate in later parts of the process than `Range` and `TopSpeed` respectively.

After this, there were still possible transformations needed for the predictors. Initial predictions were found by looking at marginal plots between the predictors and the response. Other transformations of the predictors would be tested and the more accurate predictions would be kept.

There were also possible transformations needed for the response variable. This was done by using a Box-Cox Power Transformation. In this case, we got $\lambda \approx -0.5$, which corresponds to an inverse square root transformation.

After this, there were still some insignificant predictors remaining. In order to choose the significant ones, stepwise selection — with BIC as the metric — was utilized. This resulted in the final model:

$$\begin{aligned}\frac{1}{\sqrt{\texttt{PriceinUK}}} &= \beta_0 \\ &+ \beta_1 \ln \texttt{Range} \\ &+ \beta_2 \texttt{TopSpeed} \\ &+ \beta_3 \texttt{TopSpeed}^2 \\ &+ \beta_4 \texttt{Efficiency} \\ &+ \beta_5 \texttt{Efficiency}^2 \\ &+ \epsilon\end{aligned}$$

A summary of the table can be seen below:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.01915	0.00154	12.43	3.386e-23
log(Range)	-0.0005637	0.0001617	-3.487	0.0006865
poly(TopSpeed, 2, raw = T)1	-3.424e-05	4.669e-06	-7.334	3.07e-11
poly(TopSpeed, 2, raw = T)2	4.427e-08	9.647e-09	4.589	1.12e-05
poly(Efficiency, 2, raw = T)1	-5.274e-05	1.392e-05	-3.789	0.0002394
poly(Efficiency, 2, raw = T)2	9.774e-08	3.382e-08	2.89	0.00459

Table 2: Final Model Summary

Observations	Residual Std. Error	R^2	Adjusted R^2
124	0.0003877	0.8588	0.8528

Using this model, we decided to recognize the observations that were both high residual and high leverage. These points were considered outliers.

Table 3: Outliers

Name	PriceinUK
Tesla Roadster	189000
Mercedes EQV 300 Long	70665
Tesla Cybertruck Dual Motor	48000
Tesla Cybertruck Tri Motor	68000

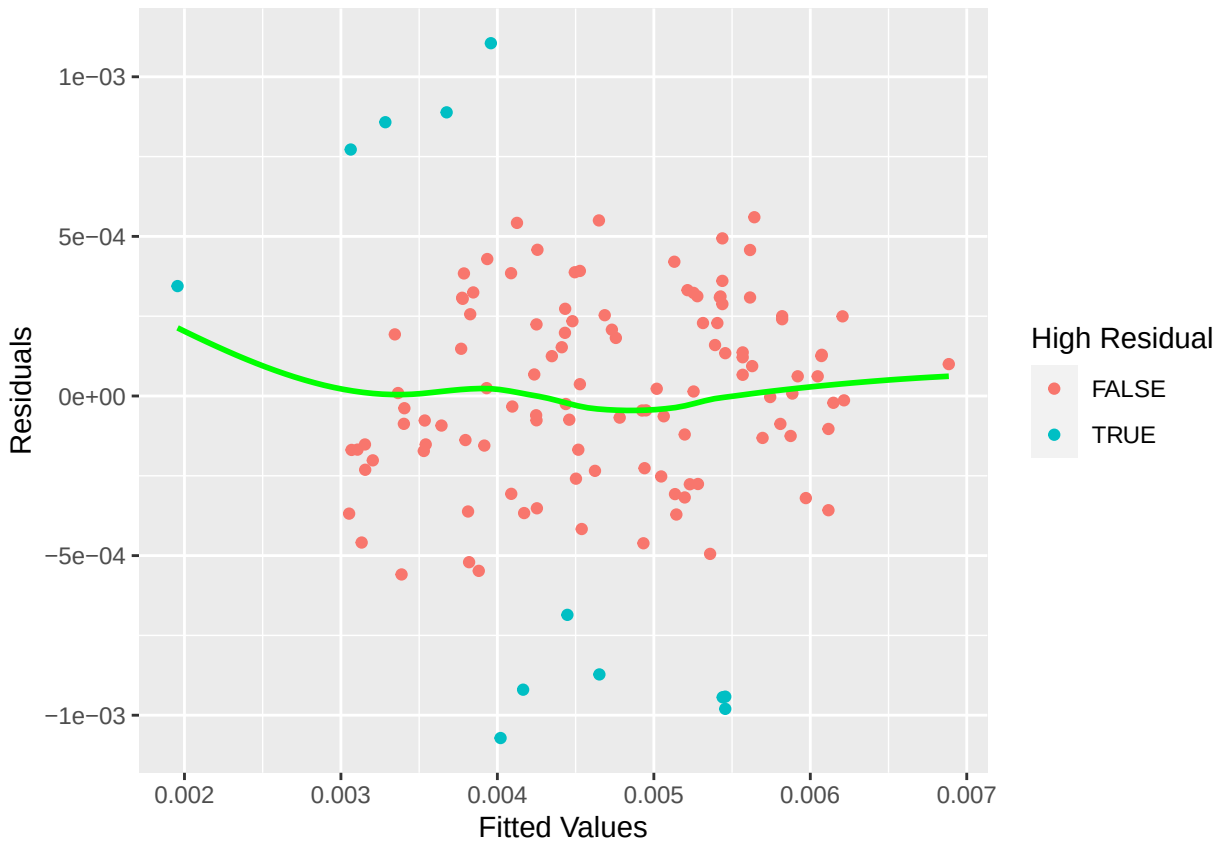

Figure 2: Residuals vs. Fitted Values Plot for final model

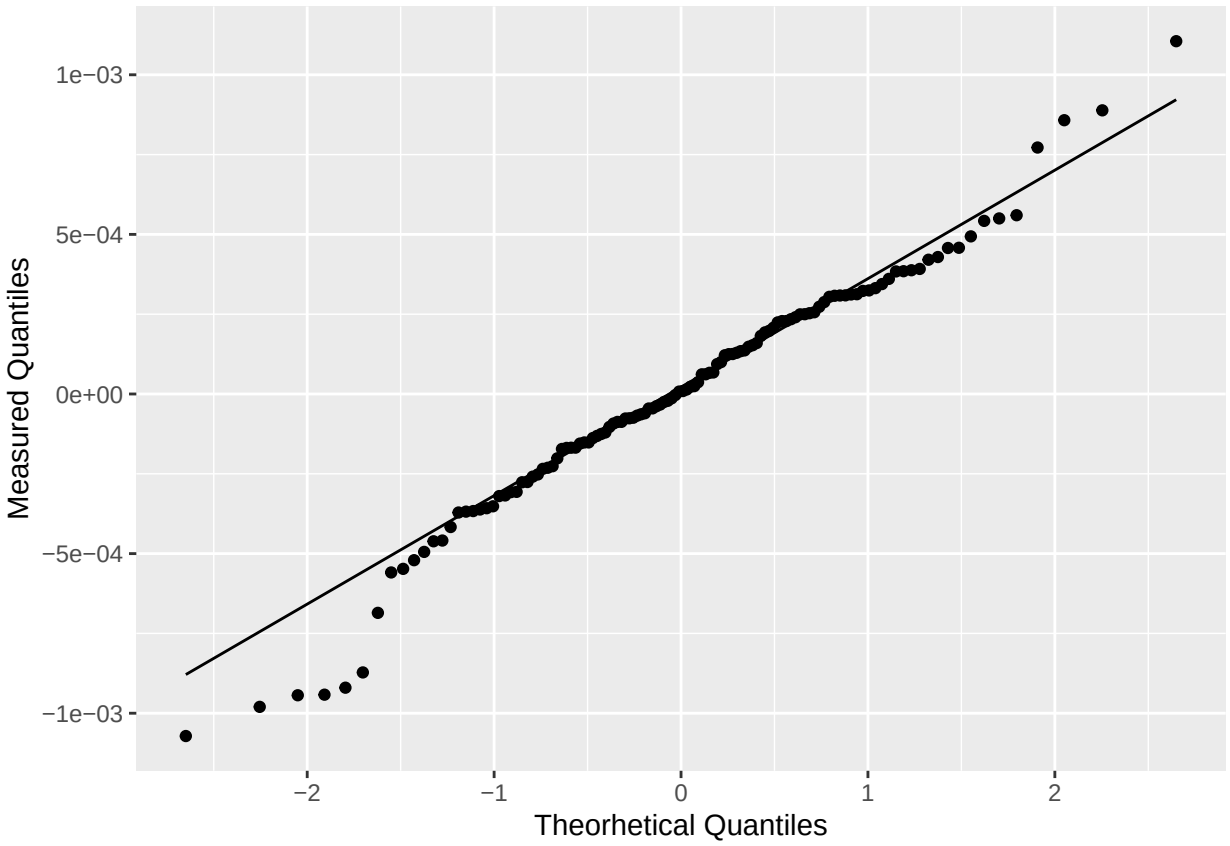

Figure 3: Q-Q Plot

Code Appendix

```
# sets up default settings for code chunks
knitr::opts_chunk$set(
  echo = F,
  message = F,
  warning = F
)

# loads necessary libraries
library(tidyverse)
library(GGally)

# load cleaned dataset and rename subtitle column
full_data <- read_csv("Dataset/cleaned_data.csv") %>%
  mutate(battery_capacity = Subtitle, .keep = "unused")

# drop incomplete observations
data <- drop_na(full_data)

# Plots figure 1
select(data, -Name, -PriceinGermany) %>%
  ggpairs(
    upper = list(

      # change font size of correlation stuff
      continuous = wrap("cor", size = 3)
    )
  )

# stepwise selection w/ BIC
final_model <- lm(
  1 / sqrt(PriceinUK) ~
    log(Range) +
    poly(TopSpeed, 2, raw = T) +
    poly(Efficiency, 2, raw = T) +
    FastChargeSpeed +
    NumberofSeats +
    Drive,
  data
) %>%
  step(trace = 0, k = nrow(data) %>% log)

# prints model summary in table format
summary(final_model) %>%
  pander::pander(caption = "Final Model Summary")

# indeces of high residual points
high_res <- rstandard(final_model) %>%
  abs() > 2

# plot for constant variance
ggplot(mapping = aes(final_model$fitted.values, final_model$residuals)) +
```

```
  geom_point(aes(color = high_res)) +
  geom_smooth(se = F, color = "green", alpha = 0.5) +
  labs(x = "Fitted Values", y = "Residuals", color = "High Residual")

# qqplot
ggplot(mapping = aes(sample = final_model$residuals)) +
  geom_qq() +
  geom_qq_line() +
  labs(x = "Theorhetical Quantiles", y = "Measured Quantiles")

# get leverage from model
fm_lev <- hatvalues(final_model)

# print desired observations
filter(
  data,
  high_res,
  abs(fm_lev) > 2 * mean(fm_lev)
) %>%
  select(Name, PriceinUK) %>%
  knitr::kable(tabel.envir = "figure", caption = "Outliers")
```