Exploration Of the Relationship of Engines & CO2 Emissions

Muhammad Abdul Mannan

Student at University of Toronto

December 17, 2021

Introduction section

Over the years, car engines have been changing in terms of design, efficiency, power and are being built with the environment taken into consideration and cars are being manufactured at a faster rate than any point in history. Some cars either require larger and more powerful engines and other cars utilize more efficient engines and due to this, we are likely going observe an impact on the CO2 emissions of the engines. For over two years now, the price of gas has been climbing and since I drive, I wanted to observe how the size of the engine and the number of cylinders influences the CO2 emissions and have developed an interest for car efficiency and it's influence on the environment. According to Fontaras, G (2017), according to lab data, there is a decline in the overall fuel consumption and CO2 emissions in Europe and this analysis is important as it will provide insight on how much CO2 emissions cars release and as a hypothesis. For this project, we will be analyzing and observing the relationship between car engine size, cylinders and CO2 emissions and observe whether the engine size and number of cylinders influences CO2 emissions.

Methods section

Initially, before doing the analysis, I had first cleaned the dataset by renaming variables and trimming down the dataset to the variables of interest. In the exploratory data analysis (EDA), we had plotted histograms as well as scatterplots of each predictor and response to check for any violated conditions.

Then we had then fitted our model and checked Condition 1 by plotting the response against the fitted values and plotting the residuals against each predictor. Then we had checked Condition 2 by viewing a series scatterplots of predictors plotted against each other and also checked the QQ-plot to check normality aswell. Based on the results, we had performed a power transformation on both the predictors and the response and then fitted the transformed model and rechecked the assumptions with the same method as before and compared the original model and the transformed model to see what had changed and improved.

A leverage point is generally a value that deviates from the rest of the data and is further away (in terms of the x- axis) and can influence the regression line resulting in inaccurate results. An outlier is a value which deviates from the rest of the data in the y direction and both leverage and outliers are capable of moving the estimated regression line towards them. Influential points can also have the potential to move the regression line but it does not need to be a leverage point or outlier. Checking for these values is important to know that there may be potential values which can cause the regression line to deviate. So, we had first checked for leverage points from our dataset by checking whether each observation's leverage, h_{ii} is greater than the cut-off, $2(\frac{p+1}{n})$ and then flagging those observations with leverages greater than that cut off and then plotting them in red with the rest of the data to check if a leverage value is flagged due to it being an extreme value or is it because it is distant from the 'center' of the observations that is derived from a combination of the predictors. We had then checked for outliers by getting each observation's standardized residual, r_i and checking whether or not it is between [-4, 4] (we used this cut off instead of the [-2, 2] since our dataset is 'large' as it has 14,253 observations. Then we plotted the outlier observations in blue along with the other observations to see if they are actual outliers or just distant from the center derived from predictors. Finally we had checked for influential points using Cook's Distance (greater than 50th percentile of F(p+1, n-p-1) cut-off), DFFITS (greater than $2\frac{p+1}{n}$ cut-off), and DFBETAS (greater than $\frac{2}{\sqrt{n}}$ cut-off).

Multicollinearity in a model result in inflated variations and misleads us to think that the model explains more variation than it actually does. However, if we are aware of multi-collinearity, we can potentially reduce or remove it all together and there are techniques that were used such as the Variance Inflation Factor, which detects multicollinearity by returning a value that indicates how inflated the variances are in each of the predictors, ideally this value should be below 5 (cut-off) for a predictor to be highly correlated and its variance to not be considered extremely inflated. Another tool is Akaike's Information Criterion (AIC) and AIC corrected (AICc) which measures how good the fit of the model is relative to its complexity and one more is the Bayesian Information Criterion (BIC) and ideally, we want a model with the smallest AIC, AICc, BIC and largest adjusted-R^2. We had checked the multicollinearity of our model using all of these techniques to check if there is multicollinearity between the two predictors.

We had used the All-Possible Subsets Selection Process and derived 3 models and checked which is the best model based on their R^2_{adj}, residuals, AIC, AICc, BIC. Finally, we had validated the model in which we divided our dataset into two halves, a training and test dataset and conducted an EDA on the training dataset, then checked the assumptions and checked if transformations were needed as done before and found the best model. Then we fitted the training model for the test dataset and checked if the regression coefficients and R^2 values were similar for both models, whether both the model had the same significant predictors, and no model violations.

Results section:
The dataset initially had 13 variables however it needed a lot of cleaning which was done in which we had renamed variables, removed observations with missing values and removed the unwanted variables and was left with 3 variables, Engine Size and Cylinders (predictors) and CO2 Emissions (response) and 14,253 observations. Figure 1 is the histograms of the variables and in the scatterplots (predictor vs response), we found that they had slight curves and were not totally linear which also indicated that a transformation was necessary but overall, the model was looking decent.

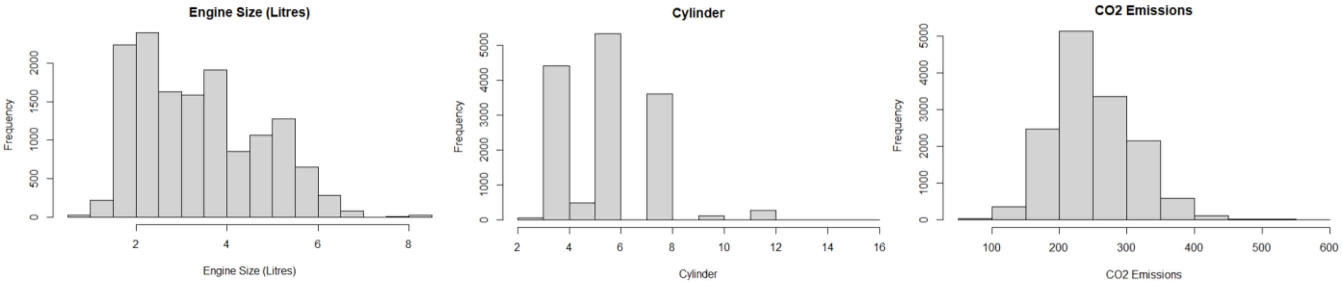

Figure 1: The histograms of engine size and Cylinder seems to have a normal distribution that is slightly left skewed indicating possible model violations and a potential transformation is needed.

While checking for model violations, we first checked condition 2 to by plotting scatterplots of the predictors to see if they were reasonably linear and then checked condition 1 by plotting the response vs fitted values and also noted that it also had a slight curve at the ends and also plotted scatterplots of the predictors vs residuals and noted that they were random with no pattern or clusters, which was good. The QQ-plot had also shown deviations at the tails which all indicated to performing a transformation to fix the violations.

A power transformation on the predictors and responses was conducted and all variables had to be transformed ((engine size)^0.18, (cylinders)^-0.12, and (co2 emissions)^0.33) and we created a transformed model (mod2) and had re-checked the assumptions as done to the original model and had observed the following plots:

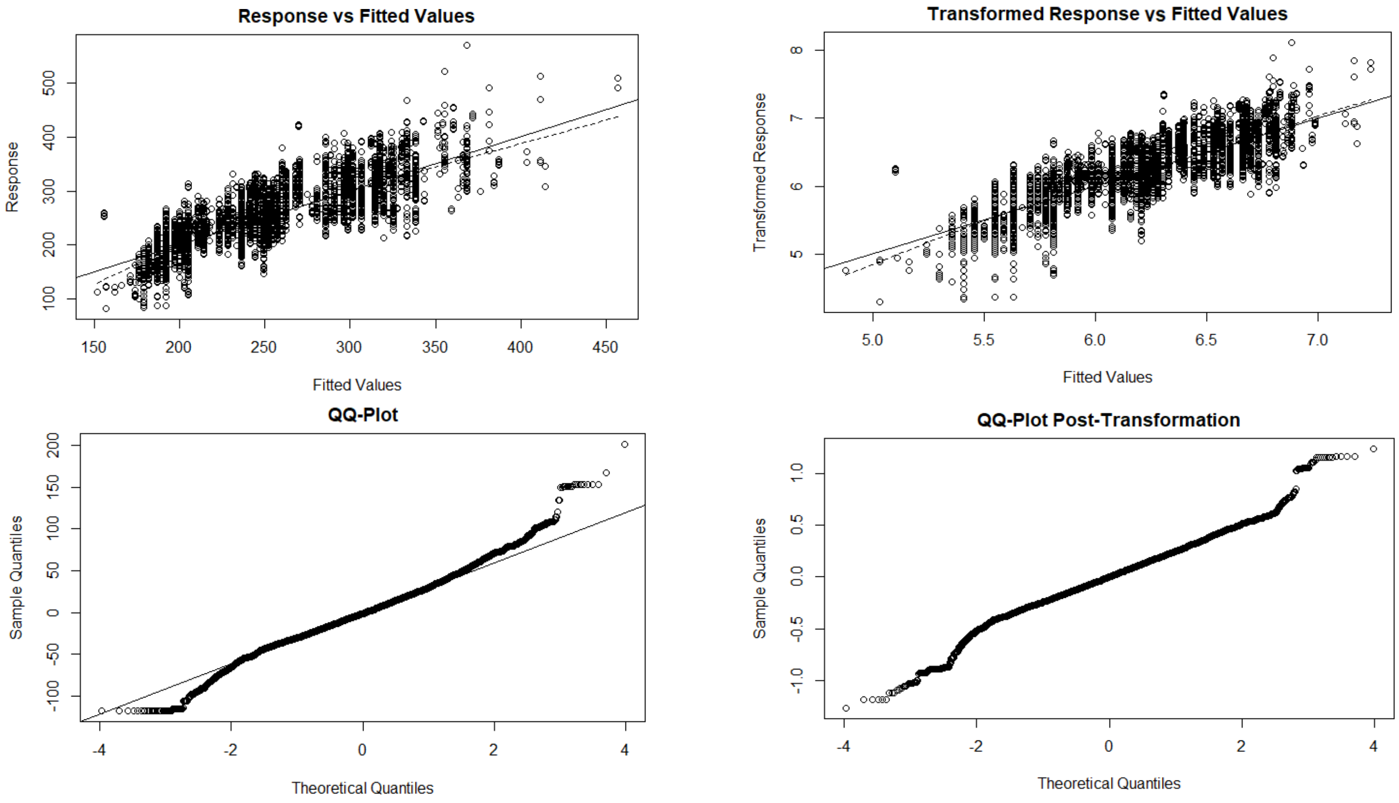

Figure 2: The Response vs Fitted Values had a slight curve as seen in the graph but once the graph was transformed, the curve had become more linear. In the QQ-plot, we could see heavy tails whereas in the transformed QQ-plot seems to follow the reference line and is more normal. That is, we found that normality and linearity after the transformation was better than before.

Next, we had checked for problematic observations. We began with leverage measurements and found that there were 782 observations that were leverage points and had deemed most of them to be non-extreme values. Then we had checked for outliers and found that we had 42 outliers and based on the graph, we can see that only ~6 can be deemed extreme. Finally, we had checked the influential points and 757 observations were flagged. We can visually see these results in Figure 3:

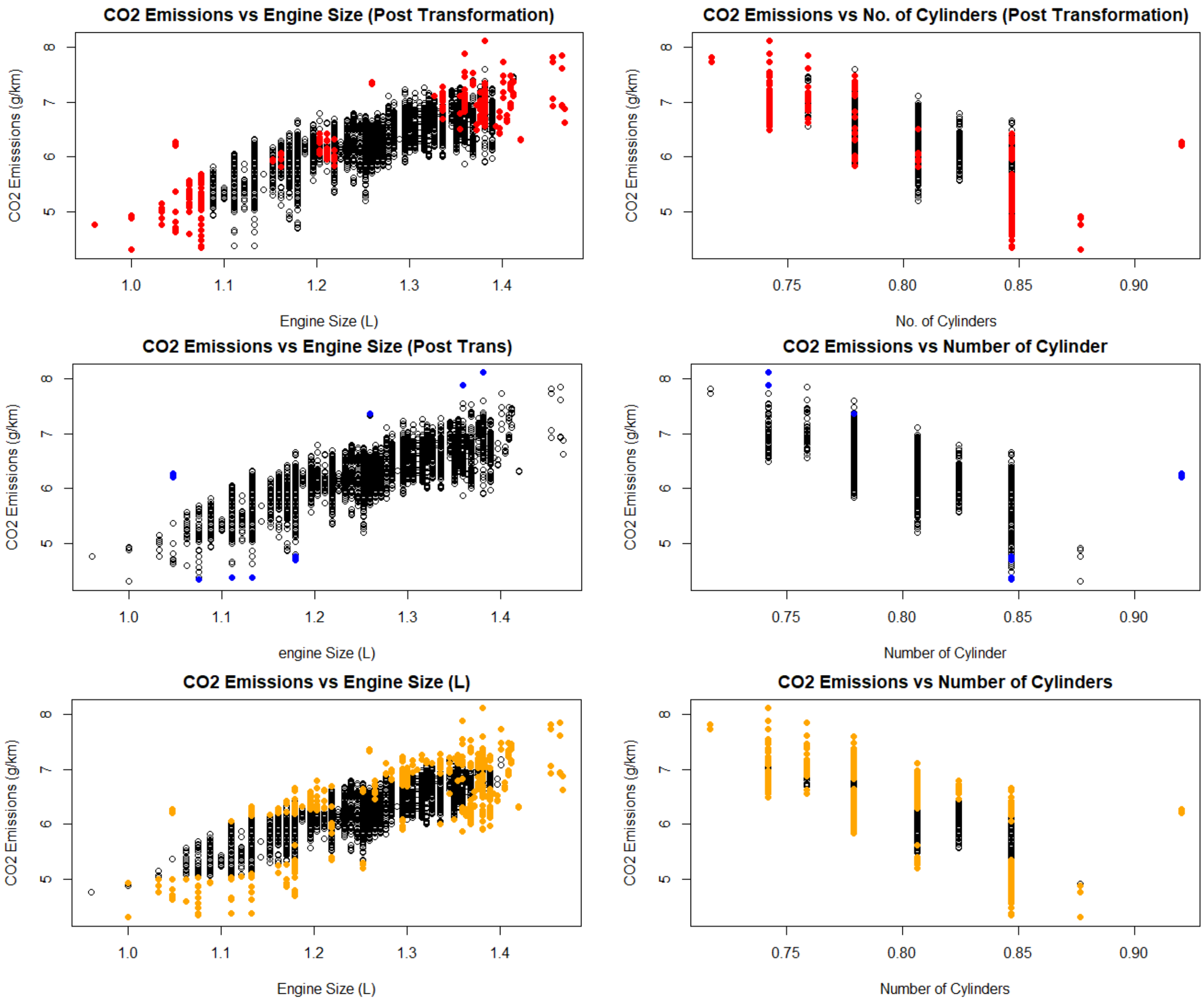

Figure 3: From the graph, we can note that most of the observations were not extreme and this is likely due to the dataset being very large and so most values are likely near one center and due to this, many points have gotten flagged as leverages, outliers, and influential and overall, the extreme values will not have a significant effect on the regression model again, due to the size of the dataset.

Multicollinearity was checked along side the all-possible subsets selection process and can see the observations in the following table:

Model	R^2	R^2_{adj}	AIC	AICc	BIC
Original Model	0.6945	0.6945	99,371	99,371	99,405
Transformed Model	0.7077	0.7077	-38,269	-38,369	-38,235
Model (cylinders omitted)	0.7038	0.7038	-38,083	-38,083	-38,057
Model (engine size omitted)	0.6328	0.6328	-35,020	-35,020	-34,992

Figure 4: We noted that removing cylinders and engine size will negatively influence the model so we had decided to not remove either one. We can see that the transformed model gives the most optimal model with the predictors explain ~71% of the variation with all assumptions held and all variables are significant via Anova t-test.

Finally, we had to validate the model so we conducted the steps as stated in the methods section and ended up with the following:

Model	**Regression Coefficients**	**R^2**	**Model Violations**	**Mean & SD (CO2 Emissions)**	**Mean & SD (Engine Size)**	**Mean & SD (Cylinders)**
Training Model	Intercept: 3.910 Engine size: 3.801 Cylinders: -3.062	0.7021	None	Mean: 6.13 SD: 0.49	Mean: 0.81 SD: 0.03	Mean: 1.24 SD: 0.09
Test Model	Intercept: 3.058 Engine size: 4.017 Cylinders: -2.331	0.7137	None	Mean: 6.14 SD: 0.48	Mean: 0.81 SD: 0.03	Mean: 1.24 SD: 0.09

Figure 5: We have chosen and we found that the training and test models were quite similar and had similar regression coefficients, mean and standard deviations (SD), significant predictors, no model violations and had similar R^2 values aswell, which fulfills the requirements for a model to be considered validated.

Hence, based on these results and our previous analysis, our model is validated.

Discussion section

Our final model is

$$(\text{CO2 Emissions})^{0.33} = 3.909(\text{Engine Size})^{0.18} - 2.697(\text{Cylinders})^{-0.12} + 3.484$$

This model depicts the relationship between CO2 emissions, engine size and cylinders. We can see that for each increase in engine size (in litres), CO2 emissions will go up by 3.909 g/km when number of cylinders is fixed. We can also note that as number of cylinders increases by one unit, CO2 emissions drop by 2.697 g/km when engine size is held constant. This shows that engine size and number of cylinders there are in a car does have a significant influence on its CO2 emissions. We can use this equation to estimate the CO2 emissions of a car given its engine size and number of cylinders.

For example:

A Toyota Corolla has a 1.8 L Engine which has 4 cylinders. Then using this equation:

$$(\text{CO2 Emissions})^{0.33} = 3.909(\text{Engine Size})^{0.18} - 2.697(\text{Cylinders})^{-0.12} + 3.484$$

$$\text{CO2 Emissions} = (3.909(1.8)^{0.18} - 2.697(4)^{-0.12} + 3.484)^{1/0.33}$$

$$\text{CO2 Emissions} = 179.63 \text{ g/km}$$

We had gotten 179.63 g/km and the official co2 emission for this car is 112 g/km (Errity, S., 2021). Although our model has over estimated it, it is still a decent estimate given that the actual calculations of CO2 emissions account for many more factors.

References

Errity, S. (2021, June 10). *Toyota Corolla Hybrid MPG & CO2 Emissions*. DrivingElectric. Retrieved December 17, 2021, from https://www.drivingelectric.com/toyota/corolla/mpg

Fontaras, G., Zacharof, N.-G., & Ciuffo, B. (2017). Fuel consumption and CO2 emissions from passenger cars in Europe – Laboratory versus real-world emissions. Progress in Energy and Combustion Science, 60, 97–131. https://doi.org/10.1016/j.pecs.2016.12.004