Group C7: Teevint Prak, Gloria Yi, Roberto Mulliadi, Emma Oh

Exploring the Impact of Vehicle Specifications on Carbon Dioxide Emissions: A Comparative Analysis of Regression Models

Introduction

Carbon footprints represent the total amount of greenhouse gasses released in the atmosphere and are one of the leading environmental indicators of global climate change. Individual carbon footprints are based on lifestyle habits by a particular person, taking account of their diet, means of transportation, household energy, consumption and much more. The transportation sector is responsible for about 28% of Canada's greenhouse gas emissions (Statistics Canada, 2021), as vehicles burn fossil fuels like gasoline and diesel, and release a range of gasses – one of which is carbon dioxide.

The dataset that will be studied throughout this report is provided by the Government of Canada's open data initiative from the open.canada.ca platform, where the following link directs to the data itself: 2024 Fuel Consumption Ratings. This dataset encompasses model-specific fuel consumption ratings and estimated carbon dioxide emissions data for new light-duty vehicles that are marketed for the year of 2024 available in the Canadian retail market. Published on February 19th 2024, this dataset contains values contained that are derived from initial ratings rather than direct vehicle testing; offering approximations for research purposes (Natural Resources Canada, 2023).

Vehicle testing was performed to approximate the data, where the process is described as follows: first, a vehicle being driven about 6,000 km before the testing, then the test vehicle is placed on a chassis dynamometer, which is a machine that acts like a treadmill for vehicles. This machine is adjusted for specifications such as weight and the aerodynamics of the respective vehicle. Then, a driver runs the vehicle through standard driving cycles in which trips in the city and on the highway are simulated. The city and highway fuel consumption ratings are obtained from the emissions generated during the following laboratory driving cycles:

- City test
- Highway test
- Cold temperature operation
- Air conditioner use
- Higher speeds with more rapid acceleration and braking

(Natural Resources Canada, 2023).

Table 1 below summarizes the specifics on what was recorded in the given dataset, outlining the name of the variable, the type – categorical or continuous, and its description. Refer to Appendix A to view the breakdown of all the levels of the categorical variable, if applicable.

Table 1: Description of Variables

Name	**Type**	**Description**
Model year	Categorical	The year used by the manufacturer to designate a model of vehicle.
Make	Categorical	The manufacturer of the vehicle.
Model	Categorical	The model name of the vehicle.
Vehicle Class	Categorical	Classification of the vehicle based on interior volume for cars and gross vehicle weight rating for light trucks.
Engine size (L)	Continuous	Total displacement of all cylinders (in litres).
Cylinders	Categorical	Number of engine cylinders.
Transmission	Categorical	The type of transmission and number of gears/speeds.
Fuel Type	Categorical	The type of fuel used to power the vehicle.
City (L/100km)	Continuous	City fuel consumption rating shown in litres per 100 kilometres. Represents urban driving in stop-and-go traffic
Highway (L/100 km)	Continuous	Highway fuel consumption rating shown in litres per 100 kilometres. Represents a mix of open highway and rural road driving, typical of longer trips
Combined (L/100 km)	Continuous	Combined fuel consumption rating shown in litres per 100 kilometres.
Combined (mpg)	Continuous	The combined rating expressed in miles per imperial gallon. Reflects 55% city driving and 45% highway driving
CO2 emissions (g/km)	Continuous	The vehicle's tailpipe emissions of carbon dioxide shown in grams per kilometre for combined city and highway driving.
CO2 rating	Categorical	The vehicle's tailpipe emissions of carbon dioxide are rated on a scale from 1 (worst) to 10 (best).
Smog Rating	Categorical	The vehicle's tailpipe emissions of smog-forming pollutants are rated on a scale from 1 (worst) to 10 (best).

The primary goal of our project is to investigate the question: *what is the relationship between certain vehicle specifications and carbon dioxide emissions (grams/km)?*

By studying the relationship between these vehicle factors and the total carbon emission, it can aid in identifying fuel-efficient vehicles and various ways we can take part in mitigating the effects on the environment.

Analysis

Exploratory Data Analysis - Continuous Variables

To perform an exploratory data analysis on the continuous variables of our dataset, we began with calculating the summary statistics, shown below in Table 2 and plotting the histograms of each depicting the distribution in Figure 1.

Table 2: Summary statistics of Continuous Variables

Continuous Variable	Min	1st Quantile	Median	Mean	3rd Quantile	Max
Engine_Size_L	1.20	2.00	2.90	3.08	3.60	8.00
City_L_100km	4.40	10.10	12.20	12.39	14.50	30.30
Highway_L_100km	4.40	7.7	9.3	9.43	10.80	20.90
Combined_L_100km	4.40	9.00	11.00	11.06	12.70	26.10
Combined_MPG	11.00	22.00	26.00	27.39	31.00	64.00
CO2_Emissions_g_km	104.00	210.00	260.00	258.80	299.00	608.00

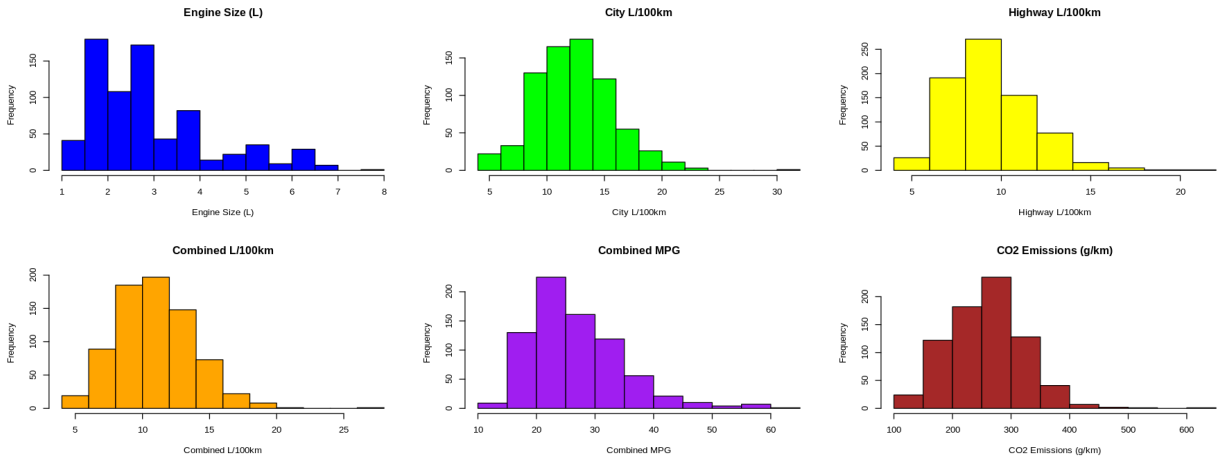

Figure 1: Distribution of Continuous Variables

From the observed histograms above, we can gather some of the following insights:

1. **Skewness:** Right-skewed distributions for engine size, fuel consumption, and CO2 emissions suggest the presence of outliers with high values that can potentially influence any predictive modeling.
2. **Fuel Efficiency:** Most vehicles are concentrated in the mid-range of fuel consumption, both in the city and on the highway, indicating that there might be a common standard of fuel efficiency among the sampled vehicles.
3. **CO2 Emissions:** There is a wide range of CO2 emissions, but with a concentration of vehicles in the lower emission range. This suggests that while there are vehicles that are not very environmentally friendly, the majority have lower emissions.
4. **Potential Transformations:** Given the skewness in some of these distributions, transformations such as logarithmic scaling might be necessary before using these variables in linear regression models to meet the assumption of normality of residuals.

Table 3: Summary Statistics for CO2 Emissions

Summary Statistic	**Value (g/km)**
Mean	258.80
Median	260
Standard Deviation	65.12

To grasp a better understanding of the continuous variables in our data, we plot the correlation matrix below in Figure 2.

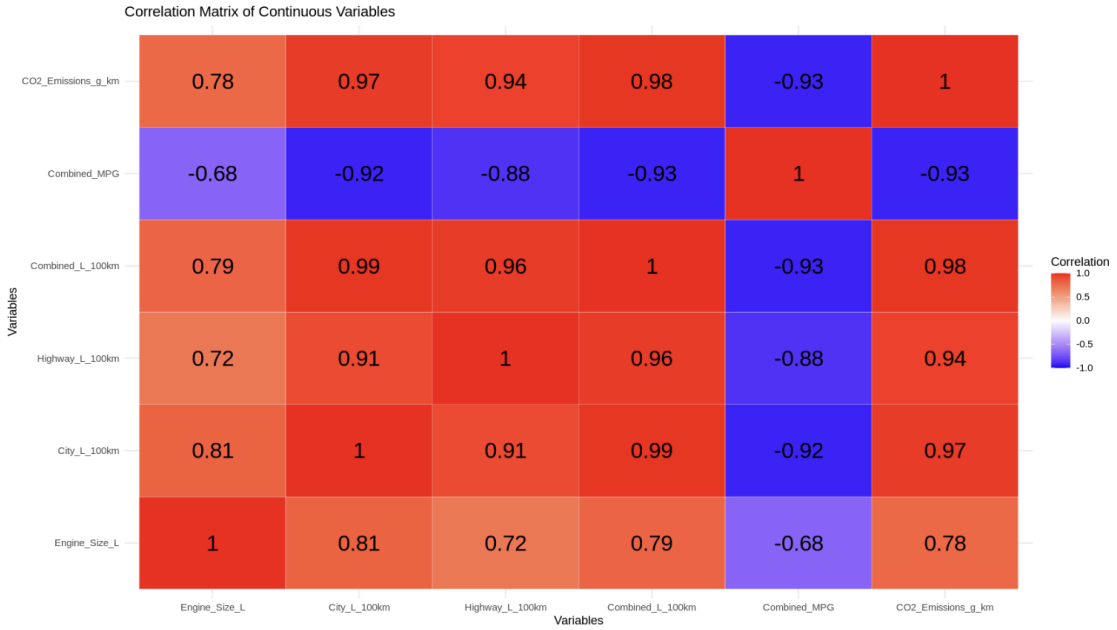

Figure 2: Pairwise Correlation Matrix between Vehicle Characteristics and CO2 emissions

To address initial multicollinearity concerns shown above in figure 2, we will be removing City_L_100km and Highway_L_100km variables from the data and only include Combined_L_100km as it represents the overall fuel consumption of each car. We observed that all cars in the data are 2024 models and that the data mostly consists of car models with one observation each, which can lead to overfitting since the model will learn noise specific to unique values instead of capturing the underlying pattern of the data to be able to generalize. Therefore we will not be including Model_Year and Model in the model selection process either. Finally, the last two variables to be removed before carrying out model selection are Smog_Rating and CO2_Rating, because they are directly derived from and highly correlated with CO2 emissions, which is the response variable we are trying to predict.

Exploratory Data Analysis - Categorical Variables

Continuing the exploratory data analysis on the categorical variables, we plotted side-by-side boxplots between CO2 Emissions and Vehicle Class, Cylinders, Transmission and Fuel Type in Figure 3.

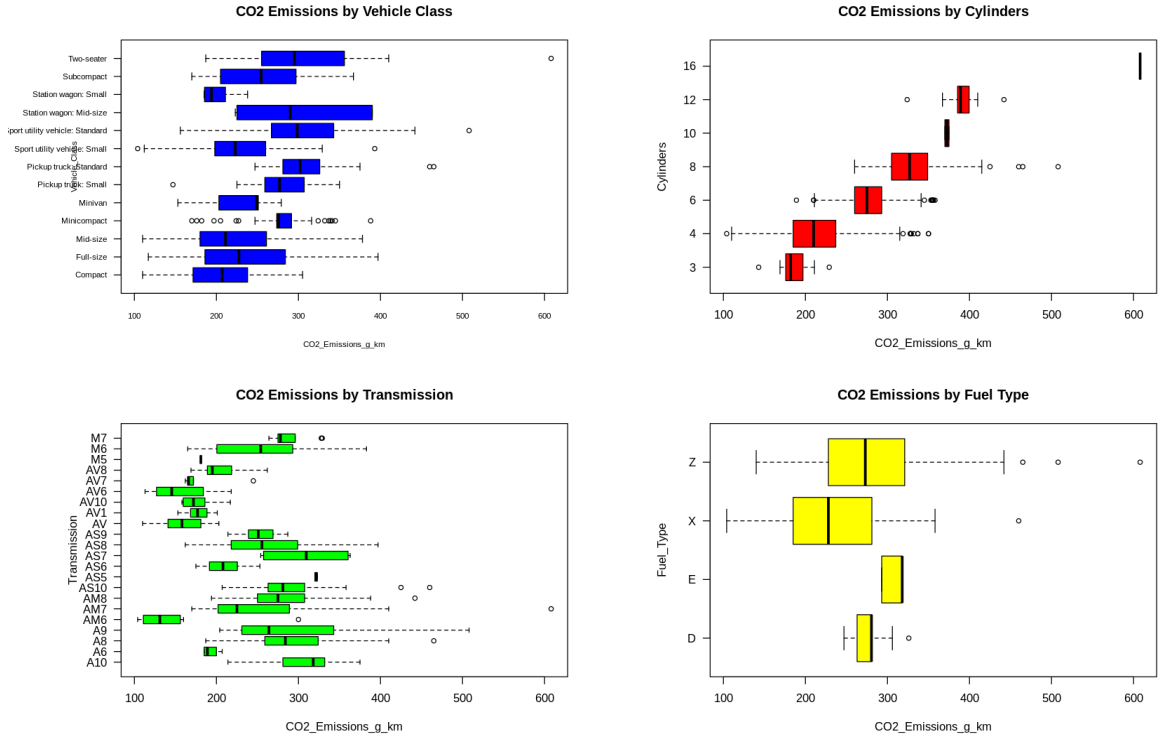

Figure 3: Box Plot between CO2 Emissions and Categorical Variables

From Figure 3 we saw that the mean CO2 emission for different levels in the categorical variables also differs a lot, suggesting that we should probably take them into account as explanatory variables in our model.

However, for the variable Transmission, we found that there are many categories that have low counts, as well as a high number of dummy variables that we can reduce to avoid unnecessary

complexity and to avoid overfitting. So we grouped the categories in Transmission into two: automatic and manual cars (represented by "A" and "M" respectively), which is a very common distinction in the real world.

It was the same case for the Make variable, where we have certain categories that are underrepresented and so we grouped these categories into two just like before, but this time we have the factors Luxury and Non-Luxury vehicles. The classification is based on the real-world market.

Model Selection: Fitting the Full Model

After removing certain features and transforming some data on Transmission and Make, we first fit CO2_emissions_g_km against all of the remaining variables, creating the full model as a baseline for our model comparisons. It was observed that the variables Intercept, Vehicle_classMinicompact, Cylinders10, Fuel_Type, Combined_L_100km and Combined_MPG were the only significant estimated coefficients. We then evaluated the VIF (Variance Inflation Factor) to be sure that we have no multicollinearity issues. We found that all variables included in the full model have values less than 5, and so we have no prominent multicollinearity problems. Hence, we will move on to carrying out the stepwise selection to consider other models that may better fit our data.

Stepwise Selection

Our data includes a high number of variables, and so we chose to carry out backward selection from the full model because it may be less prone to overfitting compared to forward selection since we are starting with the full model. We will not be doing exhaustive selection due to the high number of variables.

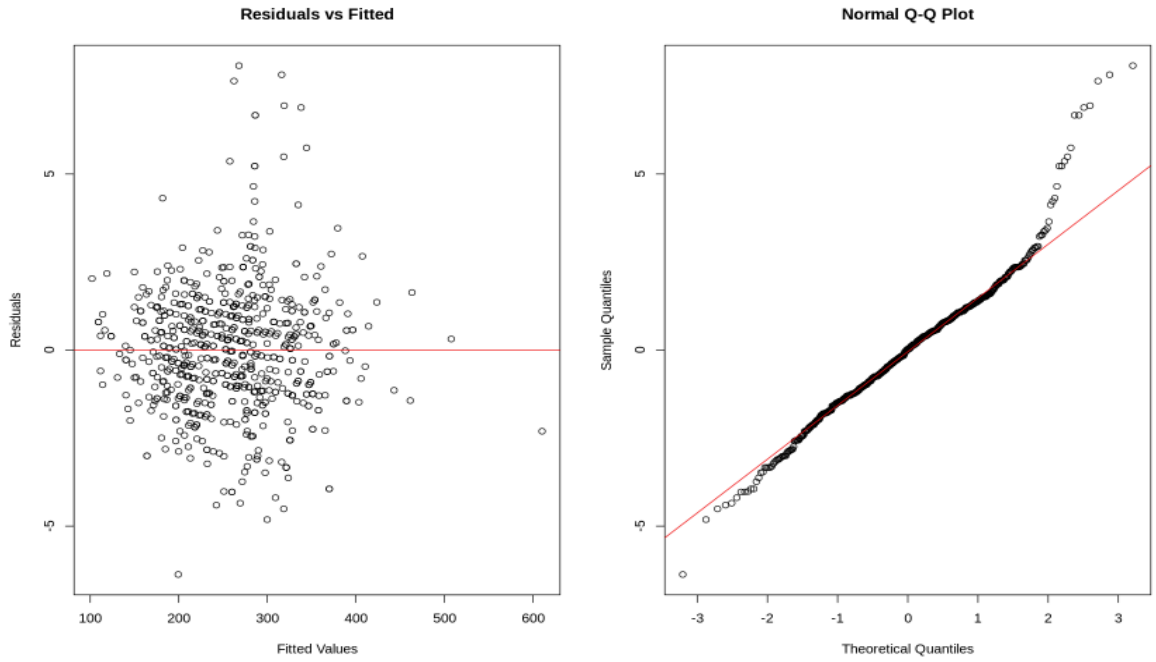

Figure 4: Residuals vs. Fitted Values for Model 10 & Normal Q-Q Plot for Model 10

We referred to Mallow's Cp and BIC in the model selection process. After doing backward selection, we reduced the number of parameters to 10. It has a Mallow's Cp value of 9.119315 which is the closest to the number of parameters p, compared to all other models of different sizes. This suggests that it has the best balance between overfitting and underfitting. It also has a very high adjusted R squared value (0.9992875), and relative to other models the BIC value is one of the lowest (-5292.763). Lower BIC values indicate superior model performance, aligning with the Bayesian principle that simpler models explaining the data are more probable.

We then decide to keep Fuel_Type , Combined_L_100km , Combined_MPG , Cylinders_10 , Cylinders_12 , Transmission , Vehicle_Class_StationSmall , Vehicle_Class_Minicompact as our explanatory variables.

After fitting model 10, we looked at the residual plot for the fitted values and the normal QQ plot. The residual plot does not show a clear systematic pattern, but the spread of **residuals appears to increase slightly** with the fitted values, indicating **potential heteroscedasticity**. We might consider using heteroscedasticity-consistent standard errors or transforming the data.

The Q-Q plot shows some deviation from the line at both ends of the distribution, suggesting that the **residuals may not follow a normal distribution**. This could be due to **outliers, heavy tails, or skewness in the data**. To account for this, we identified influential points and outliers by making use of Cook's distance to identify them.

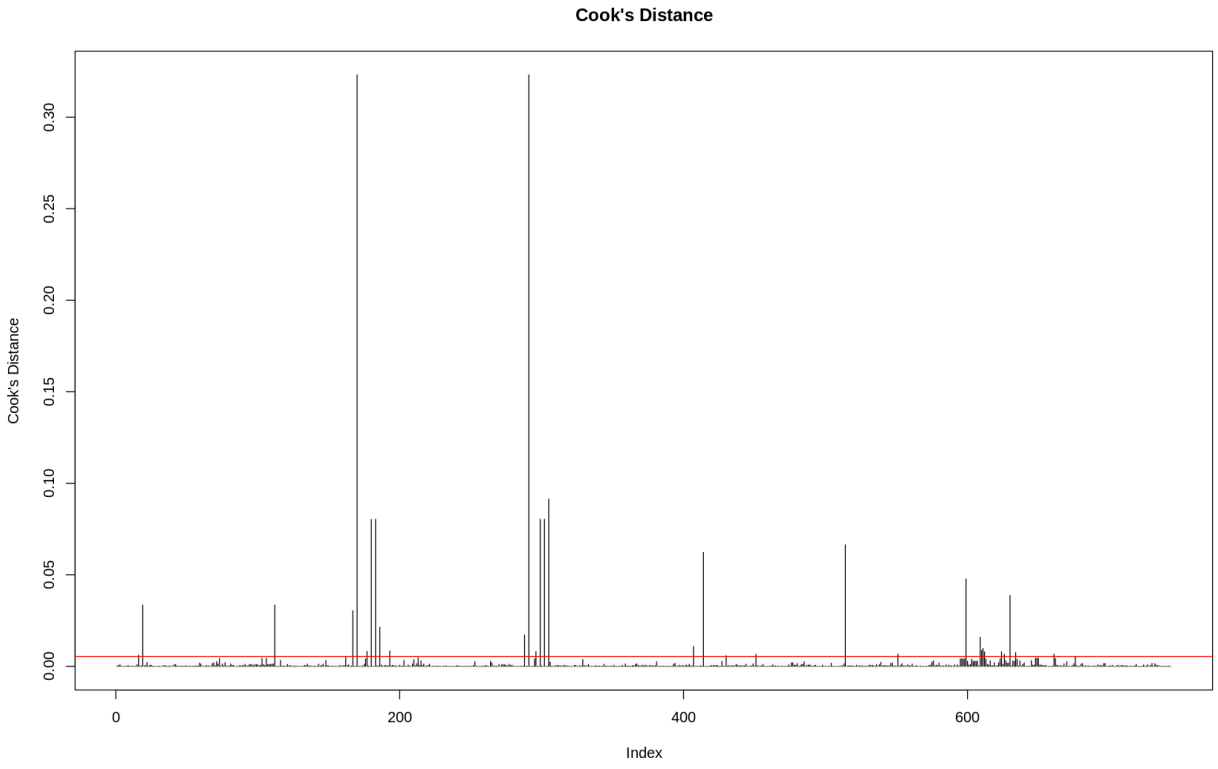

Figure 5: Cook's Distance for Model 10

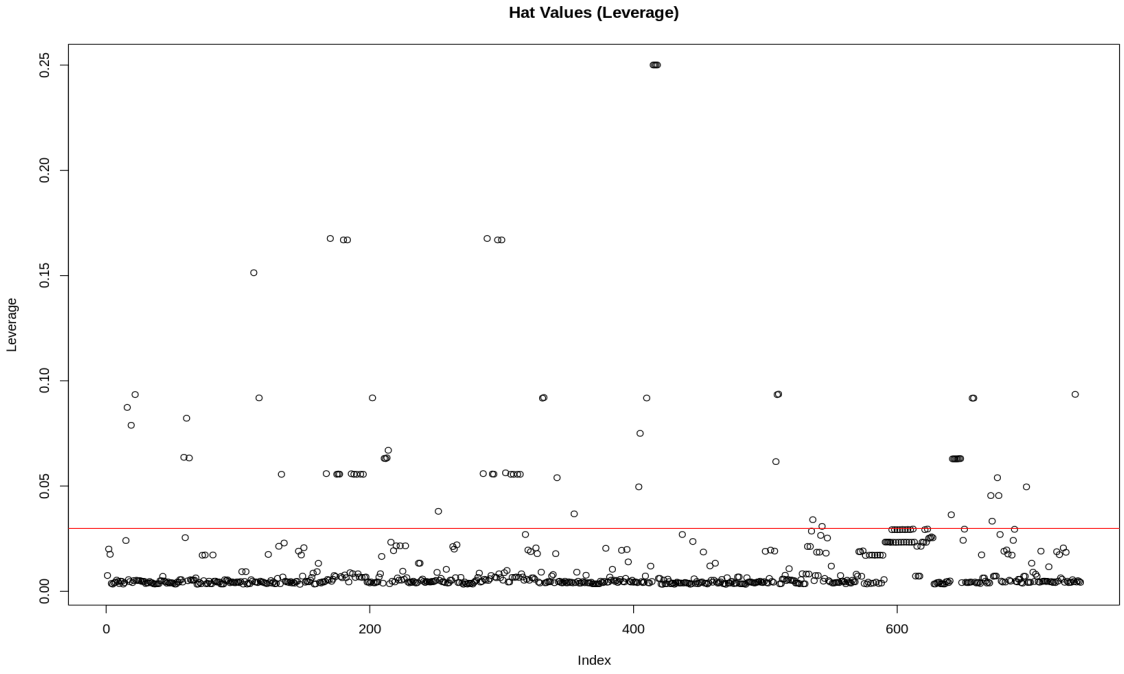

Figure 6: Leverage Values for Model 10

We plotted Cook's distance and leverage values for the fitted model 10 to identify potential outliers. Cook's distance measures how much each observation influences the model's predictions. A high Cook's distance suggests that the observation has a strong influence and may be an outlier. Leverage is a measure of how far an observation's values of the independent variables are from the mean of those variables. The red line indicates the threshold value for indicating whether a point has high Cook's distance or leverage. We then removed all the points with high Cook's distance and leverage since they are most likely to be outliers.

```
Call:
lm(formula = CO2_Emissions_g_km ~ Fuel_Type + Combined_L_100km +
    Combined_MPG + Cylinders_10 + Cylinders_12 + Transmission +
    Vehicle_Class_StationSmall + Vehicle_Class_Minicompact, data = cleaned_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4659 -0.9435  0.0434  0.9162  3.4386

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     37.41017    1.18909  31.461  < 2e-16 ***
Fuel_TypeX                     -34.74172    0.40819 -85.113  < 2e-16 ***
Fuel_TypeZ                     -35.16670    0.40792 -86.210  < 2e-16 ***
Combined_L_100km                23.33183    0.05751 405.706  < 2e-16 ***
Combined_MPG                    -0.04937    0.01995  -2.475  0.01356 *
Cylinders_10TRUE                -2.16400    0.67762  -3.194  0.00147 **
Cylinders_12TRUE                -1.92867    0.39494  -4.883  1.3e-06 ***
TransmissionM                   -0.54665    0.17711  -3.086  0.00211 **
Vehicle_Class_StationSmallTRUE  -0.99180    0.47277  -2.098  0.03629 *
Vehicle_Class_MinicompactTRUE    0.50331    0.23400   2.151  0.03184 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.322 on 679 degrees of freedom
Multiple R-squared:  0.9996,    Adjusted R-squared:  0.9996
F-statistic: 1.817e+05 on 9 and 679 DF,  p-value: < 2.2e-16
```

Figure 7: Summary of Final Model fitted

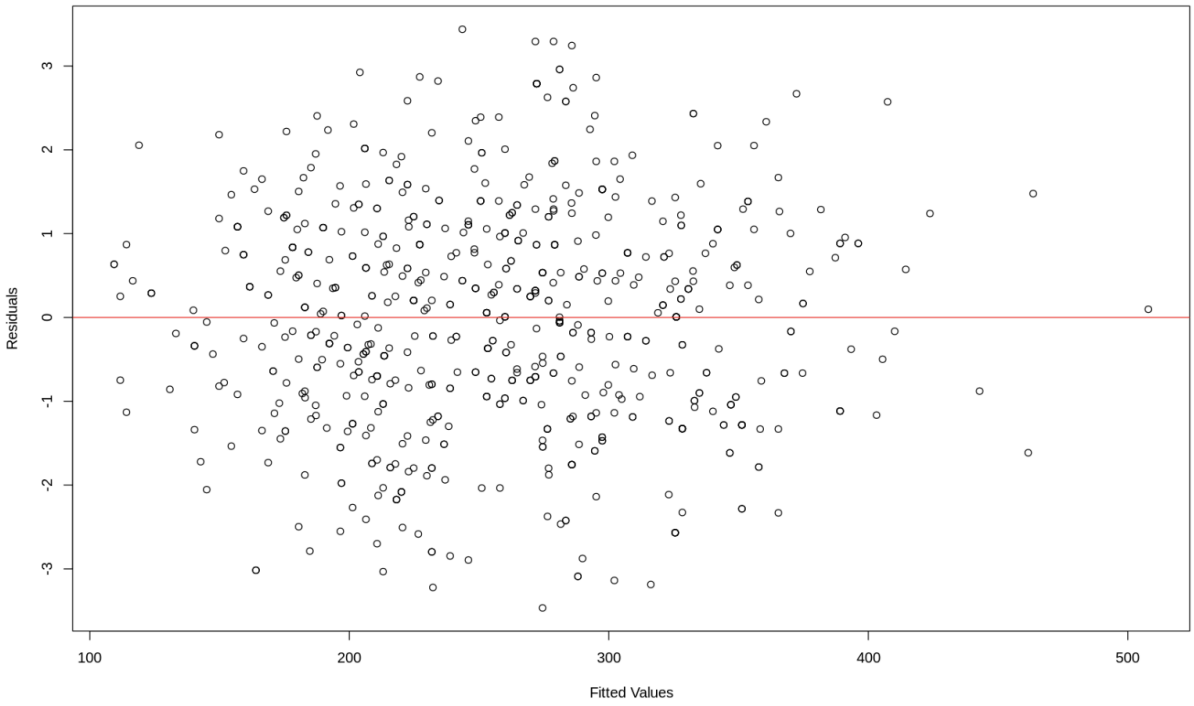

Figure 8: Residuals vs. Fitted Values for Final Model

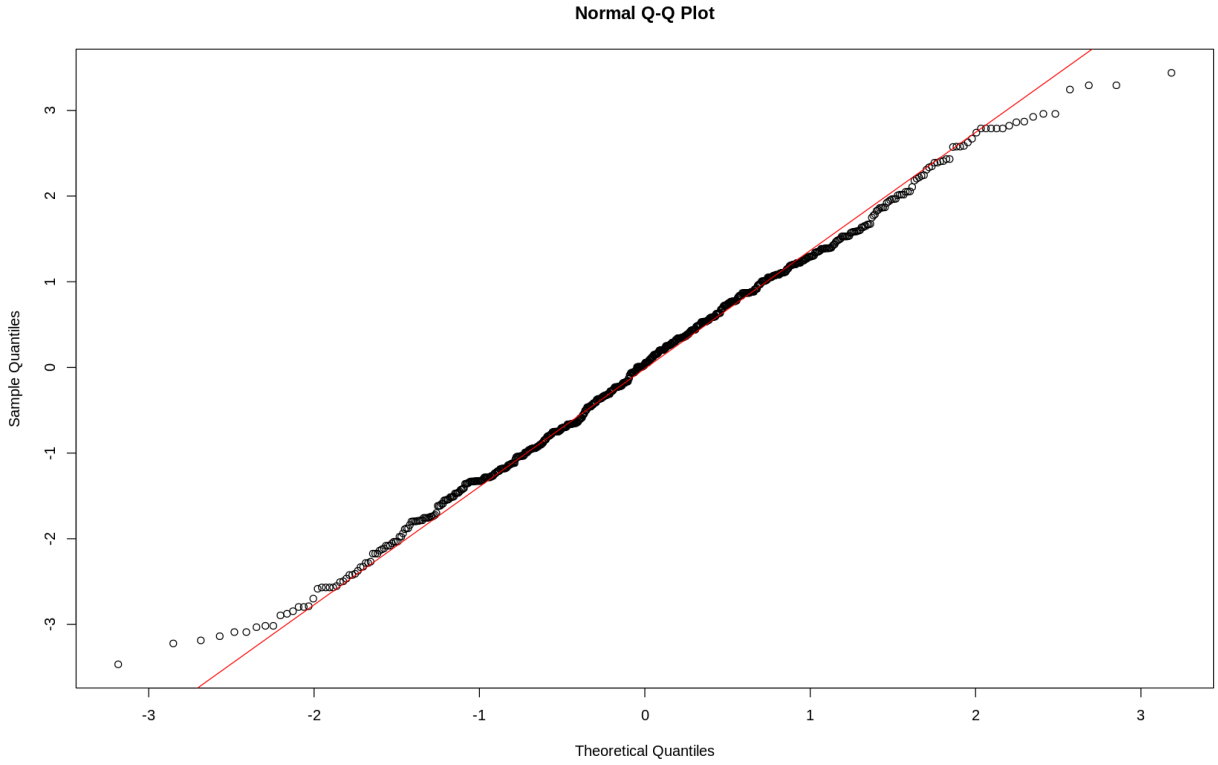

Figure 9: Normal Q-Q Plot for Final Model

After removing potential outliers, we refitted model 10 to the remaining data, and generated residual plots and Normal Q-Q Plot to observe model performance. We see that the points are scattered randomly in the residual plot, showing no obvious pattern, which indicates that the assumption of homoscedasticity is fulfilled. However from the Normal Q-Q plot we see that the data has a slightly heavier tail than normal distribution.

We also tried to apply transformations (log, exponential, sqrt, squared) for better model performance. However, all transformations make the residual plot show a clear pattern and it does not make the distribution of residuals any closer to normally distributed than it already is. The residual plot without transformation shows no patterns (suggesting homoscedasticity and linearity) and the only concern is slight non-normality in residuals.

Conclusion

The pollutants released from vehicles contain greenhouse gasses, which are known to mask the Earth and trap the sun's heat, and produce a positive climate forcing. This causes the world to warm up at a rapid rate, which over time is one of the leading threats to human health due to extreme weather patterns and the disruption of the usual balance of nature (United Nations, n.d.).

Light-tailed residuals could indicate that the model may not fully capture the variability in the data, especially with regard to rare or extreme events. However, this is not such a big concern since our research question's primary focus is on average behavior rather than the tails of the distribution. However, it does mean that the model will be less reliable when predicting extreme values. However, since we have a large sample size, the Central Limit Theorem assures that the estimation of coefficients will still be approximately normally distributed, allowing for valid inference. Therefore, non-normality of residuals may not be a significant problem.

For inference, light-tailed residuals can indicate that the model is overestimating the certainty of your predictions, which might lead to overly confident conclusions about the effects of the predictors.

The coefficients from the final model, suggest that Regular Gas, and Premium Gas has a significant negative relationship with CO2 emissions. In contrast, the combined fuel consumption in L/100km has a large positive association with CO2 emissions. The lack of interaction suggests that each variable independently affects CO2 emissions, and does not require their covariates.

By decreasing the distance that we drive and being mindful in selecting a vehicle with fuel efficient specifications to meet our needs, we can reduce our individual carbon footprint to do our part in mitigating the rapidity of global warming.

References

Natural Resources Canada. Auto$Mart - Learn the Facts: Fuel Consumption and Co, natural-resources.canada.ca/sites/www.nrcan.gc.ca/files/oee/pdf/transportation/fuel-efficient-technologies/autosmart_factsheet_6_e.pdf

Natural Resource Canada. "Fuel Consumption Ratings." Open Government Portal, Government of Canada, 11 July 2023, open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64.

Natural Resources Canada. "Links between Fuel Consumption, Climate Change, Our Environment and Health." Government of Canada, 20 Sept. 2016, natural-resources.canada.ca/energy/efficiency/communities-infrastructure/transportation/idling/4419.

Statistics Canada. "Canadians' Commutes: Still Car-Heavy, but Some Lighter Footprints." Government of Canada, 9 June 2023, www.statcan.gc.ca/o1/en/plus/3798-canadians-commutes-still-car-heavy-some-lighter-footprints

United Nations. (n.d.). "Causes and effects of climate change." United Nations. https://www.un.org/en/climatechange/science/causes-effects-climate-change#:~:text=Climate%20change%20is%20the%20single,grow%20or%20find%20sufficient%20food.

Appendix A

Table 1A: Breakdown of Dataset Variables

Name	**Levels**
Model year	“2024”
Make	"Acura", "Alfa Romeo", "Aston Martin", "Audi", "Bentley", "BMW", "Bugatti", "Buick", "Cadillac", "Chevrolet", "Chrysler", "Dodge", "Ferrari", "Ford", "Genesis", "GMC", "Honda", "Hyundai", "Infiniti", "Jaguar", "Jeep", "Kia", "Lamborghini", "Land Rover", "Lexus", "Lincoln", "Maserati", "Mazda", "Mercedes-Benz", "MINI", "Mitsubishi", "Nissan", "Porsche", "Ram", "Rolls-Royce", "Subaru", "Toyota", "Volkswagen", "Volvo". 33 Unique Values
Model	“*” Represents the name of the model followed by 8 different model type as follows: “*AWD” = All-wheel drive – vehicle designed to operate with all wheels powered; “*4WD/4X4” = Four-wheel drive – vehicle designed to operate with either two wheels or four wheels powered; “*FFV” = Flexible-fuel vehicle – vehicle designed to operate on gasoline and ethanol blends of up to 85% ethanol (E85); “*CNG” = Compressed natural gas; “*NGV” = Natural gas vehicle; “*SWB” = Short wheelbase; “*LWB” = Long wheelbase; “*EWB” = Extended wheelbase. # = High output engine 590 unique values. (which could be classified into 8 model typed as described above with the *).
Vehicle Class	“Full-size”, “Sport utility vehicle: Small”, “Sport utility vehicle: Standard”, “Mid-size”, “Minicompact”, “Two-seater”, “Subcompact”, “Compact”, “Station wagon: Small”, “Station wagon: Mid-size”, “Pickup truck: Small", “Pickup truck: Standard", “Minivan”. 22 unique values.
Engine size (L)	N/A
Cylinders	“4”, “6”, “8”, “12”, “16”, “3”, “10”

Transmission	"*" Represents the gear count (in Integers) for each transmission. "A*" = Automatic; "AM*" = Automated manual; "AS*" = Automatic with select shift; "AV*" = Continuously variable; "M*" = Manual
Fuel Type	"X" = Regular gasoline; "Z" = Premium gasoline "D" = Diesel "" = E85E "N" = Natural Gas
City (L/100km)	N/A
Highway (L/100 km)	N/A
Combined (L/100 km)	N/A
Combined (mpg)	N/A
CO2 emissions (g/km)	N/A
CO2 rating	"1", "2", "3", '4", "5", "6", "7", "8"
Smog Rating	"1", "3", "5", "6", "7", "8"

Appendix B

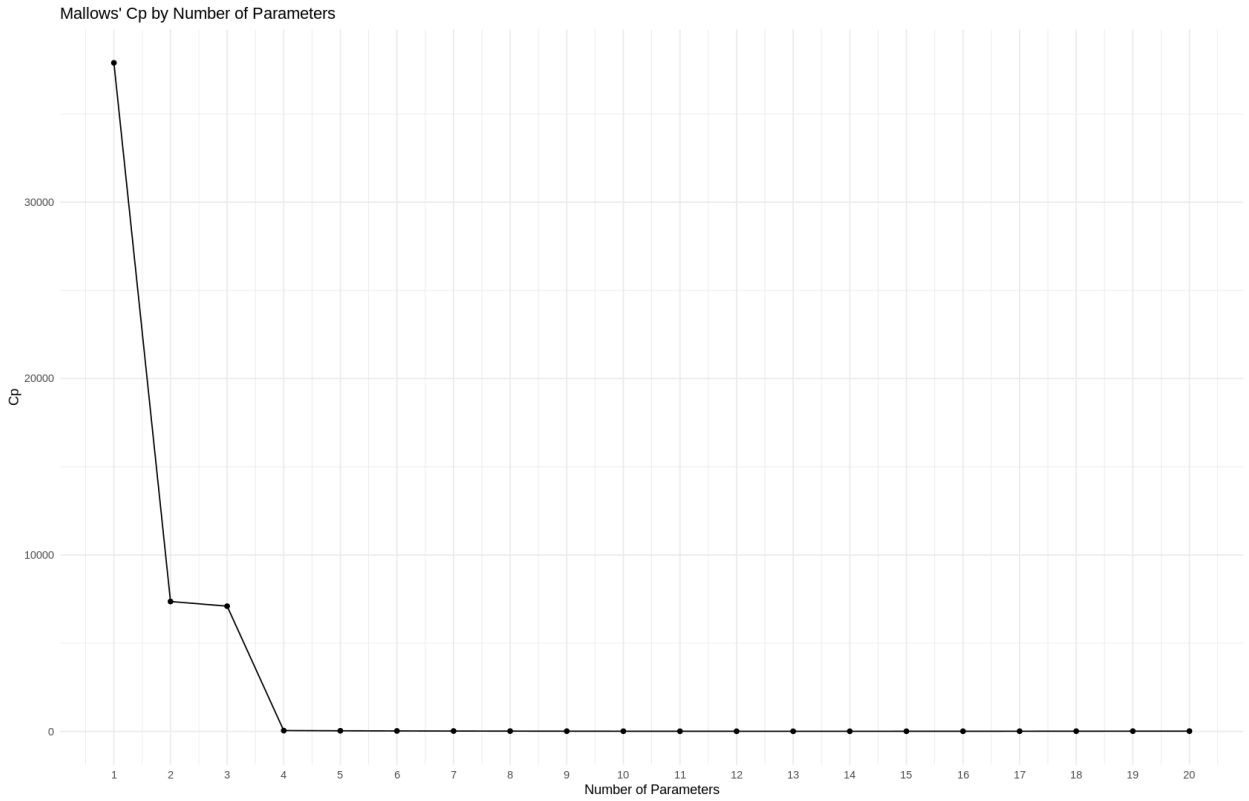

Figure 1B: Mallow's CP by Number of Parameters

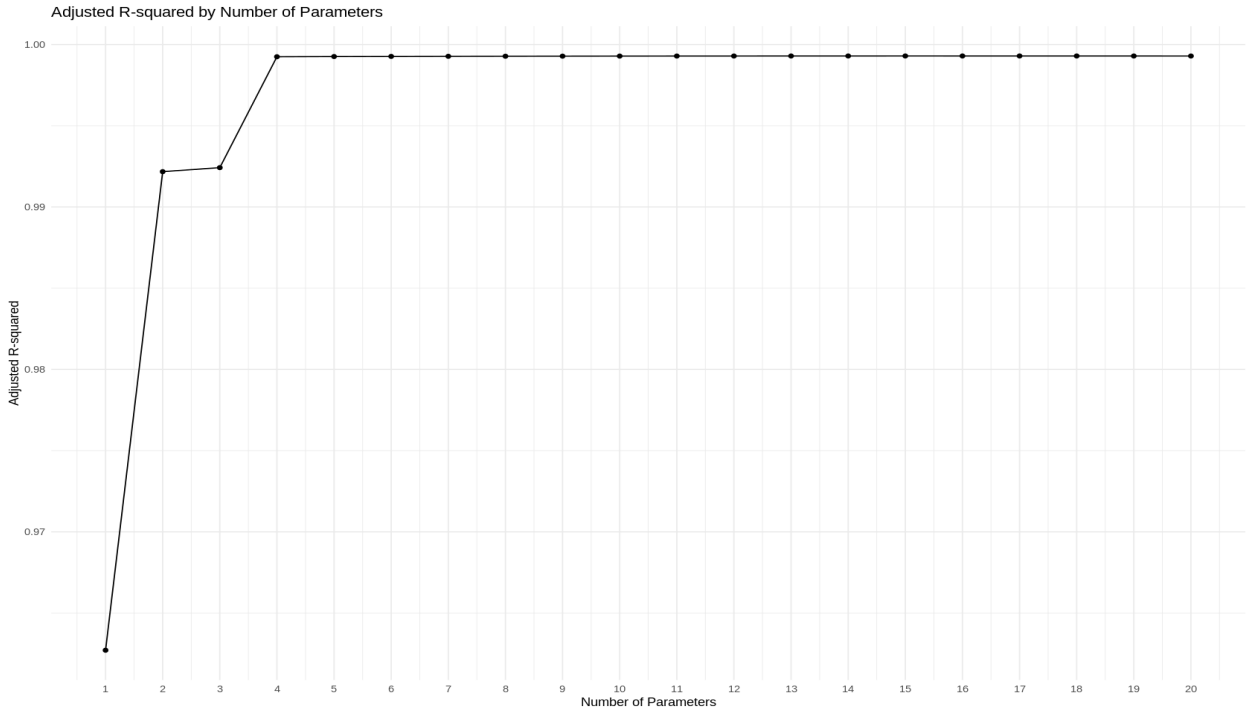

Figure 2B: Adjusted R-squared by Number of Parameters

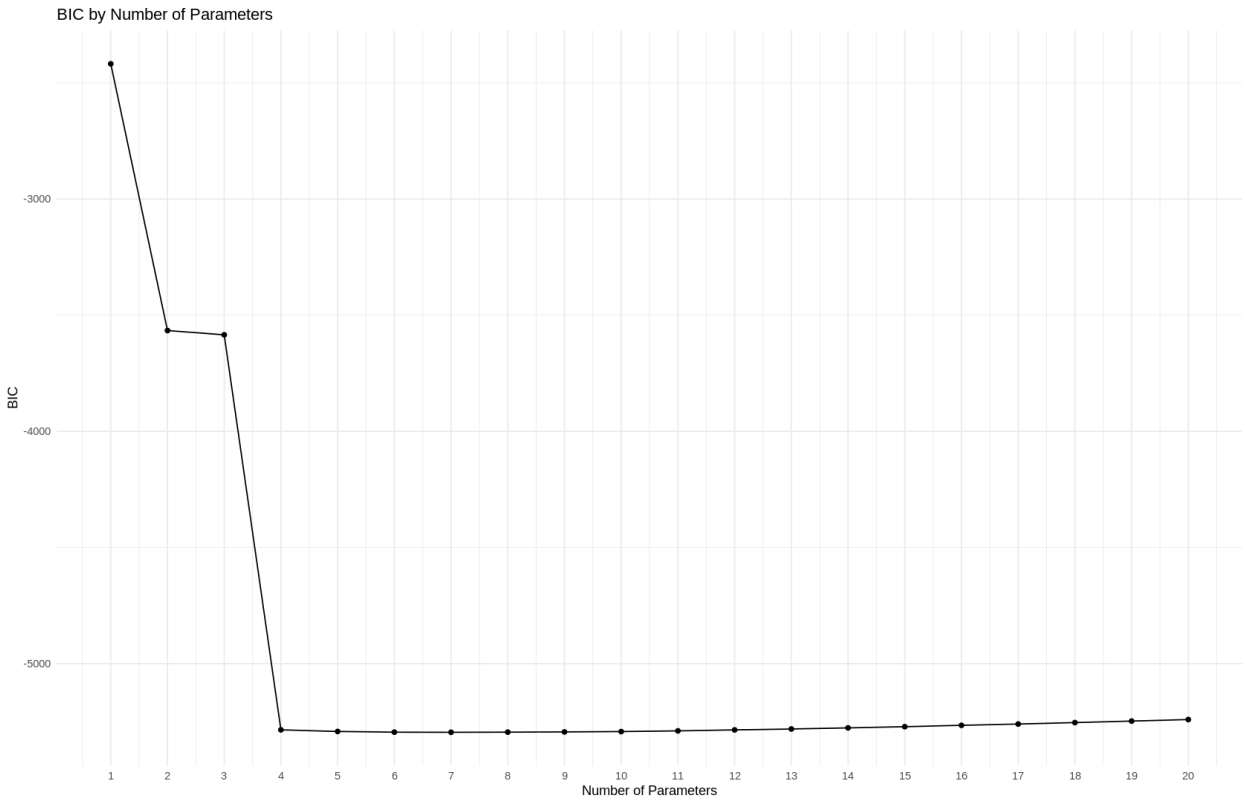

Figure 3B: BIC by Number of Parameters