Emission Insights: Analyzing CO2 Emissions in Vehicles Using R

AUTHOR
Jatin A Bomrasipeta and Sharath Reddy Muthyala

Abstract:

This report presents a detailed analysis of a comprehensive dataset on various vehicles, exploring the relationship between car features and environmental impact. It categorizes cars by brand, model, and type, such as compact or SUV, and delves into specifics like engine size and transmission. The study also examines fuel consumption in urban and highway settings, alongside CO2 emissions, to assess each vehicle's environmental footprint. Through statistical methods, patterns are discerned, revealing correlations between vehicle attributes and fuel efficiency. The findings are essential for manufacturers aiming to develop cars that align with consumer preferences and environmental standards, providing a foundation for future vehicle innovation and sustainable practices.

Introduction:

CO2 emissions have significant environmental impacts. The accumulation of CO2 in the atmosphere contributes to the greenhouse effect, trapping heat and causing global warming. This leads to various adverse effects, including rising global temperatures, sea-level rise, changes in weather patterns, and ecosystem disruptions. The long-term consequences of climate change can negatively impact human health, agriculture, biodiversity, and socio-economic systems.

Therefore, reducing CO2 emissions is crucial for mitigating climate change and minimizing its harmful effects on the environment and human well-being. This involves transitioning to cleaner and renewable energy sources, improving energy efficiency, adopting sustainable practices, and promoting conservation efforts.

Building on the analysis of the dataset provided, the development of a regression model to predict CO2 emissions is a critical step toward quantifying the environmental impact of vehicles. By identifying key features such as engine size, vehicle class, and fuel type that significantly influence emissions, the model can offer valuable predictions of CO2 output. The insights gained from the model's coefficients can guide manufacturers in designing vehicles that are both efficient and environmentally friendly. Furthermore, these findings can inform policy recommendations for regulating vehicle emissions. For instance, incentives for vehicles with smaller engines or alternative fuel systems could be a strategic approach to encourage the production and purchase of lower-emission vehicles. Such policies would not only promote innovation in vehicle design but also align with broader environmental objectives aimed at reducing the carbon footprint of the transportation sector.

Data Description:

Below is the description of the data:

Make: The brand of the vehicle (e.g., ACURA).

Model: Specific model of the vehicle (e.g., ILX).

Vehicle.Class : The category of the vehicle (e.g., COMPACT, SUV - SMALL).

Engine.Size.L.: The size of the engine in liters (continuous numeric).

Cylinders: The number of cylinders in the engine (integer).

Transmission: Type of transmission (e.g., M6, AV7).

Fuel Type: Type of fuel used (e.g., Z).

Fuel Consumption City (L/100 km): Fuel consumption in city (continuous numeric).

Fuel Consumption Hwy (L/100 km): Fuel consumption on the highway (continuous numeric).

Fuel Consumption Comb (L/100 km): Combined fuel consumption (continuous numeric).

Fuel Consumption Comb (mpg): Combined fuel consumption in miles per gallon (integer).

CO2 Emissions(g/km): Carbon dioxide emissions in grams per kilometer (integer).

Goal:

The project is focused on creating a sophisticated predictive model that accurately forecasts CO2 emissions from vehicles, utilizing a rich dataset encompassing various vehicle characteristics such as make, model, engine size, number of cylinders, type of transmission, and fuel consumption metrics. This model aims to quantify the environmental impact of different types of vehicles, which is crucial in the current context of growing environmental concerns. By analyzing and understanding the correlations between these variables and CO2 emissions, the project endeavors to provide valuable insights into how automobile design and consumer choices can be optimized for lower carbon footprints.

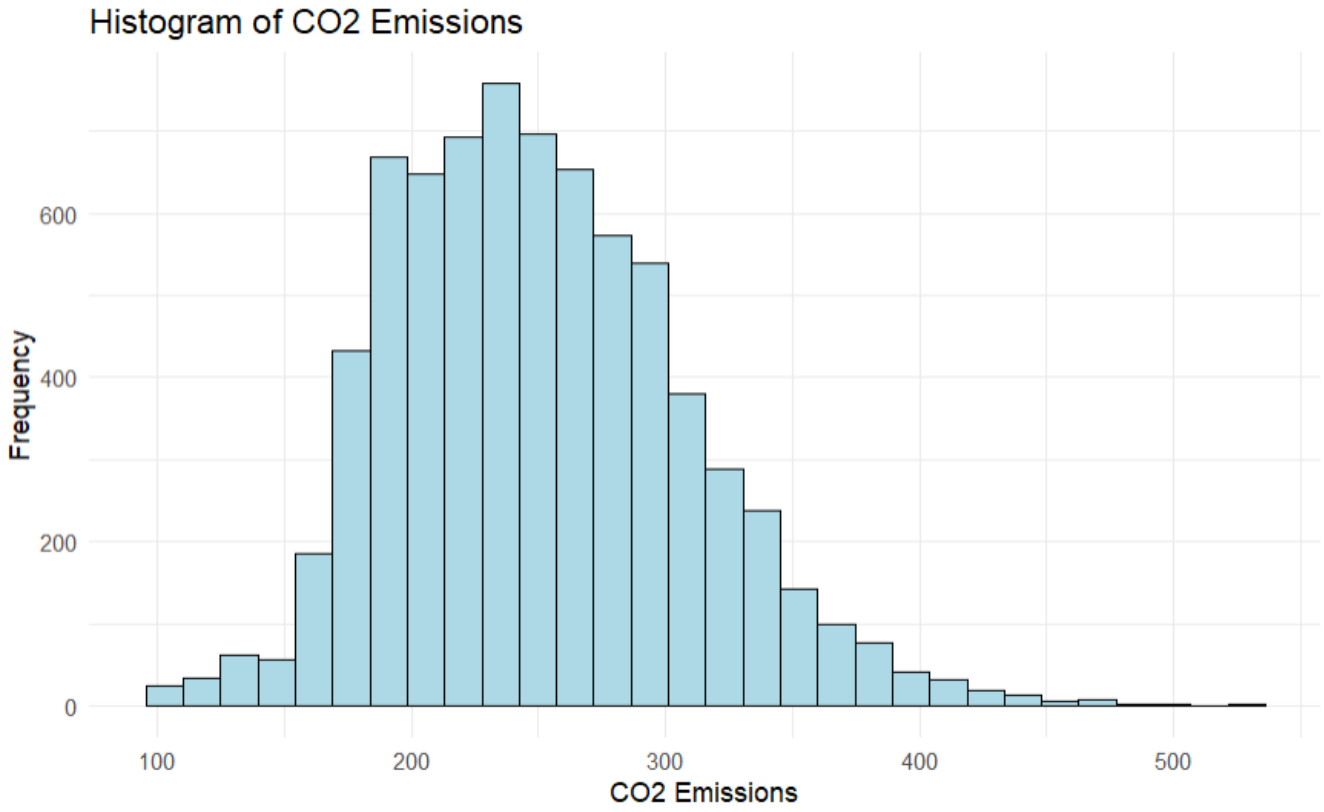
Data Analysis:

Univariate exploratory data analysis:

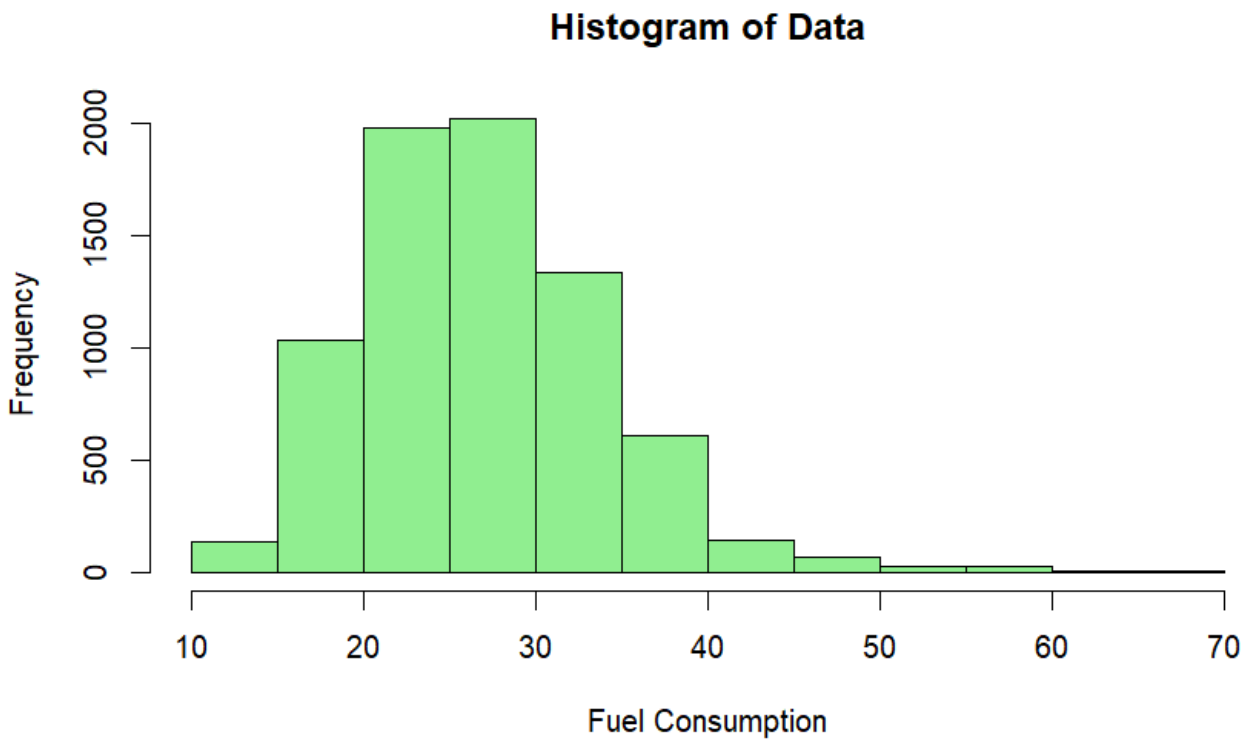
We will now perform univariate EDA to see the distribution of the data. We will check the distribution of continuous variables using boxplot and histogram. For categorical variables, we will use the table function and bar plot to check the frequency/counts..

We will also have to check if there are any missing values present in the data set. After checking the missing values, we found that there are No missing values in the data.

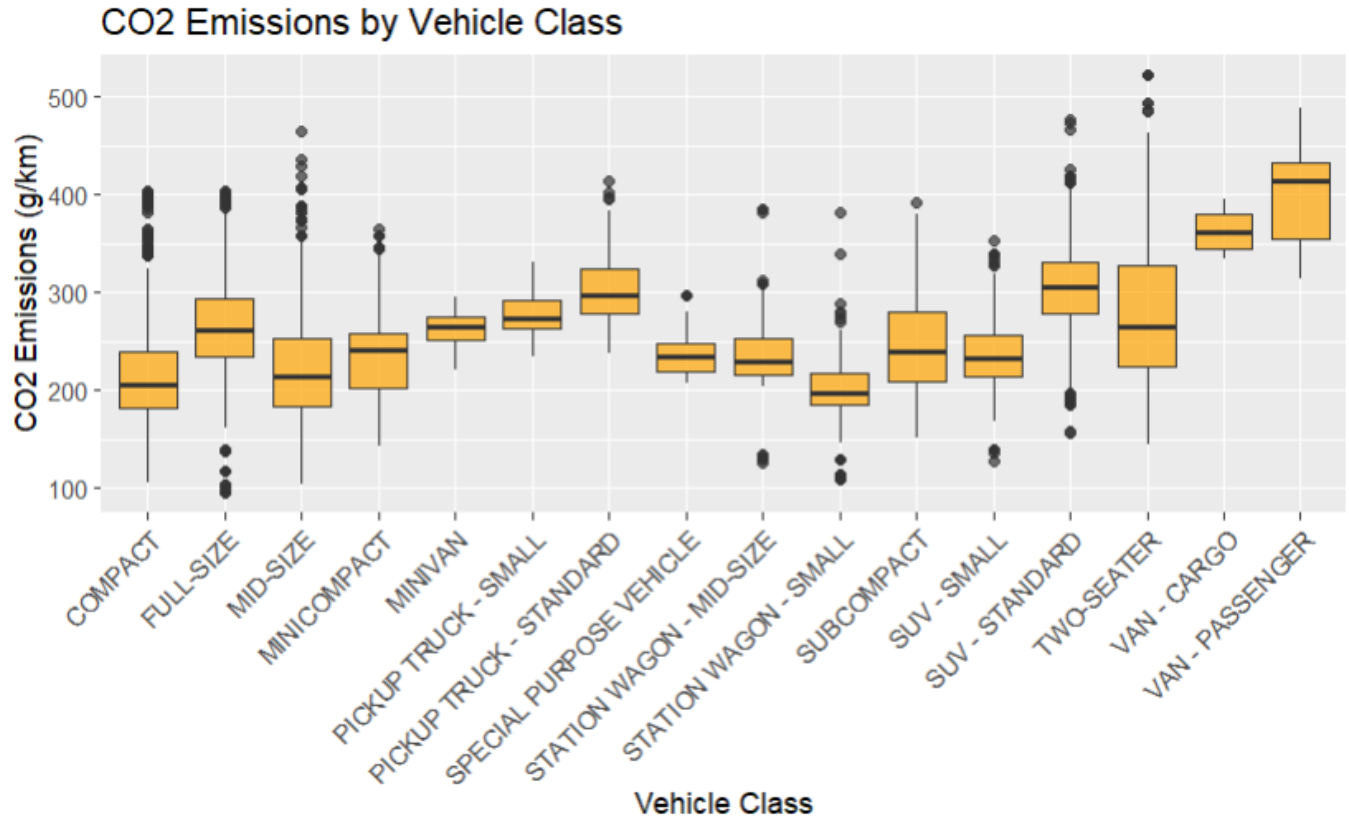
Let’s move to the plots of the variable:



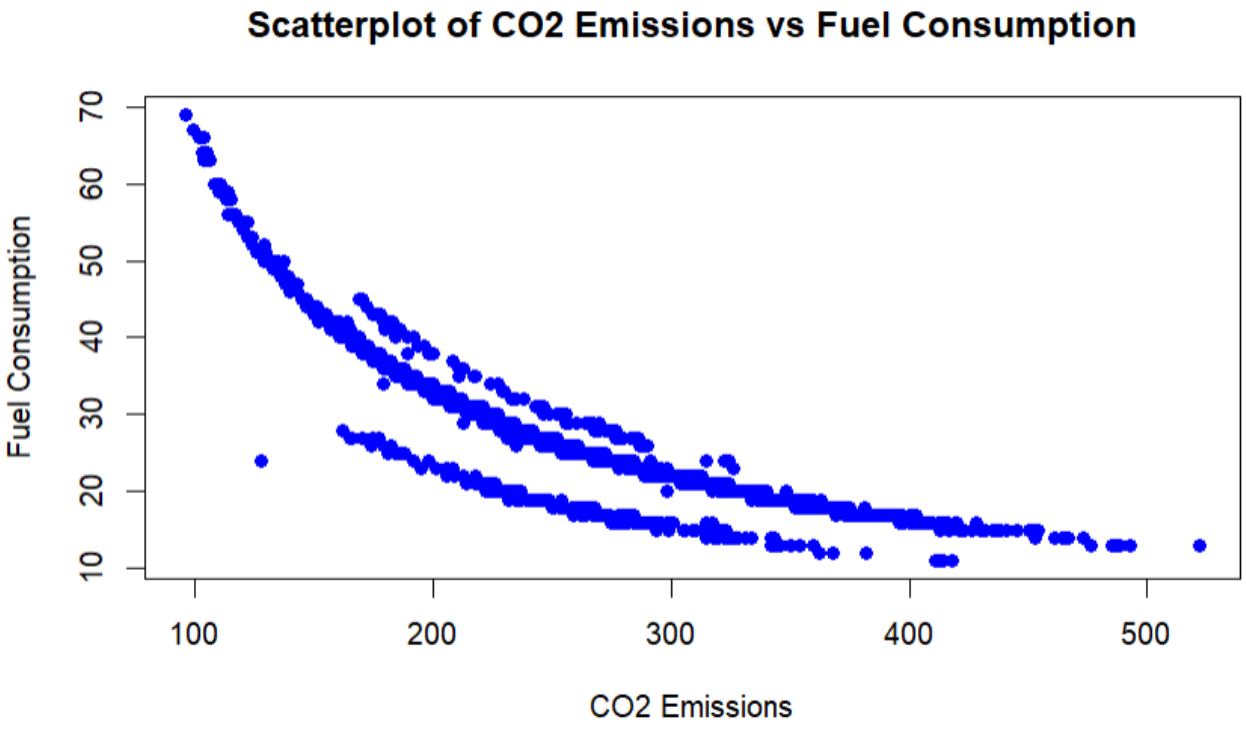
The horizontal axis (x-axis) shows the CO2 emissions, and the vertical axis (y-axis) indicates the frequency of observations. The bars represent different ranges of CO2 emissions, and the height of each bar shows how many vehicles fall within that range. The majority of vehicles emit CO2 within a middle range, as indicated by the tallest bars in the center of the histogram. There are fewer vehicles with very low or very high CO2 emissions, as shown by the shorter bars on the left and right sides of the histogram. This type of distribution is common and suggests that while most cars have average emissions, there are some that are particularly efficient or inefficient in terms of CO2 output.



The histogram represents the distribution of combined fuel consumption across the dataset's vehicle models, in miles per gallon (mpg). The majority of vehicles cluster in the lower consumption range, indicating a concentration of more fuel-efficient vehicles. The `table` function output shows the frequency of different vehicle classes in the dataset. This includes a variety of classes such as compact, SUVs, and pickups.



The boxplot analysis reveals a clear differentiation in CO2 emissions among various vehicle classes. It was observed that larger vehicle categories such as SUVs and pickup trucks are associated with higher median CO2 emissions when compared to smaller vehicles like compacts and mid-sized cars. The variability within each vehicle class, as demonstrated by the interquartile ranges and whiskers, shows considerable diversity, particularly in the standard pickup truck category, which exhibits a wide emission spectrum. Outliers are present across numerous vehicle classes, notably within the compact and small SUV segments, indicating specific models that significantly deviate from the general emission trends of their class. Additionally, the skewness in the distribution of emissions for certain classes suggests a tendency for these vehicles to emit at higher levels than the median.



The scatterplot depicts a negative correlation between CO2 emissions and fuel efficiency (in mpg). As CO2 emissions increase, fuel efficiency decreases, shown by the concentration of points forming a downward curve. This relationship is consistent with expectations: vehicles with higher fuel consumption tend to emit more CO2. The plot's pattern suggests a potential non-linear relationship, indicating that as vehicles become less efficient, the rate at which emissions increase may also change.

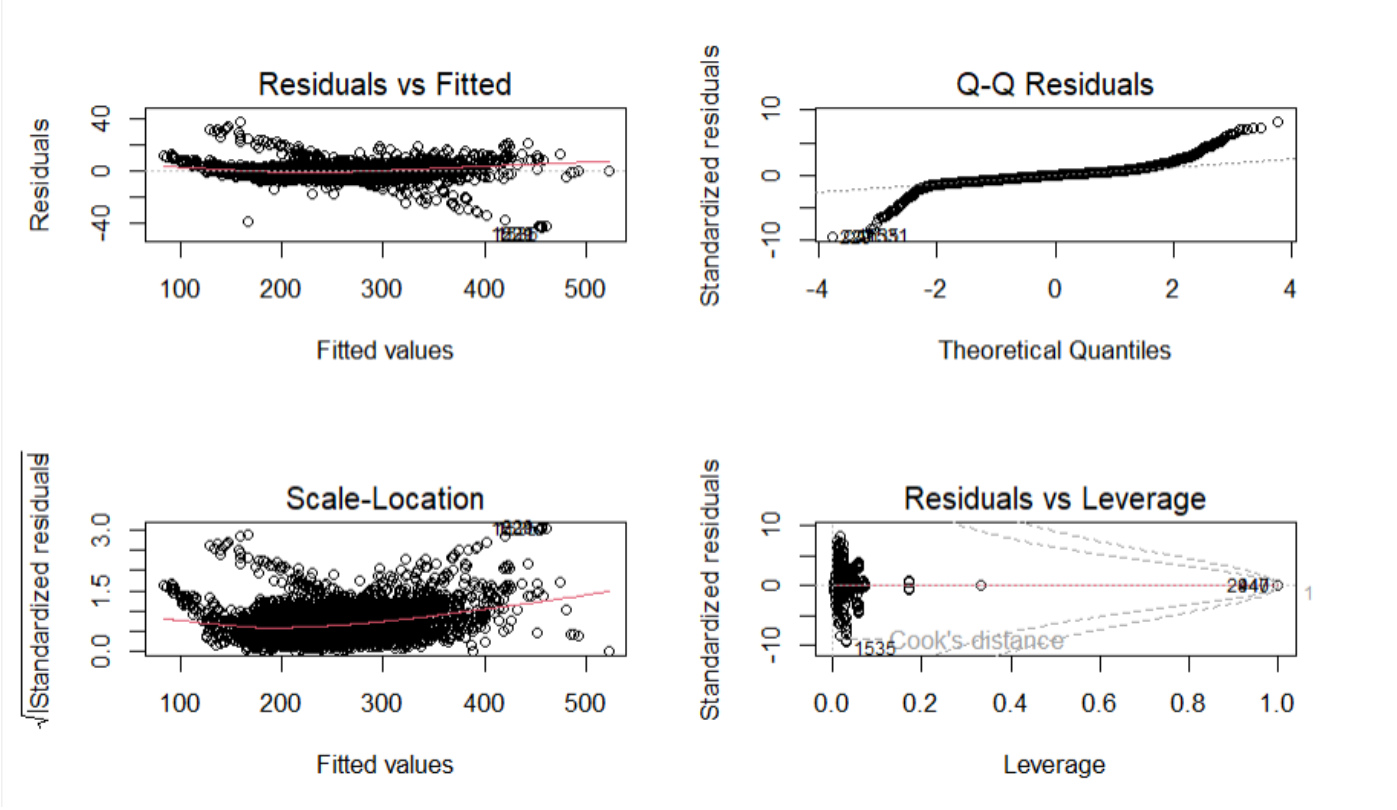
Null Model

The output is from a simple linear regression model predicting CO2 emissions for cars using only an intercept, meaning it's calculating the average CO2 emissions. The average emissions are about 250.5 grams per kilometer, which is significantly different from zero. The residuals show how much the actual emissions vary from this average, with a typical variation of about 58.7 grams per kilometer. The results are very reliable (with a very low p-value), meaning we can be confident that the average is a good representation of the dataset's emissions.

Raw Model

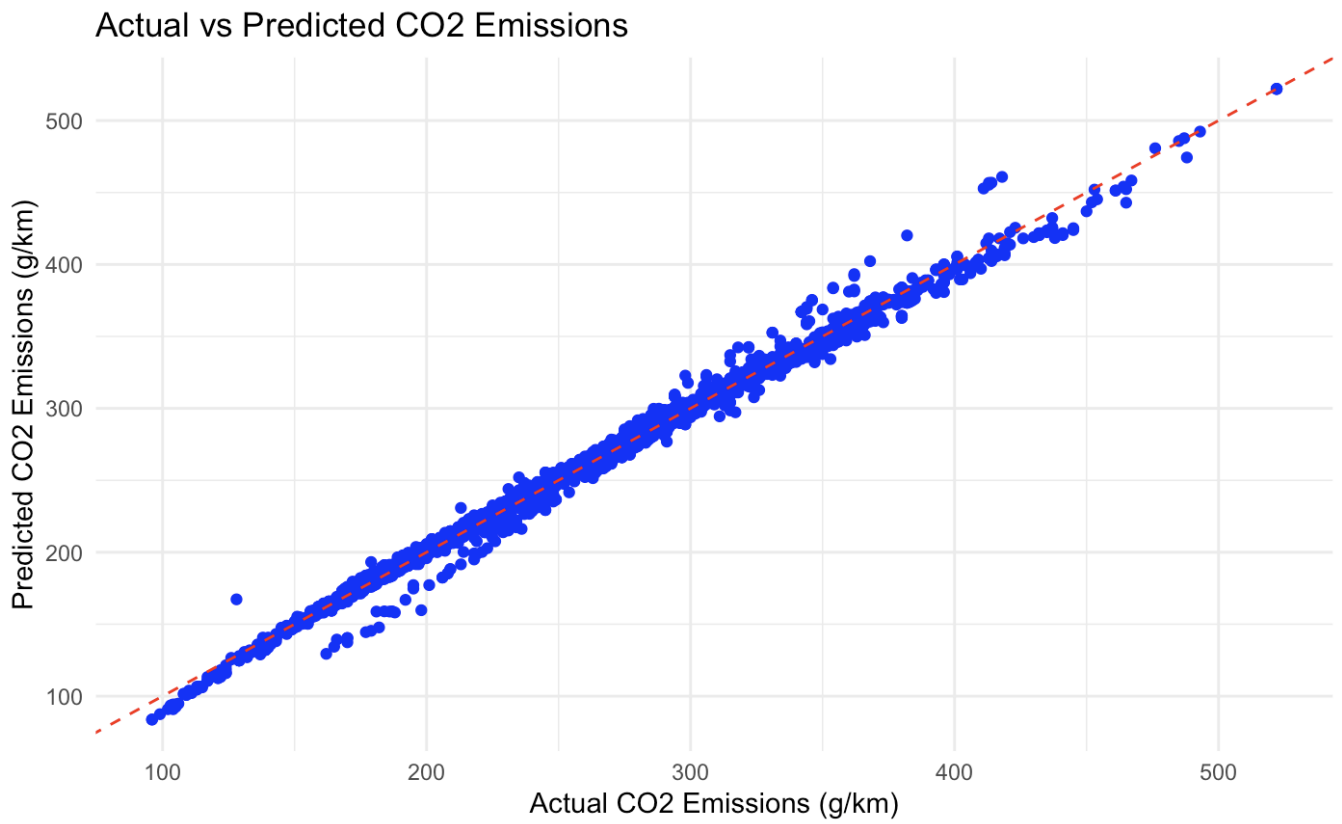
The intercept is about 85.72, which could be seen as the base CO2 emission level when all other factors are zero. Each car make has a different impact on CO2 emissions. For example, owning a BUGATTI adds about 23.97 units to the base emission level, while a SMART car reduces it by about 2.33 units. Other car features like engine size and cylinders also increase CO2 emissions. The model is very good at predicting CO2 emissions, with an R-squared value of 0.9938. This means the model explains almost all the variability of the emissions with the variables included. The residual standard error is about 4.662, meaning the typical prediction error is around 4.662 units of CO2 emissions. The p-values for most variables are very small, showing they are important in predicting CO2 emissions. The model uses data from 5816 cars, after accounting for the variables used. This model is a detailed way to understand how different factors like car brand, engine size, and fuel type impact CO2 emissions. The high R-squared value shows that these factors are very effective at predicting emissions.

Diagnostics for Raw Model:



The diagnostic plots suggest the regression model may have some issues. The residuals versus fitted values indicate potential non-linearity in the data. The Q-Q plot shows deviations from normality, especially in the tails. The scale-location plot indicates that residuals' variances are not consistent across the range of fitted values, suggesting heteroscedasticity. Lastly, the residuals versus leverage plot identifies a few potentially influential outliers. These points could disproportionately affect the model's predictions and may need further investigation.

Seeing how well data fits the regression line:



The scatterplot depicts actual versus predicted CO2 emissions of vehicles. The close alignment of data points along the dashed line indicates a strong correlation between the model's predictions and the actual data. Most points lie near the identity line (where predicted values equal actual values), suggesting the model has good predictive accuracy for CO2 emissions. However, some deviation from the line, especially at higher emission levels, suggests the model's accuracy diminishes somewhat as emissions increase.

Checking for multicollinearity:

- The Fuel Consumption in City has an extremely high VIF, which indicates that this predictor shares a lot of the variance with other predictor variables in the model.
- The Fuel Consumption on Highway also has a high VIF, suggesting it is not independent of the other predictors in the model.
- The Combined Fuel Consumption has the highest VIF of all, which is a strong indication that it is very highly correlated with one or more of the other variables in the model.

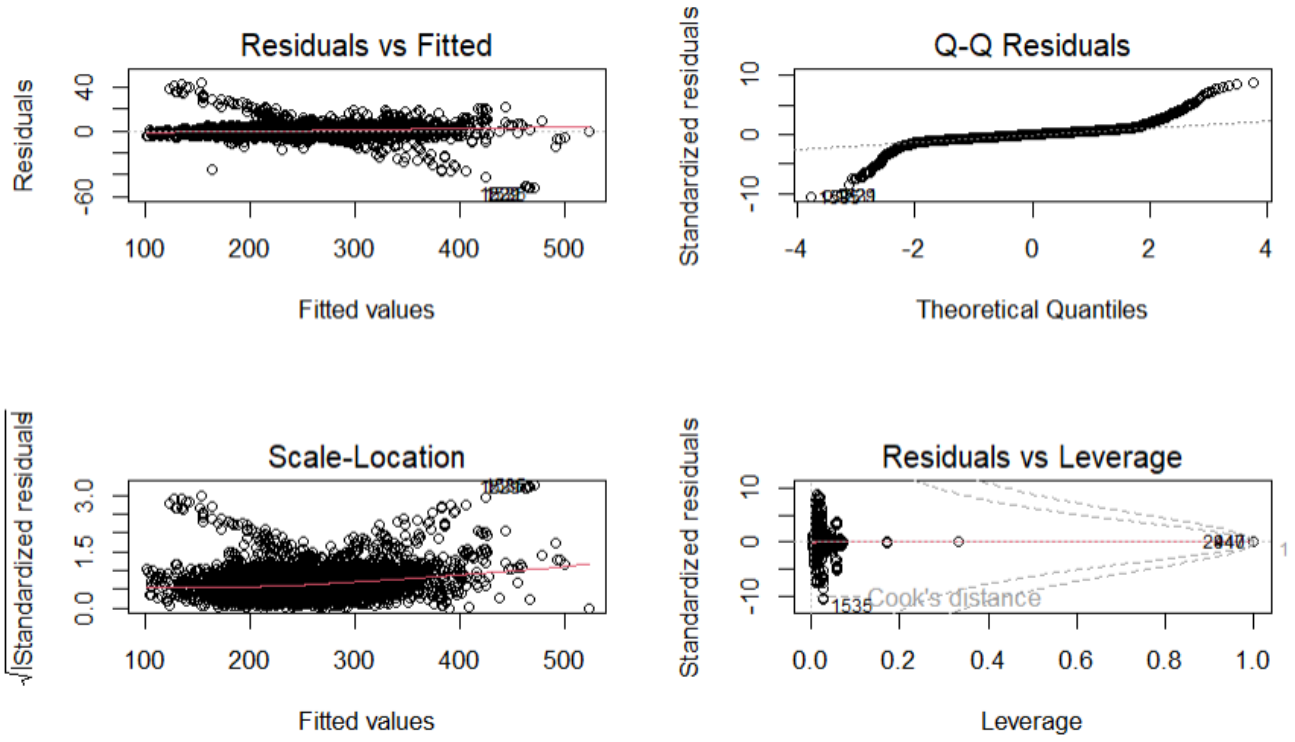
These high VIF values suggest that the variables are not just related to the response variable, but also to each other. This is not surprising since these three measures of fuel consumption are likely to be inherently related; for instance, city and highway fuel consumption will both influence combined fuel consumption.

Selecting one variable out of the 3. That has highest R^2 (Raw Model 1):

This model summary from a regression analysis is predicting CO2 emissions using car features like make, class, engine size, cylinders, transmission, fuel type, and combined fuel consumption. The base

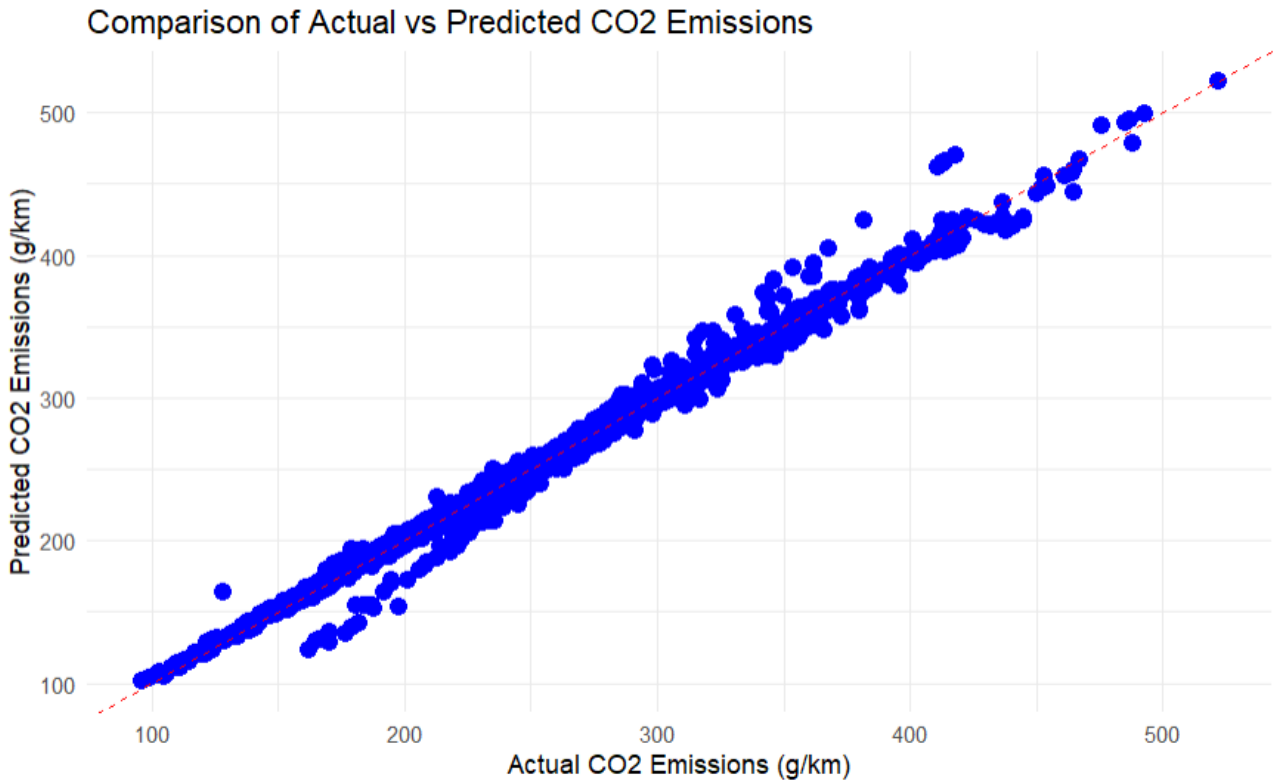
CO2 emissions (intercept) are about 35.4 grams per kilometer. Different makes and classes of vehicles affect CO2 emissions by varying amounts. For example, having a 'BUGATTI' adds approximately 12.2 grams per kilometer over the base level. The number of cylinders and combined fuel consumption are significant predictors of CO2 emissions. The model is excellent at explaining CO2 emission variations (with an R-squared value of 0.9928), indicating that these features account for nearly all the differences in emissions among the vehicles studied. The residual standard error is 5.009, showing that the model's predictions are generally close to the actual values. The coefficients for each variable show how much that factor changes the CO2 emissions when everything else is held constant. The model's F-statistic and associated p-value are practically zero, indicating the model is a good fit. This means the model can very accurately predict a car's CO2 emissions based on its characteristics.

Diagnostics of Raw Model 1:



The observations from these plots suggest that the assumptions of linear regression may not hold in this case. Specifically, there might be a problem with non-constant variance (heteroscedasticity), and the residuals do not seem to be normally distributed. Addressing these issues may require transforming the response variable or the predictors, adding missing predictors, or using a different kind of regression model that does not assume normality and constant variance of errors.

Seeing how well the data fits the regression line

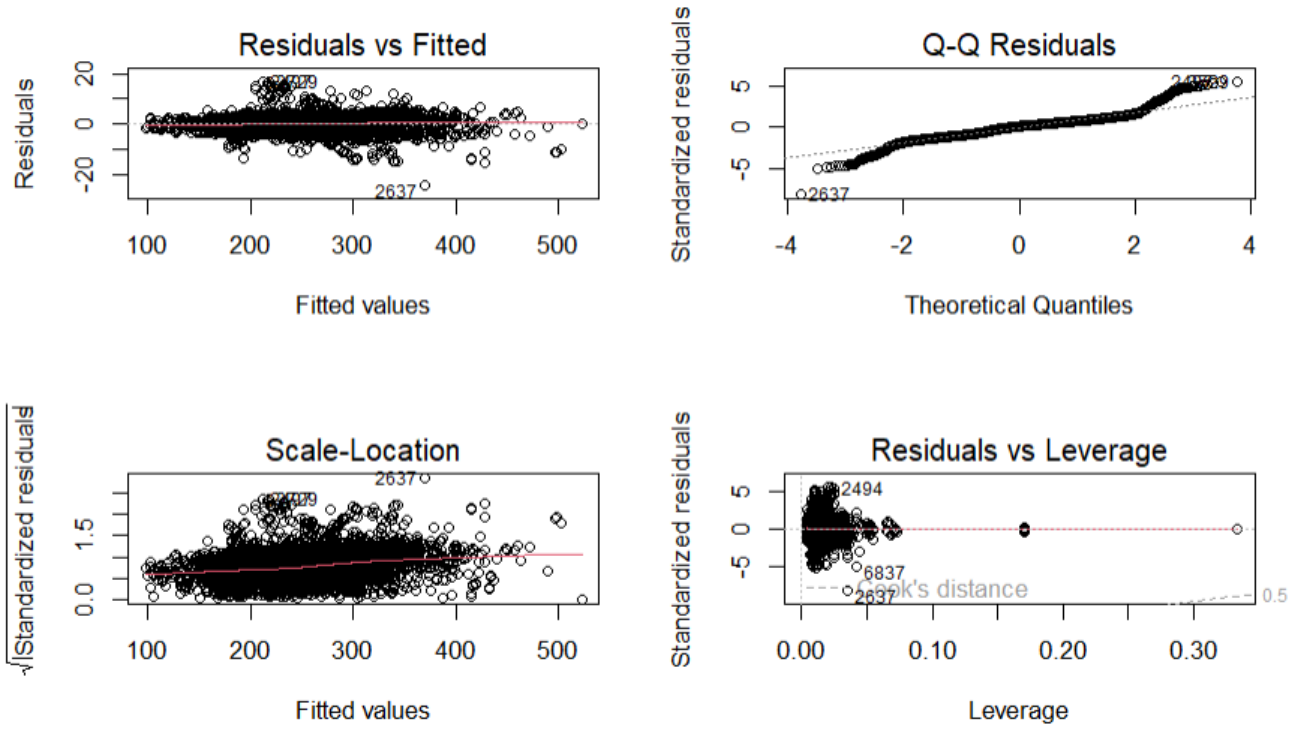


In this plot, if the blue dots are close to the red dashed line, it suggests that the model's predictions are accurate, and therefore, the independent variables chosen for the model are good predictors of CO2 emissions. However, there's a pattern to the deviations from the line (such as a funnel shape, where the variance increases with the value of the dependent variable), it might suggest that there's heteroscedasticity or that the model is not capturing all the relevant factors, especially at higher levels of emissions.

Clean Model 1 (After removing outliers):

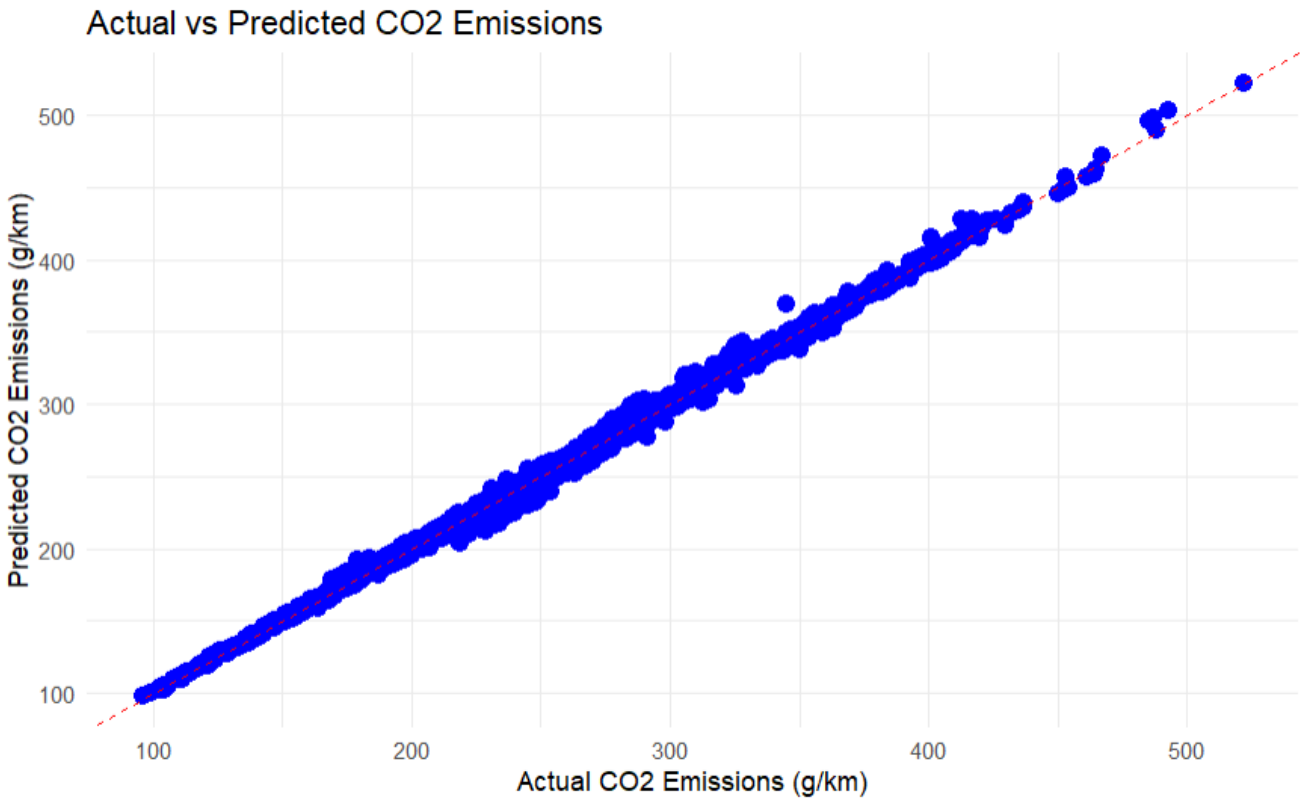
The regression analysis results show that CO2 emissions from cars can be predicted with a high degree of accuracy (R-squared of 0.9973) using the car's make, class, engine size, number of cylinders, transmission type, fuel type, and combined fuel consumption. The base level of emissions is around 32.14 grams per kilometer, and different car characteristics contribute various amounts to this base level. For example, the model suggests that cars with manual transmissions (A4, A5, A6) generally have lower CO2 emissions compared to the base model. Fuel consumption has a large positive effect on CO2 emissions, indicating that as fuel consumption increases, so do emissions. The very low p-value for the F-statistic confirms that the variables used in the model collectively have a significant effect on CO2 emissions.

Diagnostics of clean model 1:



Post-cleaning, the QQ plot of residuals more closely aligns with the theoretical quantiles, suggesting an improved adherence to normality. The Residuals vs. Fitted plot now exhibits a more random pattern, indicating better conformity with the assumptions of homoscedasticity and linearity. These enhancements suggest that the data cleaning efforts have effectively mitigated some of the prior model's deviations from the ideal regression assumptions, potentially leading to more reliable predictions from the regression analysis.

Seeing how well the data fits the regression line



The improved alignment of points around the line of perfect fit in the Actual vs Predicted values plot indicates a refined model post data cleaning. This tighter clustering suggests enhanced prediction accuracy, with the model now accounting for a greater proportion of the variance in CO2 emissions. Such convergence towards the line implies that the underlying relationship between independent variables and the dependent variable is captured more effectively, leading to increased confidence in the model's predictive capabilities and its potential applicability in real-world scenarios or further research.

Main Results:

Clean model 2 (After cleaning the data again)

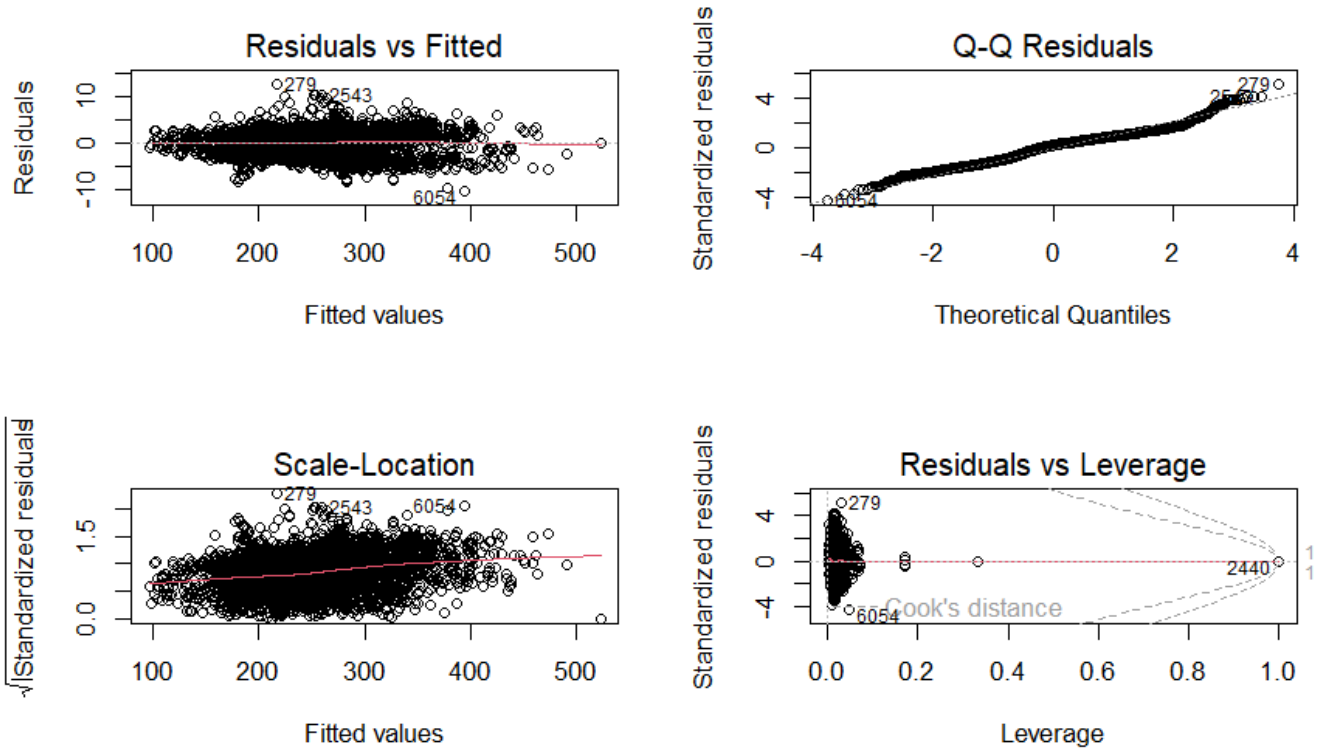
The regression model results show that we can very accurately predict a car's CO2 emissions based on certain features (Multiple R-squared of 0.9982). The intercept suggests that the starting point for CO2 emissions is around 30.88 grams per kilometer. Each make of car and its specific features contributes differently to CO2 emissions. For example, having a BUGATTI increases emissions by about 4.56 grams per kilometer compared to the base level.

Fuel consumption has a significant impact on emissions, with a 21.18 gram increase per km for every additional liter per 100 km consumed. The effect of fuel type is also substantial, with electric cars (Fuel Type E) resulting in a large reduction in predicted CO2 emissions.

The model's F-statistic is extremely high, which indicates the collective effect of all the variables in the model is highly significant in predicting CO2 emissions. The very low p-values across most predictors

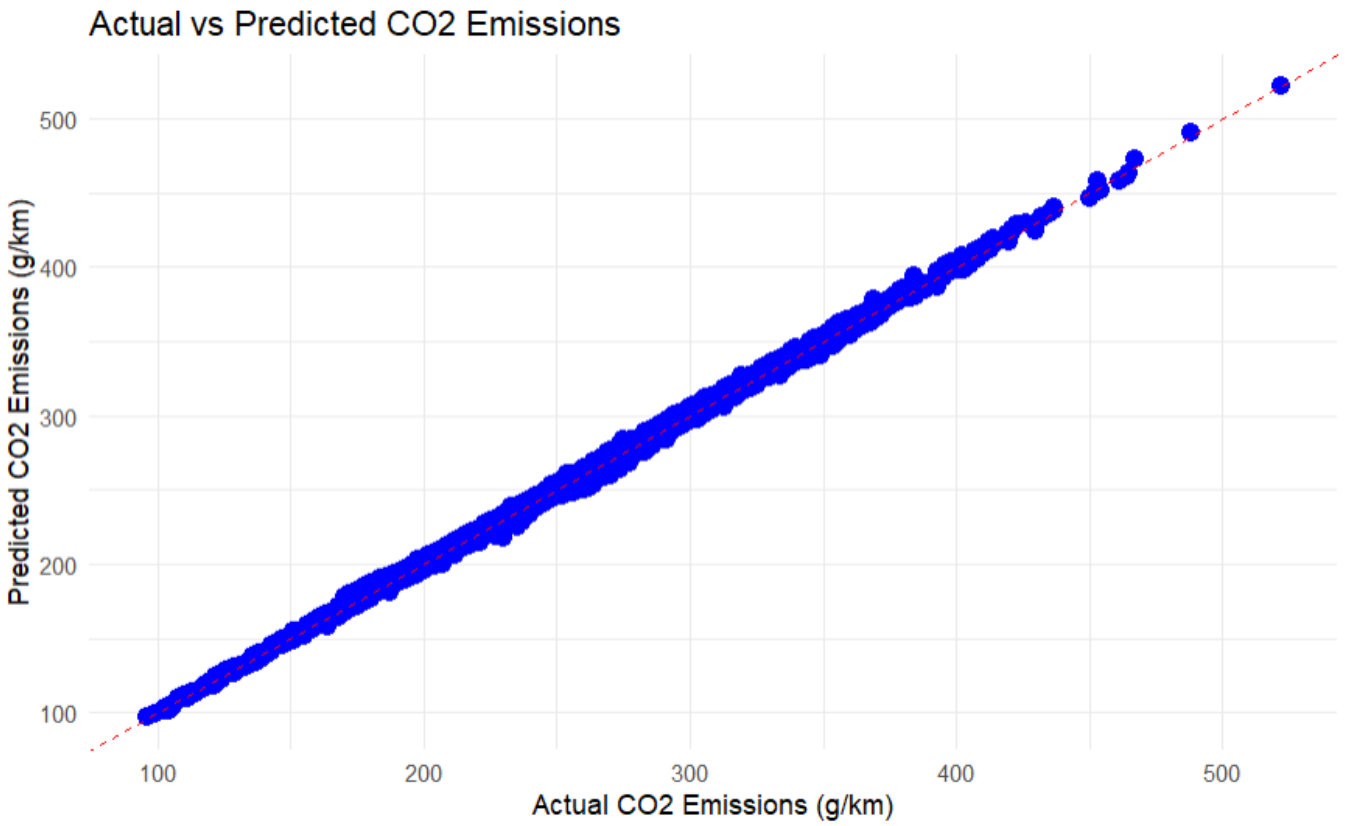
confirm their individual significance. The small residual standard error (2.46) indicates that the model's predictions are very close to the actual values.

Diagnostics of Clean model 2:



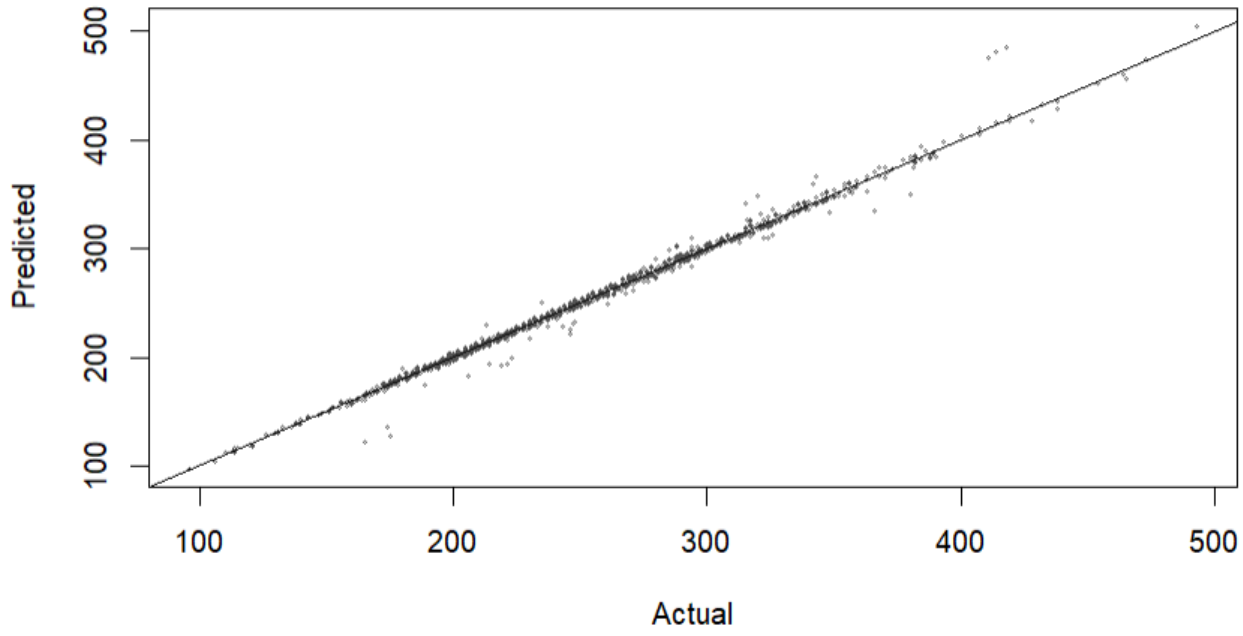
Subsequent data cleaning has yielded a QQ plot where residuals neatly follow the normal line, affirming the normality assumption of linear regression. Concurrently, the Residuals vs. Fitted plot reveals a random scatter of points with no discernible pattern, signifying that both the linearity and homoscedasticity assumptions are satisfied. These improvements in diagnostic plots reflect a robust model that reliably captures the relationship between the predictors and the dependent variable, enhancing the validity of its predictions for CO2 emissions and reinforcing its statistical soundness for inferential purposes.

Seeing how well the data fits the regression line



The Actual vs Predicted plot now shows a denser congregation of data points along the identity line, a clear sign of improved model accuracy post-data cleaning. This enhanced alignment indicates that the model's predictions are now more consistent with actual observations, reflecting a substantial increase in the explained variance of CO2 emissions. The refined model demonstrates stronger predictive power, suggesting a successful capture of the underlying data structure and a more reliable tool for forecasting and analysis within the scope of the environmental data it represents.

Seeing how well the data fits the regression line on test data



The scatter plot depicting the Actual vs Predicted values for the test set shows a tight and consistent alignment of data points along the diagonal, which indicates a strong positive correlation between the predicted and actual CO2 emissions. The proximity of points to this diagonal line of perfect prediction denotes that Model 2, after rigorous cleaning and validation, has proven to be highly effective in capturing the underlying pattern and variance in the data.

The distinct lack of systematic deviation from the diagonal and the dense clustering of points around it underscore the model's accuracy and its generalization capability to new, unseen data. Such an outcome signifies that the predictive model not only learned the training data well but also adapted effectively to the test set without overfitting or underfitting.

Hence, "clean model 2" is the best model for our chosen dataset.

Summary of Statistical Methods Used:

In this comprehensive statistics project, a variety of statistical methods were employed to analyze a large dataset containing details of 7,385 vehicles. The dataset included diverse attributes such as make, model, vehicle class, engine size, number of cylinders, transmission type, fuel type, fuel consumption, and CO2 emissions. The analysis aimed to uncover patterns and correlations that could inform vehicle production decisions and assess environmental impacts.

Descriptive Statistics and Histogram Analysis: The project began with descriptive statistical techniques to summarize the key features of the dataset. Ranges for engine sizes, cylinder counts, fuel consumption in city and highway conditions, and CO2 emissions were provided. Such summary statistics are crucial for an initial understanding of the dataset's scope and variability.

Histograms played a pivotal role in visually representing the distribution of CO2 emissions and combined fuel consumption. These histograms offered a clear visual interpretation of data frequency and distribution, highlighting the central tendency and dispersion of emissions and fuel consumption across the dataset.

Scatterplot and Correlation Analysis: To explore relationships between variables, scatterplots were utilized. One significant analysis was the examination of the correlation between CO2 emissions and fuel efficiency (measured in miles per gallon). The scatterplot indicated a negative correlation, suggesting that higher CO2 emissions are associated with lower fuel efficiency. This pattern was indicative of a potential non-linear relationship, a vital insight for understanding the environmental impact of different vehicle types.

Simple and Multiple Linear Regression Models: The core of the analysis involved linear regression. Initially, a simple linear regression model was used, calculating the average CO2 emissions with just an intercept. This provided a baseline understanding of emissions across the dataset.

The project then advanced to multiple linear regression, which included several predictors like car make, engine size, and fuel type. This model was highly effective (R-squared value of 0.9938), indicating that it could explain almost all the variability in CO2 emissions based on the included variables.

Diagnostic Plot Analysis: To ensure the reliability of the regression models, diagnostic plots were examined. These plots revealed potential issues like non-linearity, deviations from normality, heteroscedasticity, and the presence of influential outliers. Addressing these issues is crucial for improving the model's accuracy and reliability.

Variance Inflation Factor (VIF) Analysis: The project also included an analysis of multicollinearity through the computation of Variance Inflation Factors (VIF). High VIF values for variables like fuel consumption in the city, on the highway, and combined fuel consumption indicated a strong interdependence among these predictors, a common occurrence in automotive datasets.

Model Validation and Predictive Accuracy Assessment: The final phase involved validating the regression models and assessing their predictive accuracy. This was done through plots comparing actual versus predicted CO2 emissions. A tight alignment of data points along the identity line in these plots indicated high predictive accuracy, especially post-data cleaning, where the models were adjusted to address earlier diagnostic concerns.

Overall, this statistical analysis provided a comprehensive and nuanced understanding of vehicle attributes, fuel consumption, and emissions. The methodologies employed allowed for robust predictions and insights, contributing significantly to the discourse on vehicle efficiency and environmental sustainability.

A key finding of this comprehensive statistical analysis was that fuel type, particularly types N and E, emerged as the most powerful predictors in the data. This insight is particularly relevant in understanding the environmental impact of vehicles and can guide future vehicle production and policy decisions aimed at sustainability.

Conclusion:

In conclusion, this project successfully utilized statistical methods to analyze a comprehensive dataset of 7,385 vehicles, providing valuable insights into the relationship between vehicle characteristics and environmental impact. The analysis revealed that fuel types N and E were the most powerful predictors of CO2 emissions, underscoring the significant role of fuel choice in determining a vehicle's environmental footprint. Through descriptive statistics, correlation analysis, and advanced regression modeling, the study not only highlighted patterns in fuel consumption and emissions but also demonstrated the intricate interplay between various vehicle attributes. The project's findings are instrumental for automotive manufacturers and policymakers in making informed decisions aimed at reducing emissions and promoting sustainable practices in the automotive sector. The rigorous statistical approach adopted ensures that the conclusions drawn are both reliable and applicable to real-world scenarios, contributing meaningfully to the ongoing discourse on sustainable transportation.