Predicting Carbon Dioxide Emissions from Vehicles Based on Their Characteristics

STAT-632 Spring 2024: Final Paper

Humza Shah, Faizan Khan, Jaspreet Kang, Nishanth Reddy Lankala

2024-05-04

1 Introduction

Carbon dioxide (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 negative effects, including rising global temperatures, sea level rise, changes in weather patterns, and ecosystem disruptions.

In light of these events occurring, reducing carbon dioxide emissions is essential for mitigating climate change and its harmful effects on humankind. This involves transitioning to cleaner and renewable energy sources to improve the climate trends around us and our living society.

This paper focuses on developing a multiple linear regression (MLR) model to estimate CO2 emissions from different types of vehicles, using vehicle characteristics as dependent variables such as fuel efficiency, fuel type, engine size, and more. Weighted MLR and random forest models are also explored.

2 Data Description

2.1 About the Data

A dataset on vehicle CO2 emissions was sourced from Kaggle and used in this analysis. The dataset contained 7,385 rows, 12 variables, and no missing values. Each row represented a specific vehicle model and its characteristics. The first six rows of the downloaded dataset are shown in Table 3 in the Appendix section.

The response variable to be predicted is *CO2 Emissions (g/km)*. The potential predictors include the class of the vehicle, engine size (L), number of cylinders, transmission type, fuel type, and at least one measure of fuel efficiency. The variables for car brand and model were excluded to ensure the development of a robust model capable of generalizing well across the broader population of vehicles and prevent overfitting to the sampled observations. Additional information regarding the response and predictor variables is available in Table 4 in the Appendix section.

2.2 Frequencies

Figure 1 illustrates the percentage frequencies of the categorical variables in the dataset. Small and standard SUVs, mid- and full-size vehicles, and compact vehicles each comprised between 10% to 15% of the data. Approximately 47% and 33% of the dataset consisted of 4-cylinder and 6-cylinder cars, respectively. Around 90% of the dataset comprised vehicles fueled by either regular or premium gasoline. The majority of vehicles had automatic transmissions.

Figure 2 displays a density plot of the quantitative variables. All variables, including the response variable *CO2 Emissions*, exhibited right-skewness. This observation suggests that transformations may be necessary, particularly for the response variable.

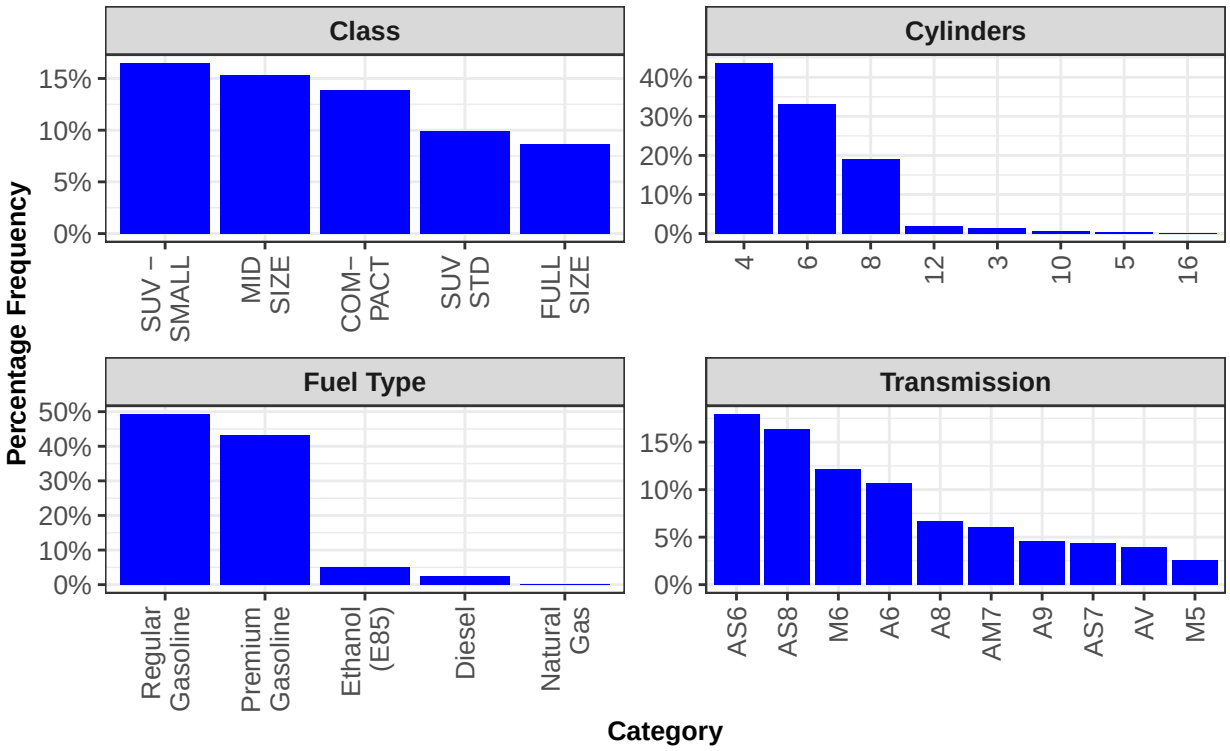


Figure 1: Percentage Frequencies of Categorical Variables

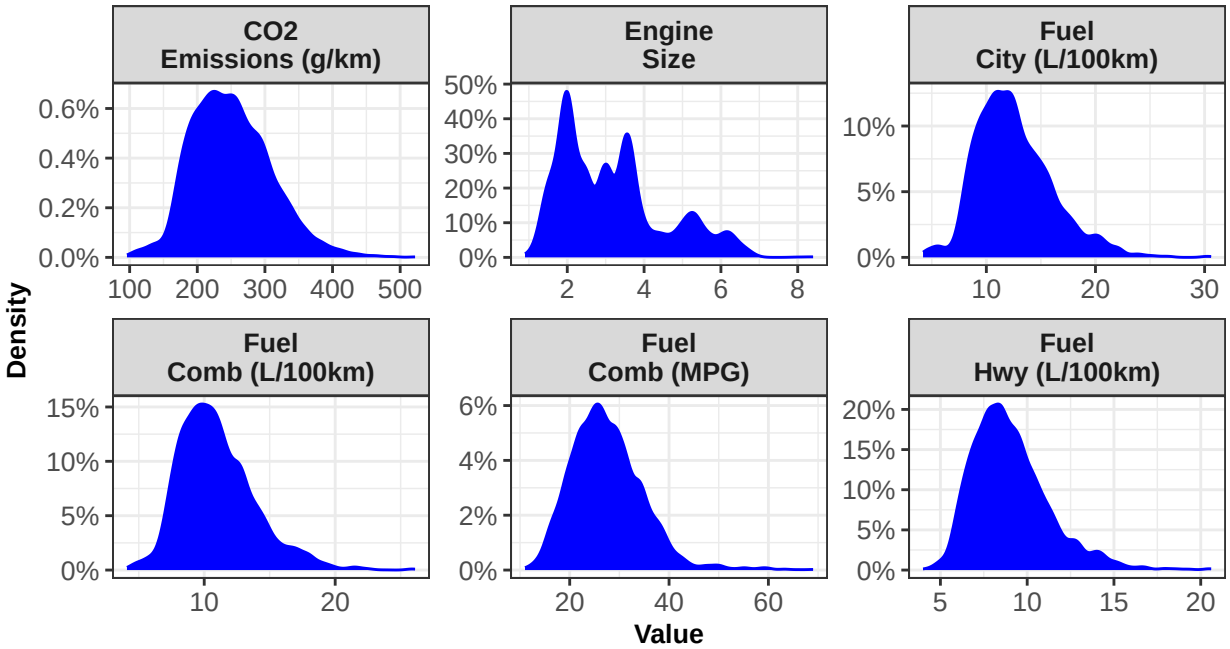


Figure 2: Distributions of Quantititative Variables

2.3 Relationships Between the Response and Categorical Predictors

Figure 3 illustrates that smaller classes of vehicles, such as small station wagons and compact vehicles, generally emit lower CO2 levels, while larger classes, including cargo and passenger vans, exhibit higher CO2 emissions. Additionally, there appears to be a positive linear association between the number of *cylinders* and *CO2 Emissions*.

Strong associations are observed between *CO2 Emissions* and each quantitative variable in Figure 4. The correlation coefficient (r) between *CO2 Emissions* and *Engine size* is 0.851, with a linear or slight parabolic relationship. Additionally, each fuel-efficiency variable exhibits an absolute value of r exceeding 0.884 with *CO2 Emissions* and over 0.90 with each other, suggesting strong correlations. Given their interrelatedness, using only one fuel-efficiency variable may mitigate potential multicollinearity issues. Notably, unrelated or parallel regression lines are observed between the fuel efficiency predictors and *CO2 Emissions*.

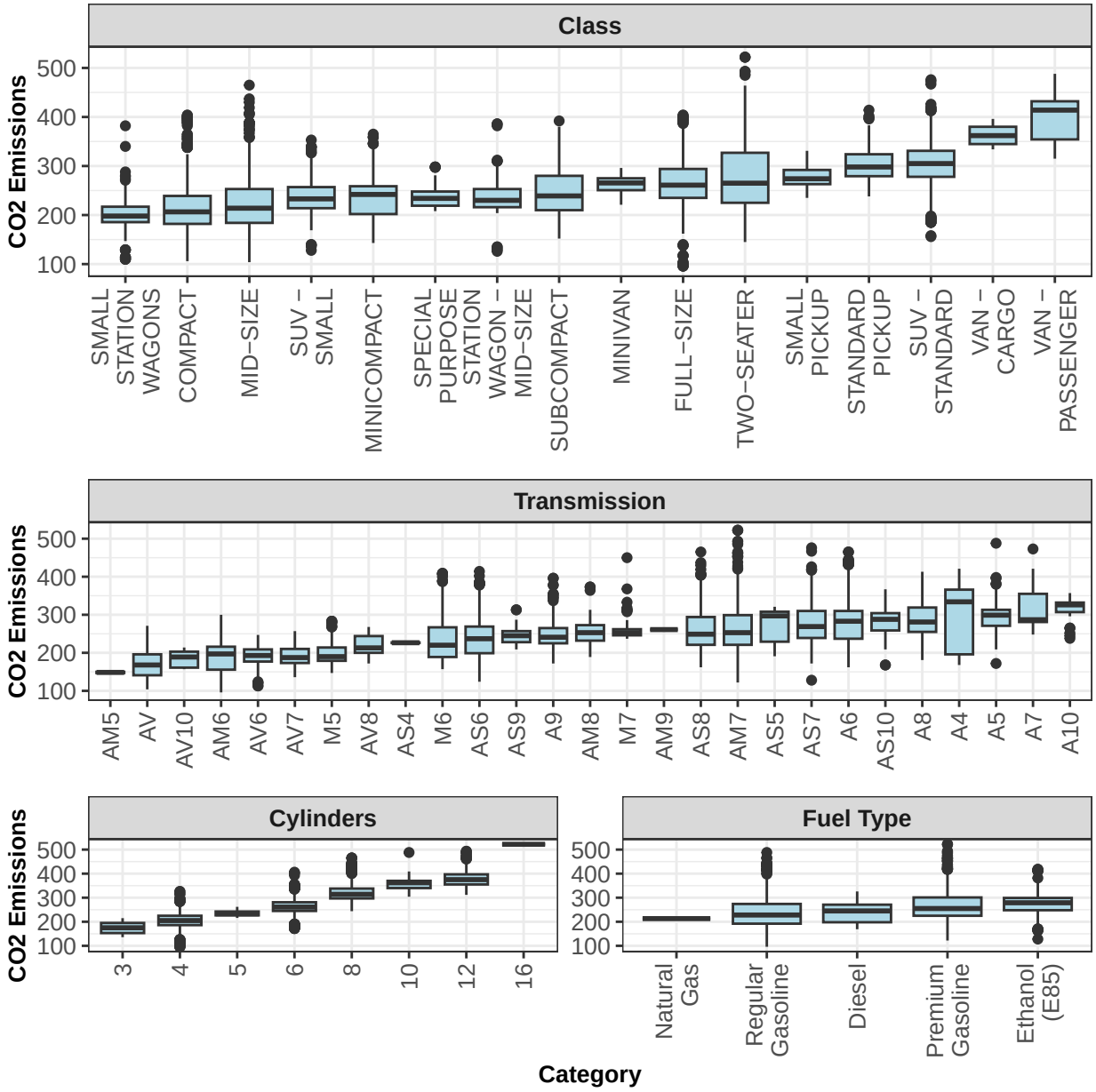


Figure 3: Summary Statistics Between each Categorical Predictor and CO2 Emissions

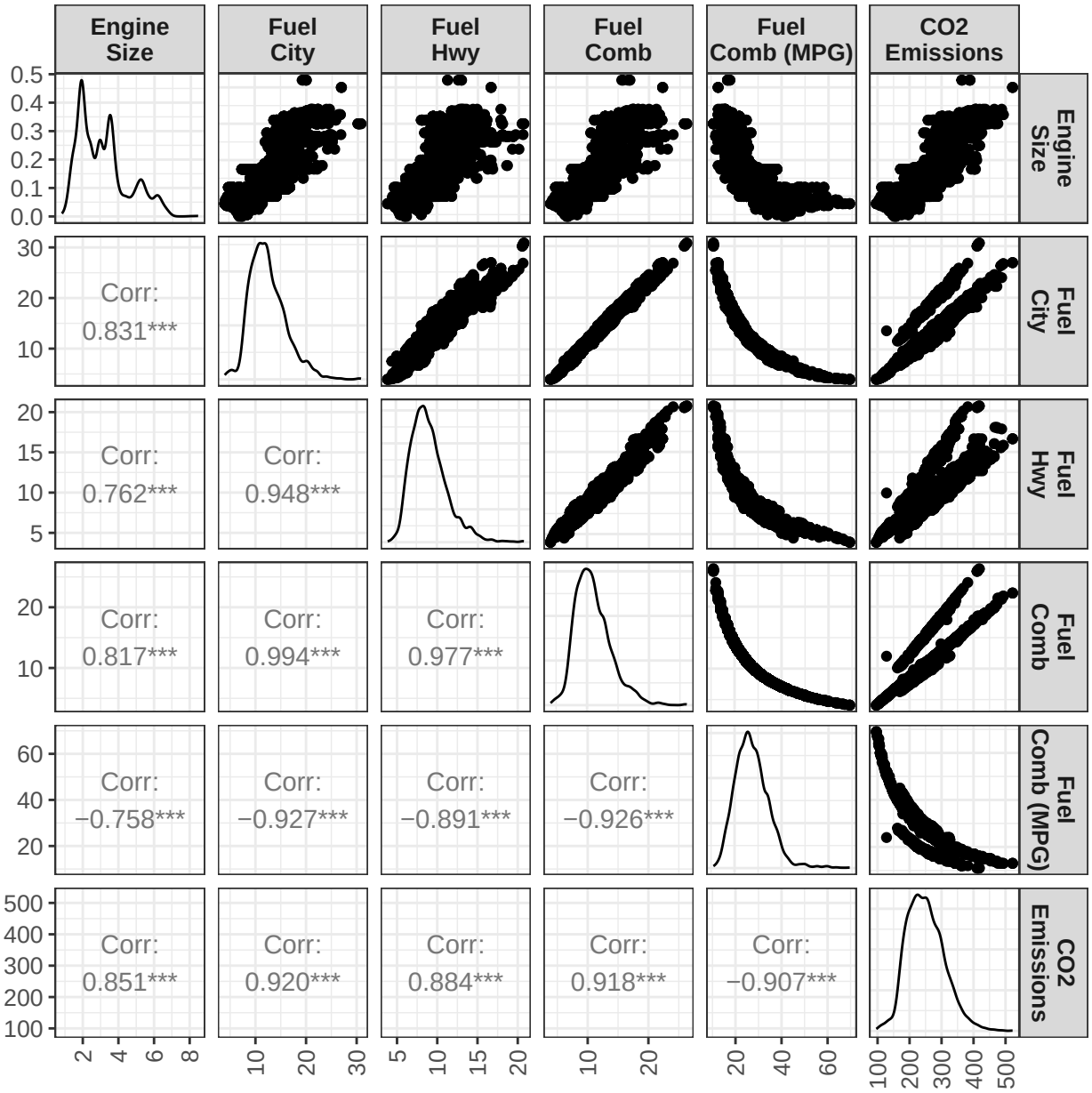


Figure 4: Scatterplot and Correlation Matrix of CO2 Emissions and Quantitative Predictors

2.4 Closer Look at the Relationships Between CO2 Emissions and Fuel Efficiency

Figure 5 clearly illustrates the existence of parallel regression lines. Distinct clusters of points with similar slopes are observed for each fuel efficiency predictors in relation to *CO2 Emissions*. The clusters, delineated by color, represent the different fuel types. This suggests that the *fuel type* predictor will introduce an additive effect on the prediction of *CO2 Emissions* by altering the intercept of the multiple linear regression model.

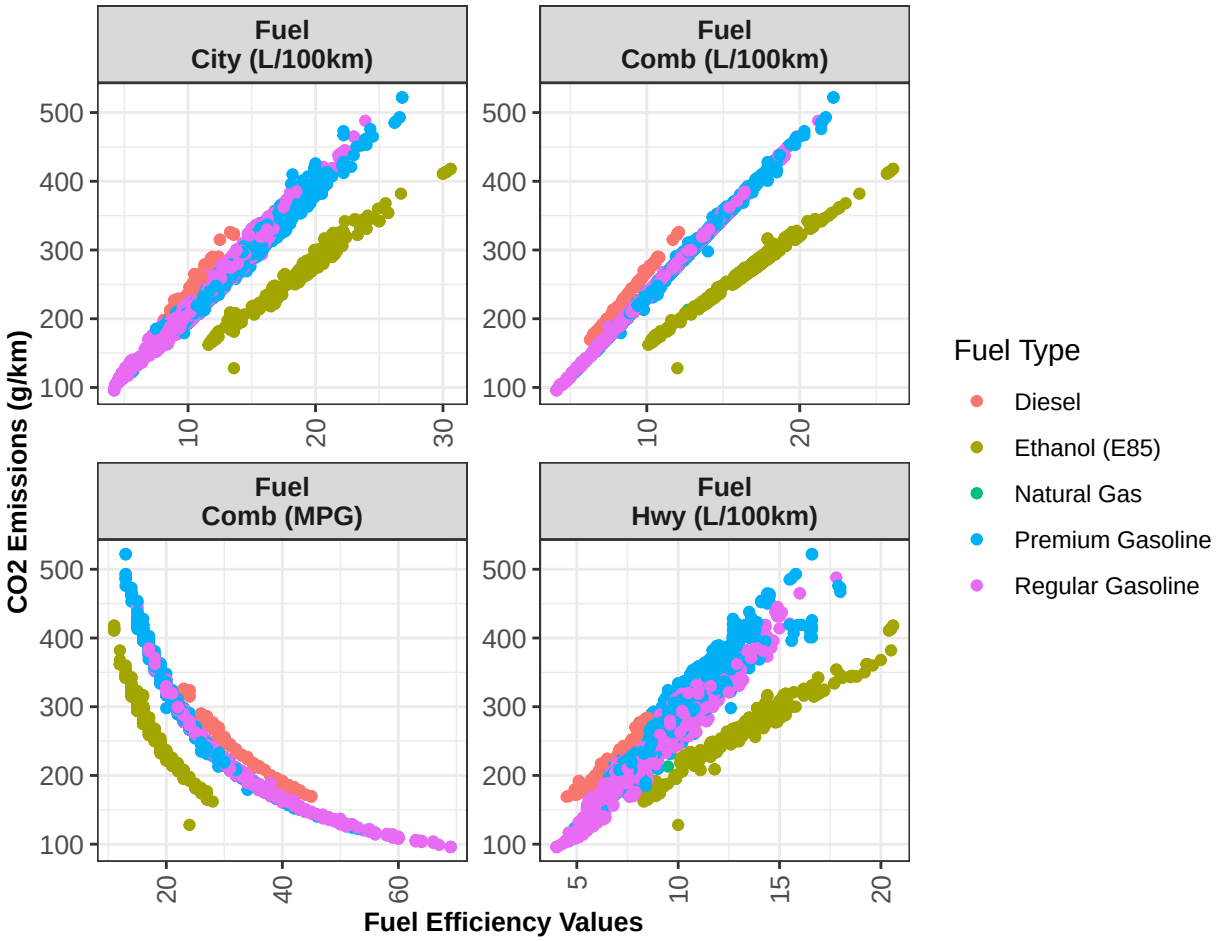


Figure 5: Scatterplots of Fuel Efficiency Predictors and CO2 Emissions by Fuel Type

3 Methods and Results

Three models are explored to predict *CO2 Emissions*:

1. **Multiple Linear Regression Model (MLR)**: Developed with several iterations of variable transformation and selection methods to create a parsimonious model with statistically significant variables. Aimed at ensuring absence of multicollinearity issues and adherence to the assumptions of a valid linear regression model.
2. **Weighted Multiple Linear Regression Model**: Created to address potential nonconstant variance of the residuals in the non-weighted MLR model.
3. **Random Forest**: A machine learning approach utilized for prediction.

3.1 Pre-processing

3.1.1 One-Hot Encoding

Prior to model training, a preprocessing technique known as **one-hot encoding** was applied to the data. This technique transforms categorical predictors (such as fuel type, number of cylinders, vehicle class, and transmission) into numerical, binary values represented by 0s and 1s. Each category within a predictor is converted into a binary feature, where 0 indicates absence and 1 indicates presence of that category. For instance, the *fuel type* predictor, which has five distinct categories (diesel, ethanol, natural gas, regular gasoline, and premium gasoline), is transformed into five binary features. Each feature represents one of the fuel types and is encoded as 1 if true and 0 if false.

3.1.2 Predictor Transformations

As seen in Figure 2, all fuel-efficiency-related variables exhibited approximately normal distributions with right-skewness. Although correcting for this skewness was not mandatory, log-transformation was applied to all fuel efficiency predictors due to improved model performance results.

3.1.3 Cross-Validation: Training and Testing

The following steps were undertaken to create training and testing sets:

1. The data was initially partitioned based on vehicle class which has 16 unique categories. Therefore, 16 subsets were created.
2. 70% of the observations from each subset was randomly sampled and combined into one training set. The remaining 30% within each subset was combined into one testing set.

As a result, the training and testing sets comprised 5,164 and 2,221 observations, respectively. Each set exhibited a comparable distribution of vehicle classes and other features.

3.2 Model #1: Multiple Linear Regression Model

To begin, a full multiple linear regression model was constructed with *CO2 Emissions* as the response variable and all predictors included. The full model yielded an R-squared (R^2) value of 0.9862, meaning around 98.32% of the variation in *CO2 Emissions* is explained by the model. The adjusted R-squared (R^2_{adj}) value was 0.986.

However, approximately half of the predictors were found to be statistically insignificant. Subsequently, Figure 6 was utilized to assess several linear regression assumptions:

1. **Linearity**: While the linearity assumption appears to be satisfied, a minor parabolic relationship is observed.
2. **Normality of Errors**: The normality assumption was violated.
3. **Constant Variance**: The assumption of constant variance in the errors was not met.

Furthermore, Figure 6 was utilized to detect outliers and high leverage points. Several outliers were identified within the dataset.

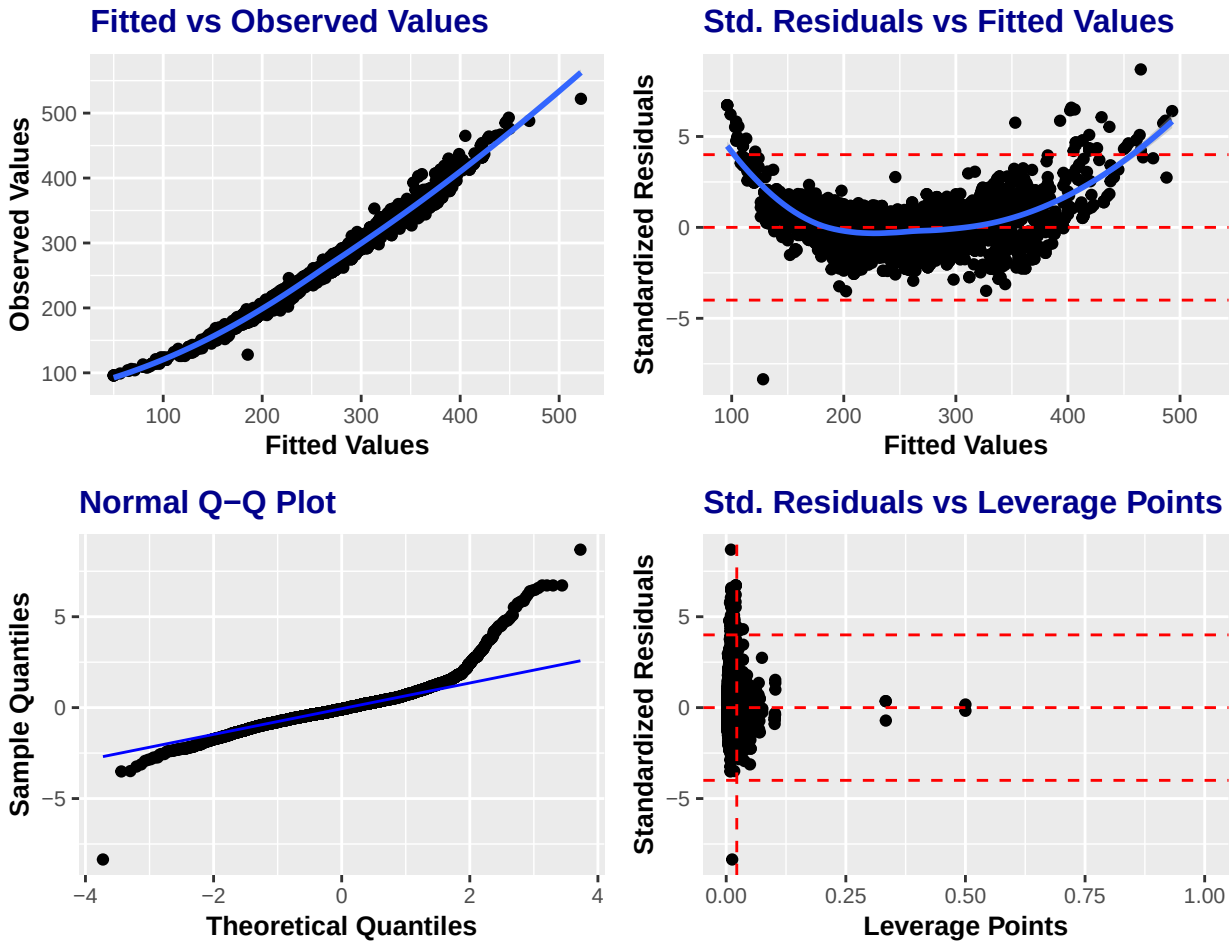
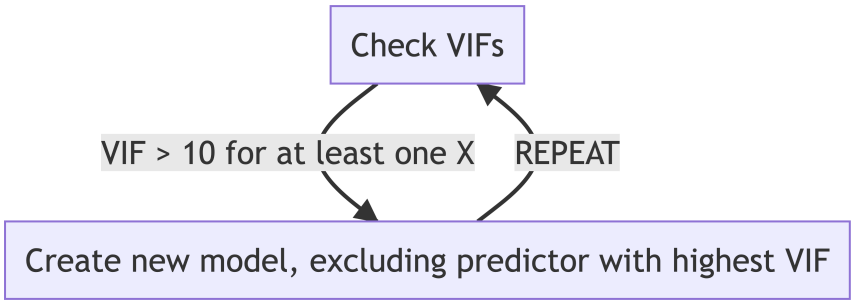


Figure 6: Summary Plots for Full MLR Model

3.2.1 Checking for and Addressing Multi-collinearity

Multicollinearity was assessed by examining the Variation Inflation Factor (VIF) of each predictor. In the full multiple linear regression model, 27 variables exhibited a VIF score exceeding 10, with 6 variables having a VIF value surpassing 100. To mitigate the influence of highly collinear predictors, the following process was repeated:



A total of eight predictors were removed from the model, including: Fuel Efficiency (City), Fuel Efficiency (Highway), Fuel Efficiency (Combined - L/100km), 4-cylinder, Transmission AS6, Regular Gasoline Fuel Type, Engine Size, and Class Small SUV. Of note, only one fuel-efficiency predictor, Fuel Efficiency (Combined - MPG), remained in the model.

3.2.2 Transformation of the Response

The Box-Cox power transformation was applied to the response variable, *CO2 Emissions*. The recommended power value of 0.05 was identified, as depicted in Figure 7. Therefore, *CO2 Emissions* was transformed to $CO2Emissions^{0.05}$.

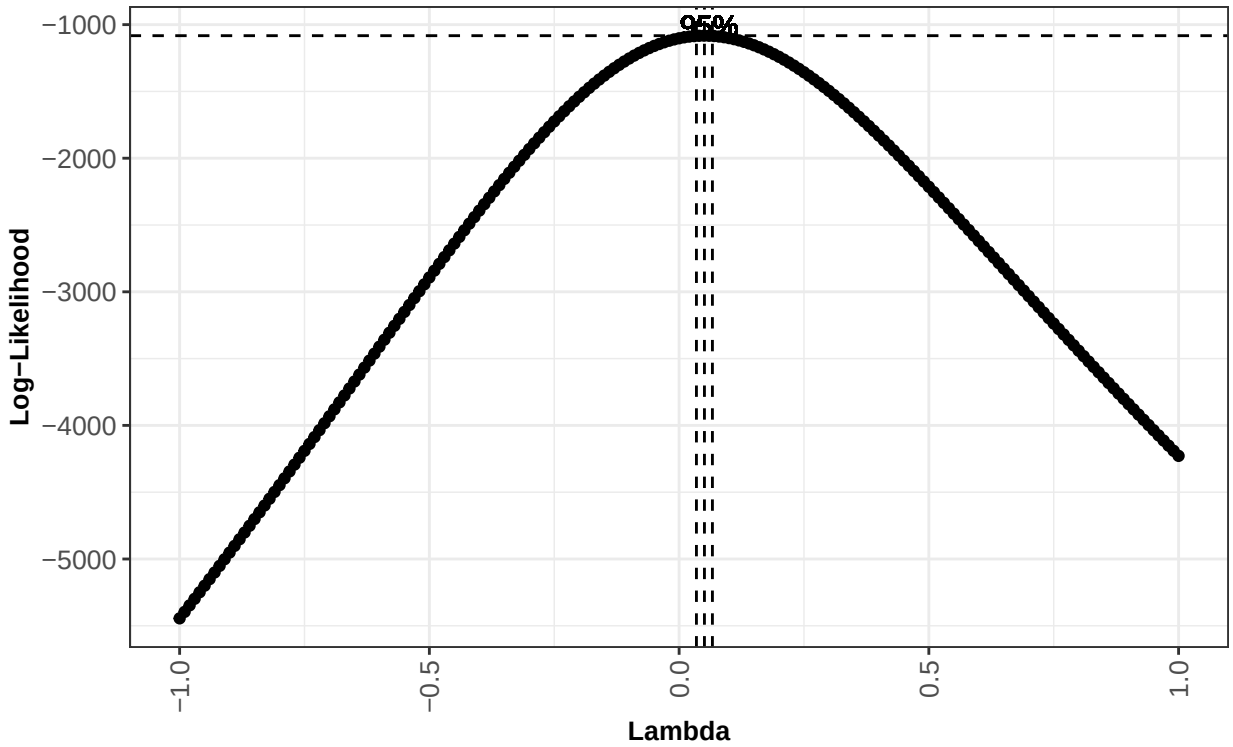


Figure 7: Log-likelihoods for the Parameter of the Box-Cox Power Transformation

3.2.3 Stepwise Variable Selection (AIC)

Backward stepwise variable selection was conducted using the Akaike Information Criterion (AIC). This iterative process involves removing predictor variables from the model based on their impact on the AIC value. The objective is to identify a model that achieves a balance between explanatory power and simplicity, thereby identifying the most significant predictors while avoiding overfitting.

During the backward stepwise selection, 17 predictors were removed, all of which were transmission and vehicle class variables.

3.2.4 Inspect and Address Outliers

Nine outliers with |standardized residuals| greater than four were identified and are listed in Table 1. Various methods were employed to address these outliers, including:

1. **Verification Using Fuel Economy**: Each vehicle model was cross-referenced with Fuel Economy, the official U.S. government source for fuel economy information. Discrepancies were noted between the fuel efficiency and CO2 emission values reported in the dataset and those listed on Fuel Economy.
2. **Comparison with Similar Vehicles**: Outliers were compared with other vehicles in the dataset that possessed similar or identical specifications.

Using the two methods above, CO2 emission values for outliers were adjusted to either match the values listed on Fuel Economy or reflect the median of similar or identical vehicles identified. If no relevant information could be found or if the outlier persisted, it was removed from the dataset. The changes made to the outliers are documented in Table 1.

Subsequently, a new version of the model was trained on the revised dataset with outliers addressed. The Box-Cox power transformation method was re-executed, yielding a new lambda value of 0.03 or $CO2Emissions^{0.03}$.

Table 1: Outliers

Vehicle	CO2 Emissions	Std Residuals	New CO2 Emissions Value	Method
MERCEDES-BENZ GLA 250 4MATIC	128	-23.03	Removed	Direct comparisons could not be made
CHEVROLET IMPALA DUAL FUEL	213	-19.36	Removed	Direct comparisons could not be made
MERCEDES-BENZ E 300 4MATIC	246	6.59	228	Median value of other identical cars in dataset
GMC SIERRA 4WD FFV	317	5.36	Removed	Direct comparisons could not be made
MERCEDES-BENZ GL 450 4MATIC	298	-5.25	Removed	Direct comparisons could not be made
CHEVROLET Malibu	189	5.08	Removed	Direct comparisons could not be made

Vehicle	CO2 Emissions	Std Residuals	New CO2 Emissions Value	Method
MERCEDES-BENZ AMG CLS 53 4MATIC+	235	-4.30	Removed	Direct comparisons could not be made
MERCEDES-BENZ AMG CLS 53 4MATIC+	235	-4.30	Removed	Direct comparisons could not be made
MERCEDES-BENZ B 250	179	-4.22	186	Fuel Economy Govt Source

3.2.5 Final Model #1: Multiple Linear Regression

After addressing multi-collinearity, transforming the response variable twice using the Box-Cox method, performing backward stepwise variable selection, and addressing outliers, the final model yields an improved R^2_{adj} value of 0.9958 with fewer predictors. For comparison, the initial model's R^2_{adj} value was 0.986. Table 5 in the Appendix section displays the remaining predictors, with the most statistically significant variables identified as *Fuel Efficiency Combined (MPG)*, *Fuel Type Ethanol*, and *Fuel Type Diesel*.

Assessments of the final model's assumptions are depicted in Figure 8. Both linearity and normality assumptions appear to be satisfied, with the Shapiro Test validating the normal distribution of residuals (p-value: 0.139).

While the variance of the residuals has improved and stabilized compared to the initial model, some evidence of nonconstant variance persists. This is corroborated by the Breusch-Pagan test, indicating the presence of heteroscedasticity ($p < 0.001$).

Due to the presence of nonconstant error variance, a weighted least squares model is attempted.

3.3 Weighted Least Squares

Weights derived from the final multiple linear regression model were utilized to construct a weighted model. Weights were obtained by:

1. Regressing the |residuals| from the final MLR model onto to final MLR fitted values.
2. Squaring the fitted values from the new regression model.
3. Calculating the inverse.

As a result, the R^2_{adj} value increased from 0.9958 to 0.9966.

Assessment of the final model's assumptions is presented in Figure 9. The linearity assumption remains satisfied, and the Shapiro Test indicates that the residuals continue to exhibit a normal distribution ($p = 0.052$)

Notably, the variance of the residuals appears more constant without discernible patterns. The Breusch-Pagan test suggests that the error variance is homoscedastic ($p = 0.406$).

In summary, the weighted least squares model satisfies all assumptions and results in an improved R^2_{adj} value.

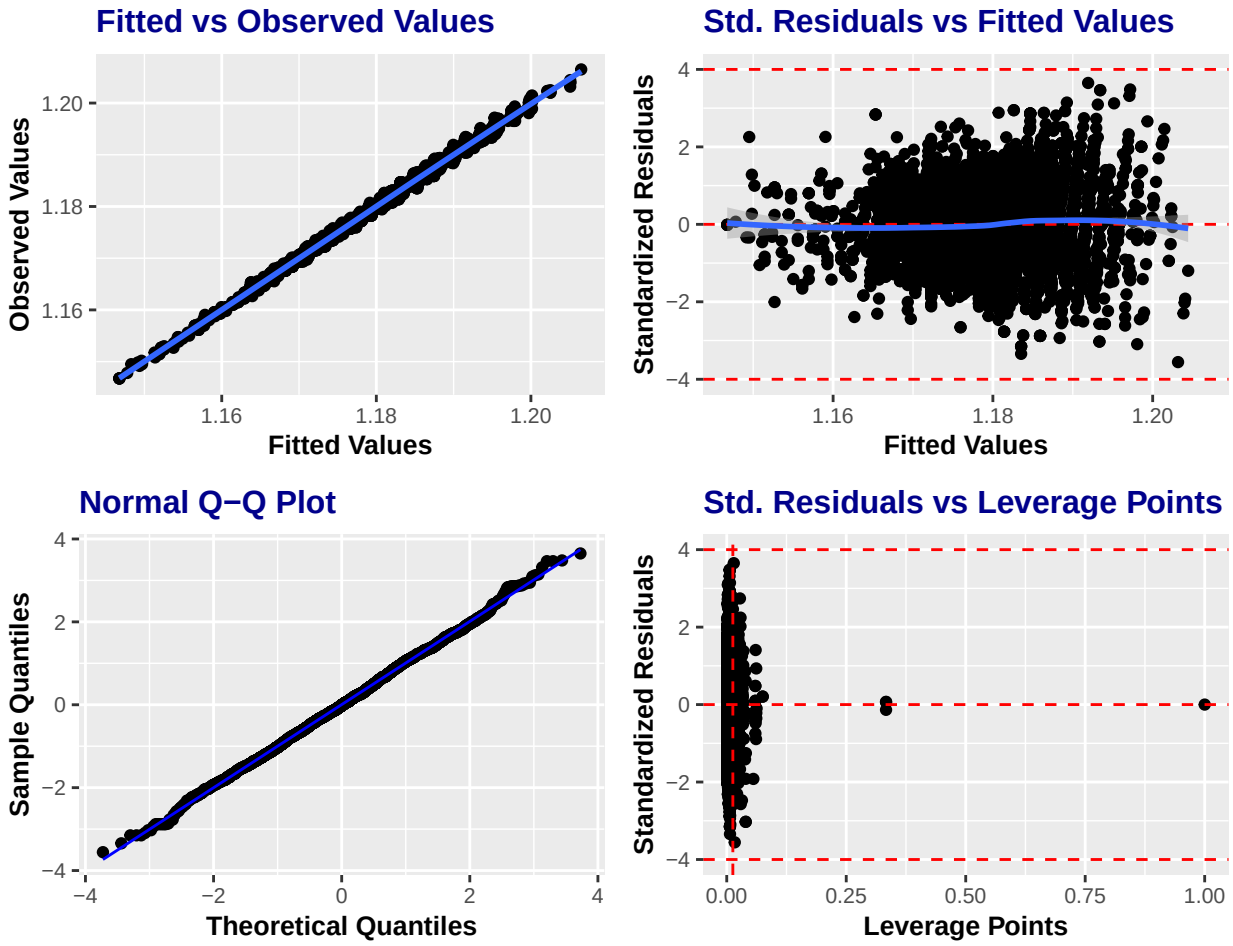


Figure 8: Summary Plots for Final MLR Model

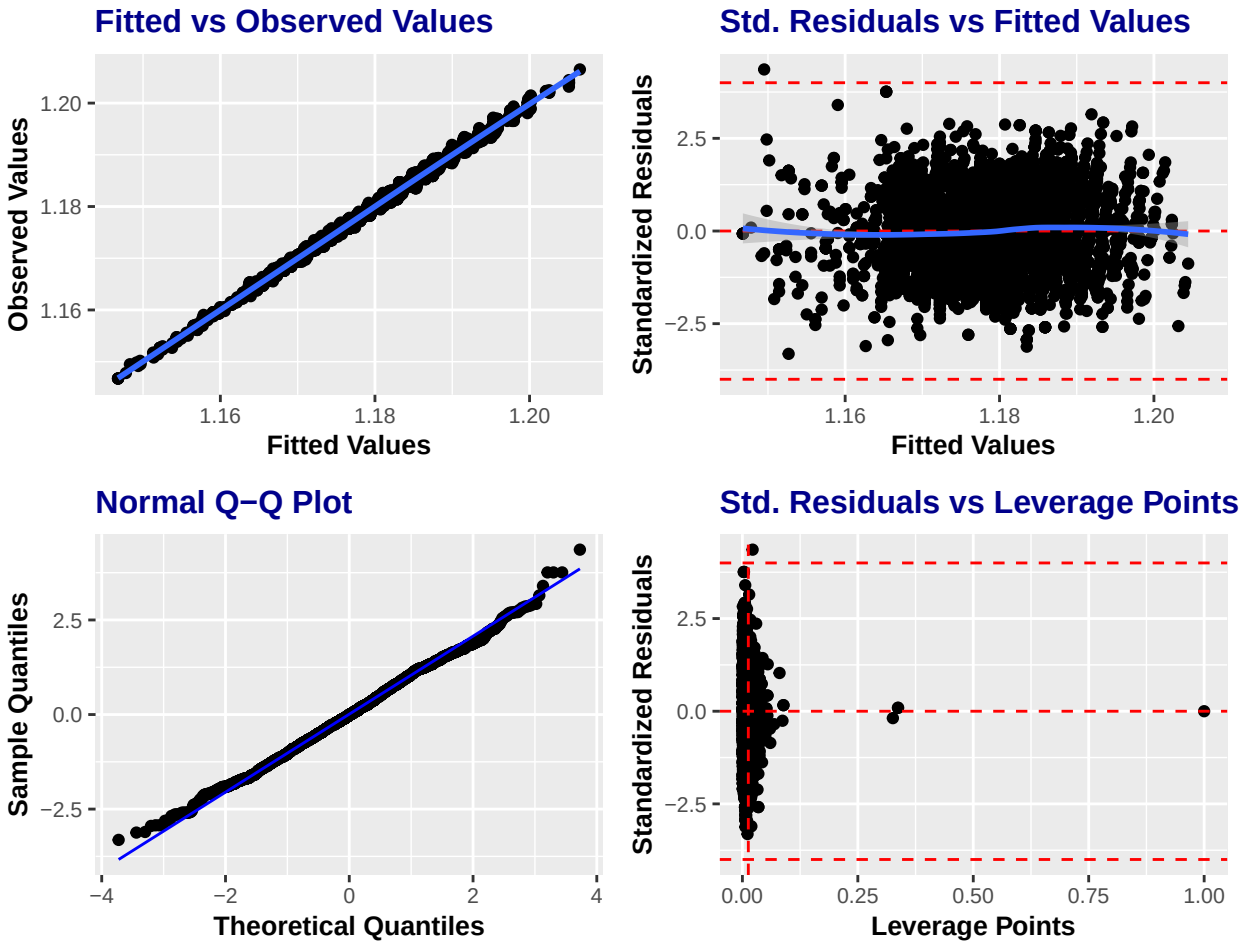


Figure 9: Summary Plots for Weighted MLR Model

3.4 Random Forest

A random forest model consisting of 500 trees was trained. To ensure fair comparisons with other models, the random forest model was fitted on the training dataset with outliers addressed.

Percentage of the variance explained from the random forest model was 99.73%. Figure 10 displays a variable importance plot, illustrating the percentage increase in mean squared error (MSE) if each variable were removed. Consistent with the non-weighted and weighted multiple linear regression models, a fuel-efficiency-related variable, *fuel type diesel*, and *fuel type ethanol* emerged as the most predictive variables.

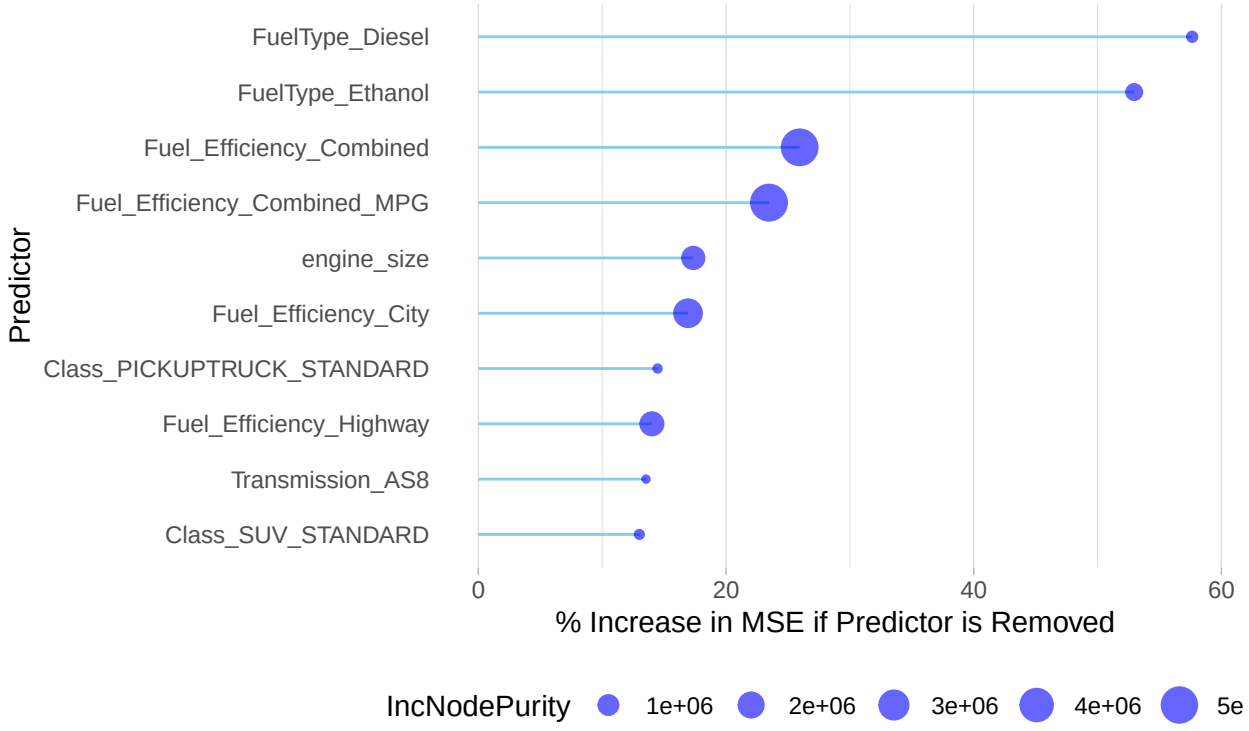


Figure 10: Variable Importance Plot

4 Conclusion

4.1 Comparing Models

A comparison of the three models developed was conducted using AIC, R^2_{adj}, and the testing root mean squared error (RMSE), as shown in Table 2. AIC and R^2_{adj} were only available for the non-weighted and weighted MLR models. Based on these two measures, the weighted MLR model outperformed the non-weighted MLR model.

The random forest model exhibited a substantially lower testing RMSE of 3.33 compared to 4.20 in both MLR models. However, the percent change from training RMSE to testing RMSE was 8.33% in the random forest model, contrasting with less than 1% in the MLR models. This suggests potential overfitting in the random forest model.

4.2 Major Findings

In the non-weighted MLR model, the variance of the residuals remained non-constant with a discernible pattern (funnel-shaped). The weighted MLR model corrected the nonconstant variance while also leading to a higher R^2_{adj} value.

In the MLR and random forest models, *Fuel Type Ethanol*, *Fuel Type Diesel*, and a fuel-efficiency variable were found to be the most predictive of the response variable *CO2 Emissions*.

4.3 Limitations

Quality issues were identified in the dataset. Numerous instances were found where either the fuel efficiency or CO2 emission values in the dataset substantially differed from credible sources such as fueleconomy.gov.

Furthermore, discrepancies were found among flex-fuel vehicle (FFV) observations. Half of the FFV entries were categorized with a fuel type of Ethanol, while the remaining half were labeled as Regular or Premium gasoline, despite no other discernible differences observed in these vehicle characteristics.

Table 2: Performance Measures for Each Model

Model	AIC	R2	R2 (Adjusted)	Training RMSE	Testing RMSE	% Change Train -> Test RMSE
Multiple Linear Regression (MLR)	-62,902.25	99.5821%	99.5794%	4.16046	4.19826	0.9086%
Weighted MLR	-63,067.72	99.6583%	99.6562%	4.16605	4.19826	0.7732%
Random Forest				3.07394	3.33003	8.3310%

5 Appendix

5.1 R Code

R code to replicate the analysis can be found in this Github repository.

5.2 Tables

Table 3: Head of the Vehicle C02 Emissions Dataset

							Fuel Efficiency			
Make	Model	Class	Engine Size (L)	Cylin-ders	Trans-mission	Fuel Type	City (L/ 100km)	High-way (L/ 100km)	Comb-ined (L/ 100km)	Comb-ined (MPG)	CO2 Emis-sions (g/km)
ACURA	ILX	COMPACT	2.00	4	AS5	Z	9.90	6.70	8.50	33	196
ACURA	ILX	COMPACT	2.40	4	M6	Z	11.20	7.70	9.60	29	221
ACURA	ILX HYBRID	COMPACT	1.50	4	AV7	Z	6.00	5.80	5.90	48	136
ACURA	MDX 4WD	SUV - SMALL	3.50	6	AS6	Z	12.70	9.10	11.10	25	255
ACURA	RDX AWD	SUV - SMALL	3.50	6	AS6	Z	12.10	8.70	10.60	27	244
ACURA	RLX	MID-SIZE	3.50	6	AS6	Z	11.90	7.70	10.00	28	230

L = Liters; L/100km = Liters per 100 kilometers; MPG = milers per gallon; g/km = grams per kilometer

Table 4: Vehicles CO2 Emissions Metadata

Variable	Variable Type	Variable Description	Data Range/Unique Categories
Make	Categorical	Brand or manufacturer of the vehicle.There are 42 unique values.	ACURA, ALFA ROMEO, ASTON MARTIN, AUDI, BENTLEY, BMW, BUGATTI, BUICK, CADILLAC,
Model	Categorical	Specific model of the vehicle. There are 2,053 unique values.	4C, A4, A4 QUATTRO, A5 CABRIOLET QUATTRO, A5 QUATTRO, A6 QUATTRO,
Vehicle Class	Categorical	General category or type of vehicle based on size and purpose. There are 16 unique values.	COMPACT, FULL-SIZE, MID-SIZE, MINICOMPACT, MINIVAN, PICKUP TRUCK - SMALL,
Engine Size (L)	Quantitative	Engine size in liters.	Ranges from 0.9 to 8.4
Cylinders	Categorical	Number of cylinders in the vehicle's engine, ranging from 3 to 16.	3, 4, 5, 6, 8, 10, 12, 16
Transmission	Categorical	The type and number of gears in the vehicle's transmission. There are 27 unique values.	A10, A4, A5, A6, A7, A8, A9, AM5, AM6, AM7, AM8, AM9, AS10, AS4, AS5, AS6, AS7, AS8, AS9, AV, AV10, AV6, AV7, AV8, M5, M6, M7

Variable	Variable Type	Variable Description	Data Range/Unique Categories
Fuel Type	Categorical	The type of fuel used by the vehicle. There are 5 unique values notated as one-letter codes: Z = Premium gasoline; D = Diesel; X = Regular gasoline; E = Ethanol (E85); N = Natural gas	Z, D, X, E, N
Fuel Consumption - City (L/100km)	Quantitative	The estimated fuel consumption rate for city driving conditions, measured in liters per 100 kilometers (L/100km).	Ranges from 4.2 to 30.6
Fuel Consumption - Highway (L/100km)	Quantitative	The estimated fuel consumption rate for highway driving conditions, measured in liters per 100 kilometers (L/100km).	Ranges from 4.0 to 20.6
Fuel Consumption - Combined (L/100km)	Quantitative	The estimated average fuel consumption rate for combined city and highway driving conditions, measured in liters per 100 kilometers (L/100km).	Ranges from 4.1 to 26.1
Fuel Consumption - Combined (mpg)	Quantitative	The estimated average fuel consumption rate for combined city and highway driving conditions, measured in miles per gallon (mpg).	Ranges from 11 to 69
CO2 Emissions (g/km)	Response - Quantitative	The estimated carbon dioxide emissions produced by the vehicle, measured in grams per kilometer (g/km)	Ranges from 96 to 522

Table 5: Summary Table for Final MLR Model

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	1.292986	0.000228	5,677.907281	0.000000	***
Class_COMPACT	-0.000113	0.000024	-4.701051	0.000003	***
Class_MID_SIZE	-0.000049	0.000023	-2.153102	0.031357	*
Class_SUV_STANDARD	0.000130	0.000030	4.378779	0.000012	***
Class_VAN_PASSENGER	0.000235	0.000087	2.696104	0.007039	**
Class_PICKUPTRUCK_STANDARD	0.000233	0.000036	6.455800	0.000000	***
Class_MINIVAN	0.000146	0.000075	1.944436	0.051898	.
Class_PICKUPTRUCK_SMALL	0.000239	0.000056	4.295783	0.000018	***
Cylinders_6	0.000134	0.000024	5.674112	0.000000	***
Cylinders_12	0.000652	0.000068	9.648241	0.000000	***
Cylinders_8	0.000241	0.000035	6.803290	0.000000	***

	Estimate	Standard Error	t value	Pr(>|t|)	
Cylinders_10	0.000444	0.000105	4.244678	0.000022	***
Cylinders_3	0.000124	0.000069	1.786310	0.074108	.
Cylinders_16	0.002075	0.000545	3.808418	0.000141	***
Transmission_AS5	-0.000361	0.000133	-2.718175	0.006586	**
Transmission_A6	-0.000101	0.000028	-3.593887	0.000329	***
Transmission_AM7	0.000104	0.000035	2.945854	0.003235	**
Transmission_AS8	0.000147	0.000024	6.139541	0.000000	***
Transmission_A4	-0.000436	0.000088	-4.979031	0.000001	***
Transmission_M5	-0.000178	0.000049	-3.625776	0.000291	***
Transmission_AV	-0.000159	0.000043	-3.663514	0.000251	***
Transmission_AS7	-0.000030	0.000039	-0.748409	0.454248	
Transmission_A9	0.000191	0.000038	5.043928	0.000000	***
Transmission_AS9	0.000161	0.000073	2.202770	0.027655	*
Transmission_AV6	-0.000219	0.000062	-3.510783	0.000451	***
Transmission_AM5	-0.000873	0.000320	-2.726495	0.006423	**
Transmission_AM8	0.000173	0.000085	2.028155	0.042596	*
Transmission_AM9	0.001155	0.000541	2.133718	0.032913	*
Transmission_AS10	0.000183	0.000053	3.480362	0.000505	***
FuelType_Premium_Gas	-0.000074	0.000021	-3.606994	0.000313	***
FuelType_Diesel	0.004809	0.000053	90.727014	0.000000	***
FuelType_Ethanol	-0.012270	0.000043	-285.856616	0.000000	***
log(Fuel_Efficiency_Combined_MPG)	-0.034536	0.000066	-524.382402	0.000000	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 0.0005409 on 5123 degrees of freedom

Multiple R-squared: 0.9958, Adjusted R-squared: 0.9958

F-statistic: 3.814e+04 on 5123 and 32 DF, p-value: 0.0000