ANLY-511 Final Project Report

ANLY-511-04 Group 3

12/12/2022

Analyzing Relationships Between Car Design and Fuel Emissions

Authors: • Mia Mayerhofer • Matt Moriarty • Natalie Smith • Madelyne Ventura • Linlin Wang

1 Introduction

As the effects of climate change become more and more apparent and dire, it is important to analyze one of the top areas contributing to the greenhouse gas emissions polluting our atmosphere. According to the United States Environmental Protection Agency (EPA), transportation, as a whole, contributed to 27% of greenhouse gas emissions in the United States in 2020 (see Figure 1). Despite the more widespread knowledge and use of electric vehicles, a surprising less than 1% of cars in the United States are electric according to Feilding Cage from Reuters Graphics. This is due to a variety of reasons some of which include costs, accessibility, hesitancy to new technologies, etc. Thus, the goal of this report will be to analyze which design aspects of non-electric (fuel-based) cars contribute to CO_2 emissions. The following research questions will be answered and analyzed further:

1. *Can we predict emissions based on other aspects of cars (weight, transmission types, etc)?*
2. *Is there a difference in the average CO_2 emissions depending on the different types of gasoline? Which type of gasoline is better for the environment?*
3. *Is there a difference in the average CO_2 emissions depending on the different car manufacturers? Which car manufacturer is better for the environment?*
4. *Is there a relationship between the brand of a car and fuel emissions? What about the relationship between sedan/SUV and fuel emissions?*
5. *How do car features (e.g., weight, transmission type) vary across the car manufacturers?*
   After the technical analyses, the final goal of the paper will be to describe the characteristics an environmentally-conscious consumer should be aware of when purchasing a non-electric vehicle in hopes to provide a more practical application of the results achieved through the report's statistical analysis.

After the technical analyses, the final goal of the paper will be to describe the characteristics an environmentally-conscious consumer should be aware of when purchasing a non-electric vehicle in hopes to provide a more practical application of the results achieved through the report's statistical analysis.

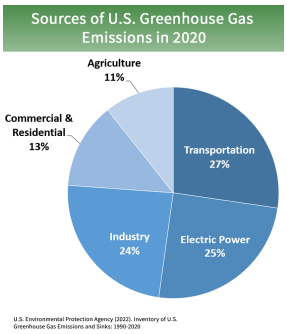


Figure 1

II. Data

Code Reference: Appendix A

In this section, we describe our data and explain the steps that we took in order to ensure that our data was prepared for use in our analyses.

Data Collection

The data used in this report contains various fuel economy metrics over a wide range of vehicles and is collected from five excel spreadsheets published annually by the EPA. The EPA collected this data from two sources: the EPA National Vehicle and Fuel Emissions Laboratory and individual data submissions from various vehicle manufacturers. The data is then combined for each year and sent to the Department of Energy (DOE), the Department of Transportation (DOT), and the Internal Revenue Service (IRS). For the initial data gathering, we chose the fuel economy reports for the years between 2018 and 2022, a five year span in total.

Data Cleaning

After collecting our data, we recognized that there were some cleaning tasks to perform before proceeding with our analyses. In this section, we outline the steps that we took to clean the data in order to transition it to a state that suits our analyses below.

Binding Datasets Together

The first step that we took in our data cleaning process was to bind our individual datasets together into one. Fortunately, the datasets that we gathered between the years 2018 and 2022 all contained the same variables, just for different individual observations. As a result, we were easily able to bind all five datasets together into one single dataset, spanning from 2018 to 2022. Note that the shape of each individual dataset was approximately 4,500 rows by 67 columns, giving us a combined dataset of size 22,616 rows by 67 columns.

Removing Unnecessary Variables

The next step that we took in order to clean our dataset was removing variables that we felt were unnecessary for our analyses. These include variables such as the unique vehicle ID, the transmission overdrive code, and the vehicle configuration number, among others. In total, there were 21 variables that we removed in this process, leaving us with a dataset of size 22,616 rows by 46 columns.

Partitioning the Dataset

An additional step that we decided to take when cleaning our dataset was to partition the dataset into two distinct groups - one for electric vehicles and one for non-electric vehicles. We felt that this split was necessary in order to extract more specific information regarding non-electric vehicles, which comprised the vast majority of our data. Additionally, we felt that this split better-allowed us to identify trends that belong to one group or the other, but not both groups. For instance, we found that the CO2 emissions for vehicles often had missing values in our dataset, but revealed themselves to be missing only for electric vehicles. The partitioning of our dataset allowed us to uncover these relationships that we otherwise may have missed with our initial, single dataset. In any case, we referred to the `Test Fuel Type Description` when doing so,

incorporating only those with `Electricity` or `Hydrogen 5` fuel types in our electric vehicles dataset. As a result, our dataset of shape 22,616 rows by 46 columns (`22,616, 46`) was partitioned into an electric vehicle dataset of shape 878 rows by 46 columns and a non-electric vehicle dataset of shape 21,738 rows by 46 columns.

Address Missing Values

As arguably the most important step of our data cleaning process, we addressed the missing values that were present in our two partitioned datasets. Note that the initial proportion of missing values across both datasets was approximately 15.64% and, as a result of our cleaning, we were able to drastically reduce this proportion to approximately 0.15%, more than a 99% reduction in missing values. In this subsection, we would like to point out the most important cleaning steps that we took here to address missing values.

The first important step that we took was removing any column that had more than 10% of its values missing. This 10% threshold is rather arbitrary, but allows us to refrain from making any bold decisions when replacing missing values with legitimate ones. Especially given the size of our dataset, we felt that replacing several thousands of missing values would have a noticeable effect on the distributions of the variables experiencing this adjustment. Applying this technique independently to each dataset, we found that we were able to remove 17 variables from the electric vehicle dataset and 9 variables from the non-electric vehicle dataset. Note that CO2 emissions and fuel economy had more than 10% of their values missing in the electric vehicle dataset, but we opted to keep them due to their relevance in our analysis.

When addressing the remaining columns containing missing values, all of which missing less than 10% of their values, we opted for a few techniques. Often, we visualized the distribution of values in order to gauge whether replacing missing values with the mean or median of existing values was sufficient. In most cases, we found that the distribution of existing values was not very symmetrical, and thus we replaced missing values with the median existing value of that variable. In other cases, such as those involving the `DT-Absolute` , `DT-Energy` , and `DT-Inertia` ratings, we found that replacing missing values with the mean existing value was sufficient, due to the symmetric distributions of those variables.

Finally, one interesting technique that we employed when replacing missing values presented itself with the `Number of Cylinders and Rotors` variable. We first noted that this variable almost exclusively took on even integer values, such as 2, 4, 6, and 8. We also noticed that this variable was very related to the horsepower of the vehicle. As a result, we opted to perform a rough linear regression, using horsepower as the predictor variable and the number of cylinders and rotors as the response variable, in order to estimate the number of cylinders and rotors for vehicles missing that value. Note that there were no missing values regarding the horsepower of a vehicle. In this case, we created a rough linear association between the number of cylinders and rotors of a vehicle and that vehicle's horsepower in order to estimate the number of cylinders and rotors of the vehicle, rounding to the nearest even integer. The scatterplot that visualizes this relationship can be found in Figure 2 below.

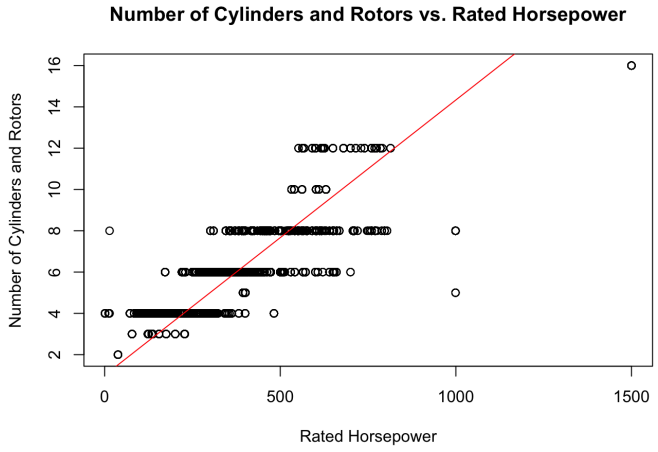


Figure 2

III. Exploratory Data Analysis (EDA)

Code Reference: All Appendices

Exploratory Data Analysis is an important first step when conducting any type of statistical analysis as it allows one to gain familiarity with the data and its features. This section contains a collection of the EDA performed for each of the methods discussed below.

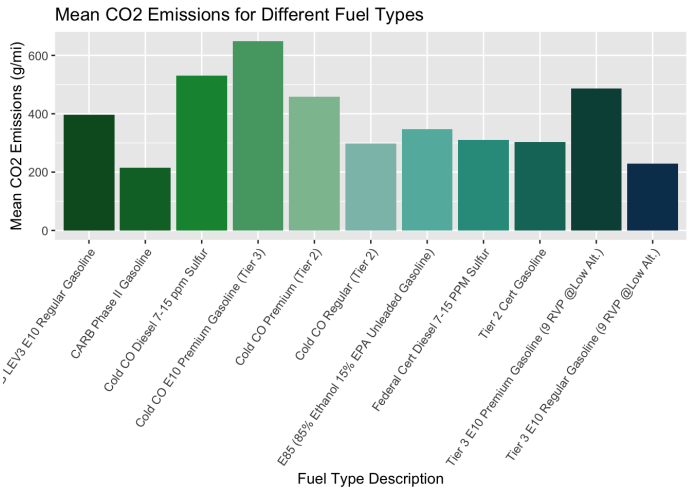


Figure 3.1

Fuel Type	Mean CO2 Emissions
Cold CO E10 Premium Gasoline (Tier 3)	648.4783
Cold CO Diesel 7-15 ppm Sulfur	530.4942

Fuel Type	Mean CO2 Emissions
Tier 3 E10 Premium Gasoline (9 RVP @Low Alt.)	486.2701
Cold CO Premium (Tier 2)	457.6009
CARB LEV3 E10 Regular Gasoline	397.1994
E85 (85% Ethanol 15% EPA Unleaded Gasoline)	346.8022
Federal Cert Diesel 7-15 PPM Sulfur	310.3833
Tier 2 Cert Gasoline	303.3148
Cold CO Regular (Tier 2)	297.7858
Tier 3 E10 Regular Gasoline (9 RVP @Low Alt.)	229.1300
CARB Phase II Gasoline	215.4994

Figure 3.2

In Figure 3.1, the bar plot of the mean carbon dioxide emissions for the different fuel types highlights which fuel types tend to produce the most and the least emissions. Figure 3.2 provides the table with the exact values plotted in the barplot in Figure 2. The Cold CO E10 Premium Gasoline (Tier 3) had the highest mean carbon dioxide emissions at around 648 g/mi while the CARB Phase II Gasoline had the lowest mean carbon dioxide emissions at around 215 g/mi over the five year time span. It is important to note, however, that the fuel types showing the lowest mean carbon dioxide emissions are also not as common in the data set while the fuel types showing higher mean carbon dioxide emissions have significantly more observations. This may be seen in the frequency table below (Figure 3.3).

Fuel Type	Frequency
Tier 2 Cert Gasoline	19235
Federal Cert Diesel 7-15 PPM Sulfur	974
Cold CO Regular (Tier 2)	648
E85 (85% Ethanol 15% EPA Unleaded Gasoline)	446
Cold CO Premium (Tier 2)	372
Tier 3 E10 Premium Gasoline (9 RVP @Low Alt.)	26
Cold CO Diesel 7-15 ppm Sulfur	12
CARB Phase II Gasoline	10
CARB LEV3 E10 Regular Gasoline	6
Cold CO E10 Premium Gasoline (Tier 3)	6
Tier 3 E10 Regular Gasoline (9 RVP @Low Alt.)	3

Figure 3.3

In the boxplots shown in Figure 3.4, it is clear that some distributions of certain fuel types are significantly skewed with several outliers. It is important to be aware of these distributions for testing the assumptions of hypothesis tests later in the report.

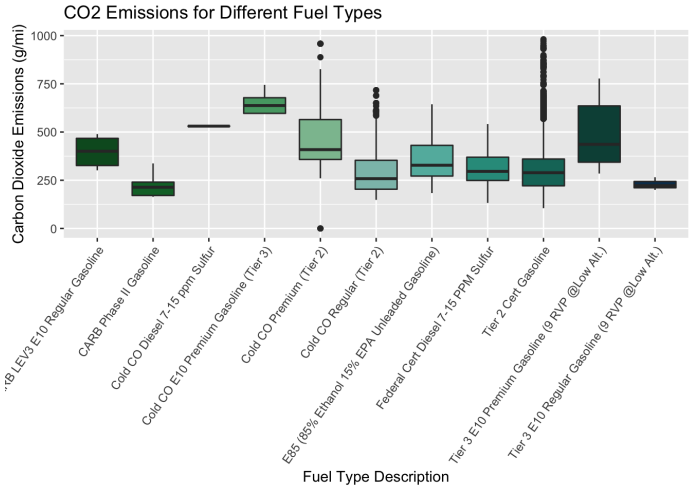


Figure 3.4

For the t-test section, specifically, only the two most common fuel types in the data set will be compared as a two sample test will be conducted. Thus, the boxplots shown in Figure 3.5 look closer at the distributions for these two most common fuel types: federal certified diesel 7-15 PPM sulfur and tier 2 certified gasoline.

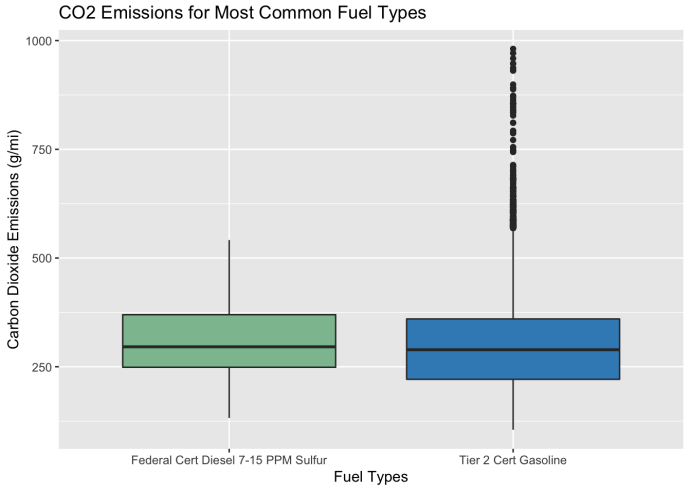


Figure 3.5

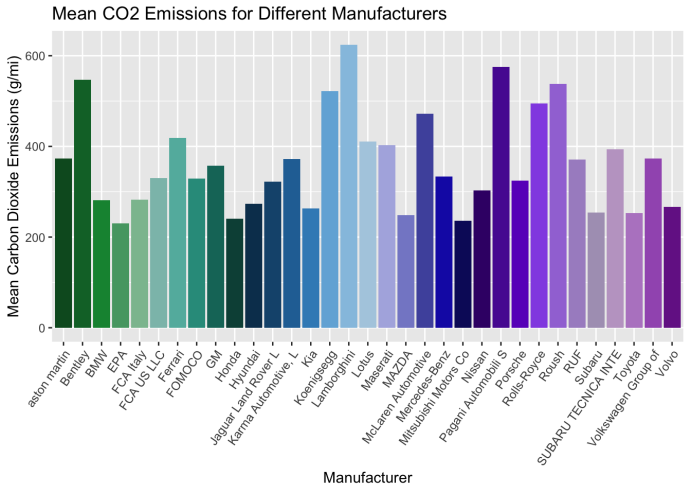


Figure 3.6

The bar plot shown in Figure 3.6 highlights which vehicle manufacturers have the highest and lowest mean carbon dioxide emissions. Figures 3.7 and 3.8 show the ten manufacturers with the highest mean carbon dioxide emissions and the ten manufacturers with the lowest mean carbon dioxide emissions during the five year time span.

Manufacturer	Mean CO2 Emissions	Manufacturer	Mean CO2 Emissions
Lamborghini	623.7461	EPA	229.7922
Pagani Automobili S	575.1250	Mitsubishi Motors Co	236.0538
Bentley	546.6000	Honda	239.9243
Roush	537.6841	MAZDA	248.8788
Koenigsegg	521.4150	Toyota	253.0496
Rolls-Royce	494.1484	Subaru	254.1140
McLaren Automotive	471.9094	Kia	263.6000
Ferrari	418.7053	Volvo	266.6802
Lotus	411.0043	Hyundai	273.2969
Maserati	402.5244	BMW	281.5546

Figure 3.7 *(left) and* ***Figure 3.8*** *(right)*

The three manufacturers with the highest mean carbon dioxide emission in the data set are Lamborghini, Pagani Automobili S, and Bentley. The three manufacturers with the lowest mean carbon dioxide emission are Honda, Mitsubishi Motors Co, and EPA. The box plots in Figure 3.9 clearly show that many of the manufacturers present in the data set have outlier vehicles with higher carbon dioxide emissions. Many of the distributions are also skewed, mostly to the right. It is also clear that some manufacturers' mean carbon dioxide emissions differ more significantly than others.

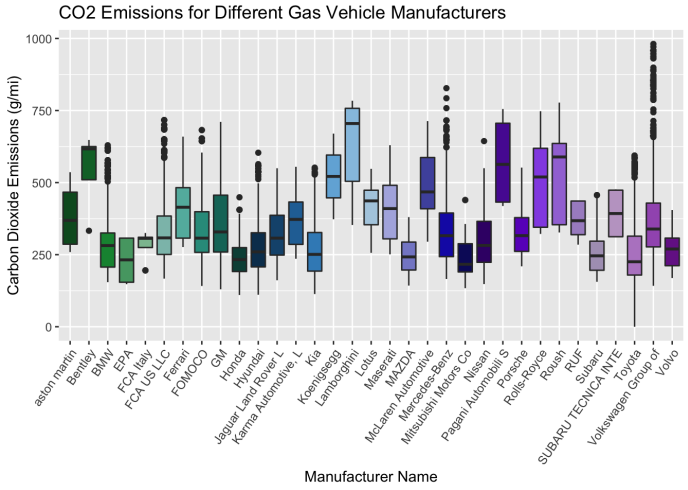


Figure 3.9

For the t-test section, specifically, only the two most common manufacturers in the data set will be compared as a two sample test will be conducted. Thus, the boxplots shown in Figure 3.10 below look closer at the distributions for these two most common manufacturers: General Motors and Toyota.

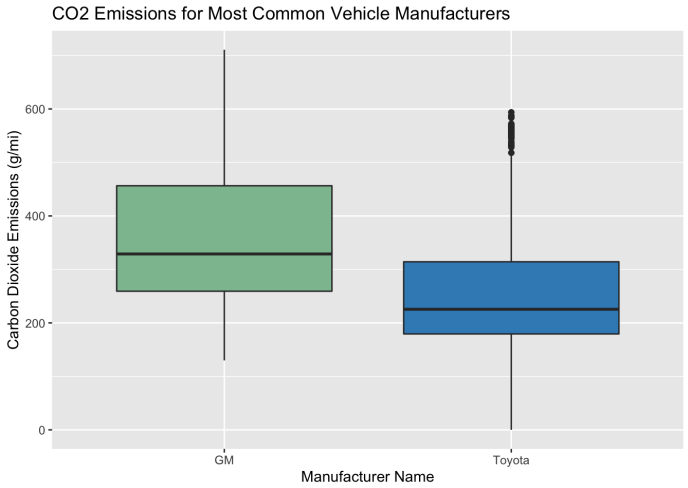


Figure 3.10

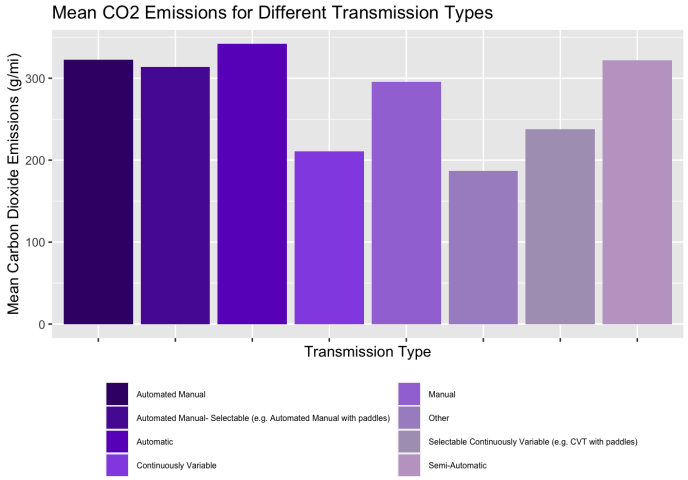


Figure 3.11

Figure 3.11 above shows a bar plot visualizing the mean carbon dioxide emissions over the five year time span for different vehicle transmission types such as manual, automatic, semi-automatic, etc. The frequency table in Figure 3.12 provides the exact values for the mean carbon dioxide emissions. It is clear from the bar plot that automatic vehicles and derivatives of automatic vehicles tend to have higher mean emissions than manual vehicles and derivatives of manual vehicles.

Transmission Type	Mean CO2 Emissions
Automatic	341.9557
Automated Manual	322.2405
Semi-Automatic	322.1310
Automated Manual- Selectable (e.g. Automated Manual with paddles)	313.5620
Manual	295.2084

Transmission Type	Mean CO2 Emissions
Selectable Continuously Variable (e.g. CVT with paddles)	237.6764
Continuously Variable	210.7348
Other	187.1900

Figure 3.12

In the boxplots shown in Figure 3.13, it is clear that some distributions of certain transmission types are significantly skewed with several outliers. It is important to be aware of these distributions for testing the assumptions of hypothesis tests later in the report. For the t-test section, specifically, only the general automatic and general manual data will be compared as a two sample test will be conducted. Thus, the boxplots shown in Figure 3.14 below look closer at the distributions for these two transmission types.

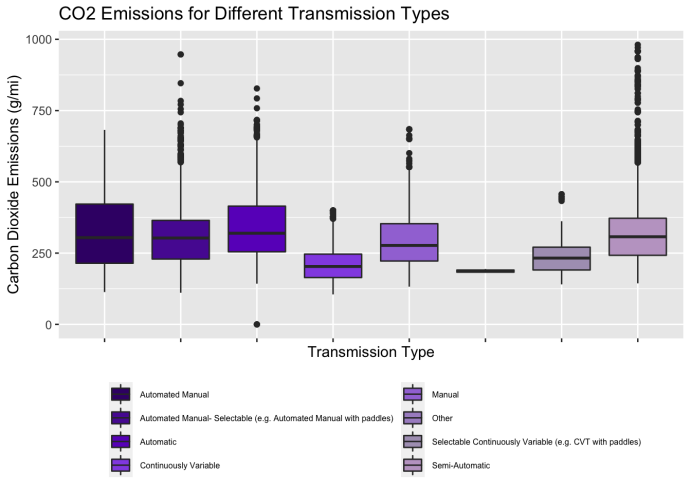


Figure 3.13

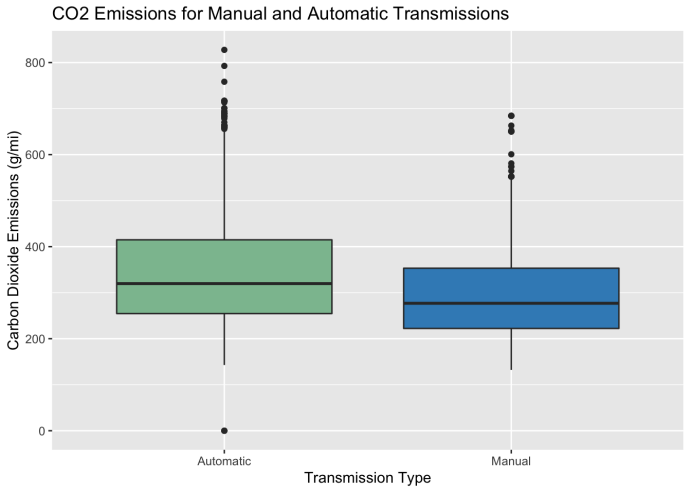


Figure 3.14

IV. Statistical Methods

In this section, we carry out the statistical methods that we used in our analysis. This is split into two sub-sections: Hypothesis Testing and Linear Regression. Our Hypothesis Testing methods incorporates several methods, including MANOVA, Chi-Squared Tests for Independence, and T-Tests.

Hypothesis Testing

Hypothesis Testing is a field within statistics that is used to determine whether a group of collected observations support a hypothesis. When a particular trend is observed, such as the average of one group of observations being higher than another other, hypothesis testing can be used to determine whether the trend is statistically significant. There are various forms of hypothesis testing that can be performed and the tests utilized within this project will be outlined throughout the report.

The process for conducting hypothesis testing is usually the same across the different types of tests. First, a null hypothesis must be defined within the context of the problem (e.g., the average values are the same for both groups of observations). Then, an alternative hypothesis is defined that is in contrast to the null hypothesis (e.g., the average values are greater for one group than the other group). The appropriate statistical test for the problem is then performed to generate a p-value, which is a measure of how likely the observations would occur if the null hypothesis were true. The p-value is then compared to a significance level α (e.g., 0.05), which is a measure of strength the evidence must have before the null hypothesis can be rejected. If the p-value is smaller than the significance level, the null hypothesis can be rejected in favor of the alternative hypothesis. The smaller the p-value, the higher chances of rejecting the null hypothesis.

MANOVA

Code Reference: Appendix B

Multivariate analysis of variance (MANOVA) is a generalized form of univariate analysis of variance (ANOVA) that includes at least two dependent variables to evaluate the mean differences on two or more dependent variables. Our purpose of using MANOVA is to analyze if there is statistically significant difference in CO_2, CO, and THC (total hydrocarbon), which are three main emissions in our dataset, between different

independent variables, such as the type of vehicles, vehicle manufacturers, or fuel types. Furthermore, MANOVA uses omnibus Wilk's Lambda, Roy's Largest Root, Hotelling-Lawley's test, or Pillai's Trace test, which is most robust to departures from assumptions. More importantly, Pillai's Trace has the highest statistical power.

- *Question 1: Is there a statistically significant difference in CO_2, CO, and THC between the type of vehicles?*

We firstly perform the exploratory data analysis. According to the boxplot, we notice there is a big difference between CO_2 and other two emissions. However, the mean difference in CO and THC is slightly unclear. In order to get more accurate results, we perform MANOVA test. The independent variable is the type of vehicles, and the dependent variables are CO_2, CO, and THC emissions. Our hypothesis is defined below:

- *H_0: There is no significant difference in CO_2, CO, and THC between the different types of vehicles.*
- *H_A: There is a significant difference in CO_2, CO, and THC between the different types of vehicles.*

We performed all four different MANOVA tests we mentioned above.In Table 4.1 below, we can see the p-values for all four different MANOVA tests are smaller than the significance level 0.05. So we reject the null hypothesis at 5% level of significance and conclude that there is significant difference in CO_2, CO, and THC between different types of vehicles.

Vehicle Type	Pillai's Trace	Hotelling-Lawley	Wilks	Roy
Test Statistic	0.04618	0.048126	0.95395	0.044949
P-Value	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16

Table 4.1

The next test we performed is univariate ANOVAs to find out how exactly each emissions are affected by the type of vehicles.

- *Question 2: Is there a statistically significant difference in CO_2, CO, and THC between vehicle manufacturers?*

The hypothesis we used for research question 2 are:

- *H_0: There is no significant difference in CO_2, CO, and THC between the different vehicle manufacturers.*
- *H_A: There is a significant difference in CO_2, CO, and THC between the different vehicle manufacturers.*

Performing Pillai's Trace, Hotelling-Lawley, Wilks, and Roy separately give the following results, shown in Table 4.2. We will reject the null hypothesis at 5% level of significance since the p-values for four tests are extremely small.

Vehicle Manufacturer	Pillai's Trace	Hotelling-Lawley	Wilks	Roy
Test Statistic	0.22328	0.26919	0.78298	0.23569
P-Value	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16

Table 4.2

Doing univariate ANOVAs again is to evaluate does vehicle manufactures has significant effect on all three emissions.

- *Question 3: Is there a statistically significant difference in CO_2, CO, and THC between different fuel types?*

Our last research question for MANOVA is focusing on the independent variable fuel type. The null hypothesis and alternative hypothesis are defined below:

- *H_0: There is no significant difference in CO_2, CO, and THC between fuel type.*
- *H_A: There is a significant difference in CO_2, CO, and THC between fuel type.*

According to the MANOVA results we performed, the p-values are significant since all of them are smaller than significance level. We can reject the null hypothesis, concluding that there is a significant difference in CO_2, CO, and THC between fuel type as well. These results can be seen below, in Table 4.3

Fuel Type	Pillai's Trace	Hotelling-Lawley	Wilks	Roy
Test Statistic	0.52439	0.9944	0.49279	0.95805
P-Value	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16

Table 4.3

Chi-Squared Test of Independence

Code Reference: Appendix C

One of the hypothesis testing methods performed throughout this analysis was the Chi-squared test of independence. The Chi-squared test of independence compares two variables and tests whether there is a relationship between them. The hypotheses for this test are defined below, where H_0A represents the alternative hypothesis.

- *H_0: There is no relationship between the two variables; They are independent*
- *H_A: There is a relationship between the two variables; They are dependent*

The test statistic for this hypothesis is the Chi (X^2) test statistic and it is computed using a contingency table, which details the frequency distribution for both kinds of variables in the test. The test statistic is computed using the observed values from the contingency tables and the expected frequency values.

When assessing the data for this analysis, the relationships between various car characteristics and fuel emissions level were analyzed. Since Chi-square testing requires the categorical variables only, a new categorical variable for fuel emissions level was created. Observations with CO_2 g/mi values lower than 250 were labeled as Low, greater than 250 but less than 500 were labeled as Medium, and greater than 500 were labeled as High. These fuel emission levels can now be used to conduct Chi-squared testing.

- *Question 1: Is there a relationship between car manufacturers and fuel emissions levels?*

When performing exploratory data analysis, there appeared to be car manufacturers that had only Low and Medium fuel emission category cars, such as Honda. Other car manufacturers, such as General Motors, had a large amount of High fuel emission category cars. The Chi-squared test for independence was utilized to test whether there was a relationship between car manufacturers and fuel emissions. The null and alternative hypotheses are defined below.

- *H_0: There is no relationship between car manufacturer and fuel emission level; They are independent*
- *H_A: There is a relationship between car manufacturer and fuel emission level; They are dependent*

The contingency table for this hypothesis shows categories with values close to zero. Table 4.4 shows the contingency values for six car manufacturers out of over 30 manufacturers as an example contingency table. An underlying assumption of Chi-squared testing is that all of the expected values are at least five. When low expected values occur, the Yates Continuity Correction and Fisher's Exact Test can be applied. Yates Continuity Correction is applied to the calculation of the Chi statistic and is used to compensate for the deviations from the theoretical probability (Giannini, 2005). Fisher's Exact Test, on the other hand, is used when one or more of the cell counts in a contingency table is less than 5 and generally is better suited when dealing with small cell counts (Leon, 1998). Therefore, when conducting the hypothesis test for this question, both the Yates Continuity Correction and Fisher's Exact Test were used.

Car Manufacturer vs. Fuel Emission Level	Low	Medium	High

Car Manufacturer vs. Fuel Emission Level	Low	Medium	High
Aston Martin	0	42	8
General Motors	569	1668	387
Honda	878	567	0
Hyundai	548	661	15
Ferrari	0	211	61
Toyota	1517	864	91

Table 4.4

- *Question 2: Is there a relationship between drive system and fuel emission level?*

The next question analyzed was whether there was a relationship between drive system type and the fuel emission level. When analyzing the drive system category against the fuel emission level, 2-Wheel Front systems appeared to have Low and Medium fuel emission categories only while 2-Wheel Rear systems appeared to have all three fuel emission categories. To test the relationship using Chi-squared testing, the following null and alternative hypotheses were used:

- H_0*: There is no relationship between drive system and fuel emission level; They are independent*
- H_A*: There is a relationship between drive system and fuel emission level; They are dependent*

The full contingency table for drive system and fuel emission level is displayed in Table 4.5. The table also contains instances where the cell values are smaller than five and thus the Yates Continuity Correction and Fisher's Exact Test were implemented.

Drive System vs. Fuel Emission Level	Low	Medium	High
Two-Wheel Drive, Front	5332	3830	7
Two-Wheel Drive, Rear	1603	6136	866
Four-Wheel Drive	78	536	197
All-Wheel Drive	643	2085	350
Part-Time Four-Wheel Drive	2	81	1

Table 4.5

- *Question 3: Is there a relationship between transmission type and fuel emission level?*

The last question analyzed for Chi-squared testing was whether there was a relationship between transmission type and fuel emission level. Exploratory analysis of the data appeared to show automatic and semi-automatic cars held the highest amount of high emissions cars compared to manual and variable cars. To test the relationship using Chi-squared testing, the following null and alternative hypotheses were used:

- H_0*: There is no relationship between transmission type and fuel emission level; They are independent*
- H_A*: There is a relationship between transmission type and fuel emission level; They are dependent*

The full contingency table for drive system and fuel emission level is displayed in Table 4.6. The table also contains instances where the cell values are smaller than five and thus the Yates Continuity Correction and Fisher's Exact Test were implemented as well.

Transmission System vs. Fuel Emission Level	Low	Medium	High
Two-Wheel Drive, Front	255	430	79
Two-Wheel Drive, Rear	553	1115	151
Four-Wheel Drive	1411	4081	603
All-Wheel Drive	2168	671	0
Part-Time Four-Wheel Drive	634	932	72
Four-Wheel Drive	4	0	0
All-Wheel Drive	454	308	0
Part-Time Four-Wheel Drive	2179	5112	516

Table 4.6

T-Tests

Code Reference: Appendix D

T-tests are a type of hypothesis testing that uses a t-distribution when calculating probabilities in hopes to compare two population means. First, a null hypothesis and an alternative hypothesis are defined, H_0 and H_A. The null hypothesis is typically a statement in the following form: there is no significant difference in the two sample means. The alternative hypothesis is typically a statement in the following form: the mean of one sample is greater than/less than/or different from the mean of the other sample. Next, a t-statistic is calculated from the samples' statistics: the sample means, standard deviations, and sizes. After the t-statistic is calculated, the p-value is computed based on the area below the t-distribution to the left or right of the calculated t-statistic. The p-value is then compared to a chosen significance level: 0.05, 0.01, and 0.001 are common chosen significance values. If the p-value is less than the significance level, the null hypothesis is rejected. On the other hand, if the p-value is greater than the significance value, we fail to reject the null hypothesis. Additionally, a confidence interval is calculated for the difference between the means from the sample statistics.

For two sample t-tests, specifically, the key assumptions are that the variables are normally distributed and the two samples are random and independent of one another. If the normality assumption does not hold, the Mann-Whitney U test is a better option for the hypothesis testing. The Mann-Whitney U test also follows the same steps as a t-test. Thus, we will check below if the variables are normally distributed and perform the correct test accordingly. The results of the above hypothesis testing could potentially yield important insights into which manufacturers, fuel types, and transmission types produce less carbon dioxide emissions and if this is statistically significant. The purpose of the t-tests and/or Mann Whitney U tests in this report will be to answer the following questions:

- *Question 1: Is there a significant difference in the amount of carbon dioxide emissions between types of fuel, specifically between the two most common fuel types in the data set: Tier 2 Certified Gasoline and Federal Certified Diesel 7-15 PPM Sulfur?*

From the exploratory data analysis, specifically the boxplots seen in figure 6, it appears that the means are relatively similar for these two fuel types; however, Federal Certified Diesel 7-15 PPM Sulfur appears to be slightly greater. The test results will determine whether or not the difference is statistically significant. The following null and alternative hypotheses are defined:

- H_0*: The mean carbon dioxide emissions is the same for Tier 2 Cert Gasoline and Federal Cert Diesel 7-15 PPM Sulfur.*

- *H_A: The mean carbon dioxide emission is greater for Federal Cert Diesel 7-15 PPM Sulfur than Tier 2 Cert Gasoline.*

The chosen significance level is 1%. The main assumption that must be verified is the normality of the samples.

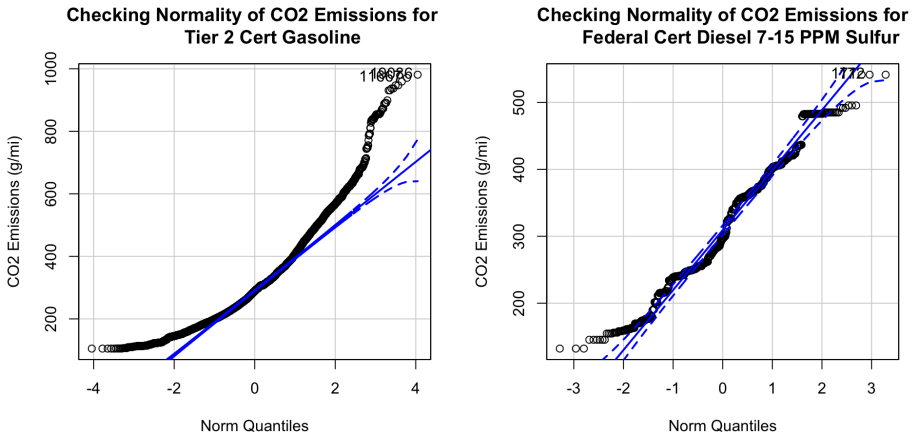


Figure 4.7

It is clear from the two QQ-plots in Figure 4.7 above that the samples do not pass the normality assumption. A shapiro test for normality verifies this result with a significantly small p-value. One method of normalizing the data is a logarithmic transformation; however, this did not improve the normality unfortunately. Thus, a Mann Whitney U test will be performed in place of a standard two sample t-test.

- *Question 2: Is there a significant difference in the amount of carbon dioxide emissions between vehicle manufacturers, specifically between the two most common vehicle manufacturers in the data set: General Motors and Toyota?*

From the exploratory data analysis, specifically the boxplots seen in Figure 3.10, it appears that the mean carbon dioxide emission is greater for GM vehicles than Toyota vehicles. The test results will determine whether or not the difference is statistically significant. The following null and alternative hypotheses are defined:

- *H_0: The mean carbon dioxide emissions is the same for GM and Toyota gasoline vehicles.*
- *H_A: The mean carbon dioxide emission is greater for GM gasoline vehicles than Toyota vehicles.*

The chosen significance level is 1%. The main assumption that must be verified is the normality of the samples.

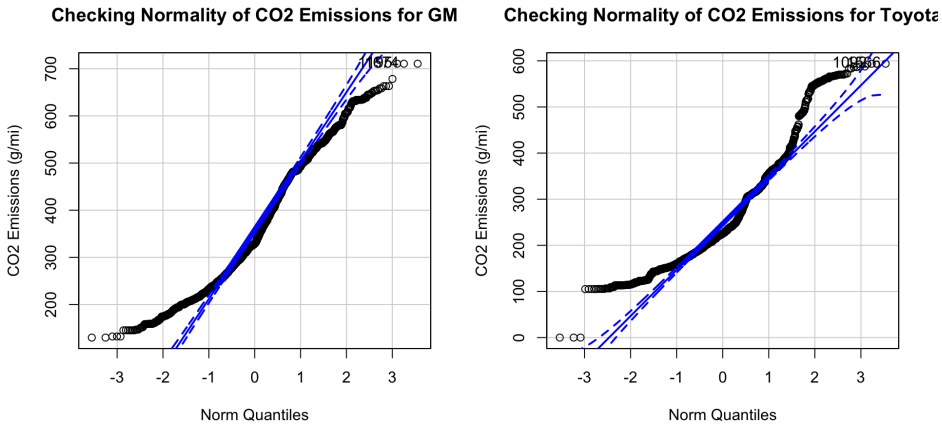


Figure 4.8

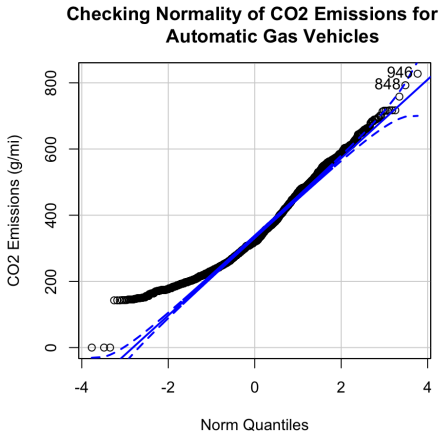
Unfortunately, the same result is true for question 2: the samples are not normally distributed. A logarithmic transformation did not help to normalize both of the samples. Thus, another Mann-Whitney U test will be performed in place of a t-test to answer the question posed above. QQ-plots expressing this relationship can be seen above in Figure 4.8.

- *Question 3: Is there a significant difference in the amount of carbon dioxide emissions between vehicle transmission types, specifically between manual and automatic transmission vehicles?*

From the exploratory data analysis, specifically the boxplots seen in Figure 3.14, it appears that the mean emissions for automatic vehicles is greater than for manual vehicles. The test results will determine whether or not the difference is statistically significant. The following null and alternative hypotheses are defined:

- *H_0: The mean carbon dioxide emissions is the same for manual and automatic gasoline vehicles.*
- *H_A: The mean carbon dioxide emission is greater for automatic gasoline vehicles than manual gasoline vehicles.*

The chosen significance level is 1%. The main assumption that must be verified is the normality of the samples. As with the other two tests, the normality assumption does not pass here as seen below in Figure 4.9; additionally, a logarithmic transformation does not improve the normality. Thus, a Mann-Whitney U test will be used again as a replacement for the t-test.



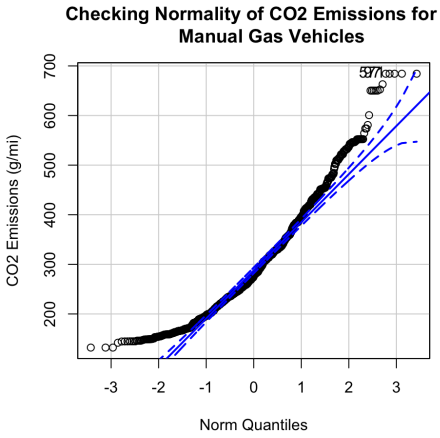


Figure 4.9

Linear Regression

Code Reference: Appendix E

In short, linear regression is a type of linear model that involves using a set of independent variables, or predictors, represented by Xi to predict a dependent variable, or response, represented by Y. For this project, multiple linear regression, involving the use of multiple independent variables to predict a response, will be used to predict a car's CO_2 emissions. Equations for multiple linear regression models take the following form, where ϵ represents the error present in the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Before delving into the research questions for this section, it is important to understand the assumptions underlying linear regression models, which are:

1. Individual observations are independent from each other
2. A linear relationship exists between the independent predictor variables X_i and the dependent response variable Y
3. Homoscedasticity, or homogeneity of variance
4. The residuals of the model are normally distributed

The goal of using multiple linear regression in this project is to answer the two following research questions:

1. Can the elements of a car's design be used to predict its CO_2 output?
2. Which elements of a car's design are best at predicting CO_2 output?

Due to the fact that electric cars do not give off emissions, only fuel-based cars will be considered for the multiple linear regression model. Because the fuel-based dataset has been cleaned already, as discussed earlier in this report, the only transformation required is to ensure the year the car was made is considered as categorical rather than numerical in the model. Once completed, the dataset was checked for any missing values before proceeding to modeling. This process highlighted that two predictors, the abbreviation and description for "aftertreatment device", which is a system that reduces harmful exhaust within the engine, had a number of missing values. This was mitigated by removing these 35 rows from the dataset. Finally, the data was split into training and testing sets, with 80% of the data being included in the training set and 20% of the data being included in the test set.

- *Emissions Model 1: Full model with (nearly) all columns as predictors*

There are some initial unnecessary variables that were identified as either repetitive (an abbreviation of another column) or completely irrelevant to the regression model to predict CO2 emissions (index for the dataset and whether or not the car was a police vehicle or not.) Additionally, the two variables of Vehicle.Manufacturer.Name and Represented.Test.Veh.Model, which detail the make and model of each respective car, need to be left out. The reason for this is that the multiple linear regression model is unable to predict emissions for makes and models of cars that appear in the testing set but *not* in the training set. Other than these variables, all other terms will be used to predict a full model and will be tweaked based on results for additional models.

The results of the first multiple linear regression model are pictured below, in Figure 4.10:

```
Residual standard error: 42.4 on 17186 degrees of freedom
Multiple R-squared:  0.856,  Adjusted R-squared:  0.8555
F-statistic:  2003 on 51 and 17186 DF,  p-value: < 2.2e-16
```

Figure 4.10

Interestingly enough, aside from a few of the initial predictors, almost all of the predictors in the model appear to be significant in predicting emissions for a car. However, it is very possible there may be multicollinearity in the current model, which occurs when at least two of the predictor variables in a model are highly correlated and result in redundancy, skewing the results and making the model unstable. To detect the presence of multicollinearity, the variance inflation factor (VIF) score can be computed. Typically, predictors that exceed 5.0 can be considered to be highly correlated with other predictors. Since there are already many significant predictors, we will be extra conservative and remove the predictors of DT Inertia Work Ratio Rating and DT Absolute Speed Change Rating from the model, which have VIF scores around 4.0. Combining this with the predictors that did not meet the 0.05% significance level, the predictors we will be removing to create a more "tuned" model to compare to the original are:

- DT Inertia Work Ratio Rating
- DT Absolute Speed Change Rating
- Transmission Lockup
- CO g/mi emissions

In terms of categorical variables, if at least one dummy variable for a categorical variable is significant, all will be kept at this stage of model tuning.

- *Emissions Model 2: Removing multicollinearity from model and initial insignificant terms*

The results of the second multiple linear regression model are pictured below, in Figure 4.11:

```
Residual standard error: 42.64 on 17190 degrees of freedom
Multiple R-squared:  0.8543, Adjusted R-squared:  0.8539
F-statistic:  2145 on 47 and 17190 DF,  p-value: < 2.2e-16
```

Figure 4.11

At this stage of model tuning, the last of the insignificant variables below the 0.05% significance level, as well as those categorical variables where less than half of the dummy variables are significant, will be removed. As a result, the variables Set.Coef.C..lbf.mph..2. (the measure of force, speed, and power required to operate a car divided by miles per hour) and Aftertreatment Device are removed for the final linear model.

- *Emissions Model 3: Removing all insignificant terms*

The results of the third multiple regression model are pictured below, in Figure 4.12:

```
Residual standard error: 42.71 on 17196 degrees of freedom
Multiple R-squared:  0.8538, Adjusted R-squared:  0.8534
F-statistic:  2449 on 41 and 17196 DF,  p-value: < 2.2e-16
```

Figure 4.12

Satisfied that this seems to be the best-performing model of the bunch when considering its reduced number of predictors, the next step is to check for any outliers or high leverage points present.

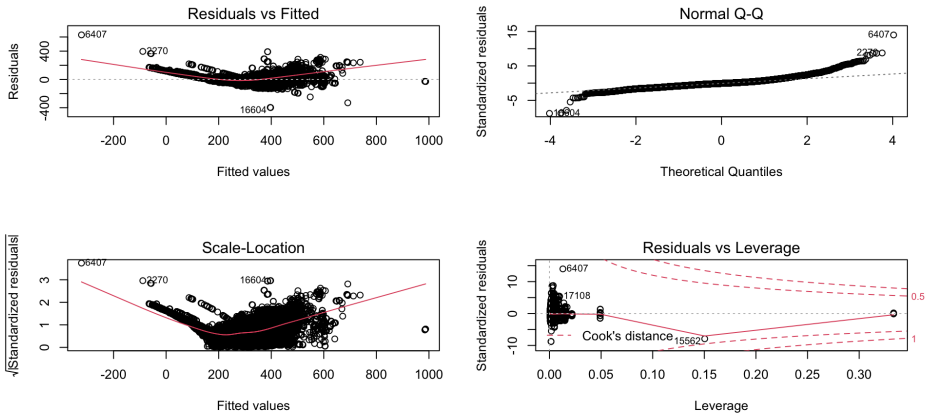


Figure 4.13

Looking at Figure 4.13, particularly the Residuals vs Fitted and the Scale-Location plots, it can be seen that the model appears to violate the assumption of linearity. Due to the parabola shape of the data, it is possible that a quadratic regression model, which finds the equation of the parabola that fits the data rather than the line, may be a better fit for this data.

- *Emissions Model 4: Quadratic Regression Model*

To see if quadratic regression could improve this model, a single squared regression term will be added to the predictor variables. Because the predictor with the highest influence on the model (or, largest F-statistic) is RND_ADJ_Fe, or miles per gallon, with an F-statistic of -174.25, a quadratic term for this predictor will be added to see if it improves the model.

The results of the quadratic regression model are pictured below, in Figure 4.14:

```
Residual standard error: 25.49 on 17195 degrees of freedom
Multiple R-squared:  0.9479, Adjusted R-squared:  0.9478
F-statistic:  7454 on 42 and 17195 DF,  p-value: < 2.2e-16
```

Figure 4.14

The final conclusions on the best regression model to predict CO_2 emissions as well as the most influential aspects of a car's design on emissions will be discussed in the Results section below.

V. Results

In this section, we report the results of our statistical methods from section IV.

MANOVA

According to the result table of univariate ANOVAs below, the p-values for response CO<sub2, CO, and THC based on all three independent variables are all extremely small, which indicates that vehicle type, vehicle manufacturer, and fuel type have statistically significant effects on CO_2, CO, and THC.

Independent Variable	Response CO2 P-Value	Response CO P-Value	Response THC P-Value	Significant?
Vehicle Type	< 2.2e-16	0.00166	3.751e-16	Yes
Vehicle Manufacturer	< 2.2e-16	7.695e-05	< 2.2e-16	Yes
Fuel Type	< 2.2e-16	2.379e-16	< 2.2e-16	Yes

Chi-Squared Test of Independence

The Chi-squared test for independence was conducted to test whether there was a relationship between fuel emission level and the following car features: car manufacturer, drive system type, and transmission type. Figure X below summarizes the results for the three hypothesis questions and shows the p-values for their respective statistical tests (Chi-squared test of independence with Yates Continuity Correction, Fisher's Exact Test).

Hypothesis	Chi-Squared P-Value	Fisher's P-Value	Reject Null Hypothesis?
Question 1: Is there a relationship between car manufacturer and fuel emission level?	< 2.2e-16	< 4.9e-4	Yes
Question 2: Is there a relationship between drive system and fuel emission level?	< 2.2e-16	< 4.9e-4	Yes
Question 3: Is there a relationship between transmission type and fuel emission level?	< 2.2e-16	< 4.9e-4	Yes

At a significance level of 0.05, all three sets of p-values are small enough to reject the null hypothesis in favor of the alternative hypothesis. Therefore, there is a relationship between each car feature mentioned above and the fuel emissions type. The car features and the fuel emissions type are dependent on each other.

T-Tests

Mann-Whitney U tests were conducted in place of t-tests due to the failure of the normality assumption in all three cases. Figure X below summarizes the results for the three hypothesis questions by showing the p-values, whether or not the null hypothesis was rejected, the 95% confidence interval, and the final conclusion. As a point of reference, the significance level of 1% was chosen to compare the p-value to.

Hypothesis	P-Value	95% Confidence Interval	Reject Null / Conclusion
Question 1: Is there a significant difference in the amount of CO2 emissions between the two most common fuel types: Tier 2 Cert Gasoline and Federal Cert Diesel 7-15 PPM Sulfur?	2.1e-07	[10.61379, 23.05800] g/mi	YES: Federal Cert Diesel 7-15 PPM Sulfur Emissions > Tier 2 Cert Gasoline Emissions
Question 2: Is there a significant difference in the amount of CO₂ emissions between the two most common manufacturers: GM and Toyota?	2.2e-16	[94.60287, 106.32853] g/mi	YES: GM Emissions > Toyota Emissions
Question 3: Is there a significant difference in the amount of CO₂ emissions between manual and automatic transmission vehicles?	2.2e-16	[38.75607, 49.62847] g/mi	YES: Automatic Transmission Emissions > Manual Transmissions Emissions

Each of the three tests resulted in small p-values below the 1% significance level; thus, the null hypothesis is rejected for each of the questions.

For question one, the conclusion is that the mean carbon dioxide emissions for Federal Certified Diesel 7-15 PPM Sulfur is significantly greater than the mean carbon dioxide emissions for Tier 2 Certified gasoline. With 95% confidence the Federal Certified Diesel 7-15 PPM Sulfur produces between 10.6 g/mi to 23.1 g/mi more carbon dioxide than Tier 2 Certified gasoline.

For question two, the conclusion is that the mean carbon dioxide emissions for GM gasoline vehicles is significantly greater than the mean carbon dioxide emissions for Toyota gasoline vehicles over the five year time span. With 95% confidence GM vehicles produce between 94.6 g/mi to 106.3 g/mi more carbon dioxide than Toyota vehicles on average.

For question three, the conclusion is that the mean carbon dioxide emissions for automatic gasoline vehicles is significantly greater than the mean carbon dioxide emissions for manual gasoline vehicles over the five year time span. With 95% confidence automatic vehicles produce between 34.8 g/mi to 49.6 g/mi more carbon dioxide than manual vehicles on average.

Linear Regression

The results for all four regression models tested to predict CO2 emissions are pictured below.

Model Name	RMSE	R2	F-Statistic	Adj R2	RSE
Model 1	67.23582	0.6917589	2229.906	0.8413123	44.44045
Model 2	67.98031	0.6864206	2435.042	0.8393518	44.71413
Model 3	67.65622	0.6878452	2818.738	0.8351897	45.28966
Model 4	362.84165	0.0780696	6224.317	0.9203404	31.48662

As can be seen in the model results, the quadratic regression model, Model 4, appears to be a huge improvement in every way on the best linear regression model, Model 3. However, the RMSE for Model 4 is extremely large, suggesting that this quadratic regression is badly overfitting, and therefore is not a good predictor of CO2 emissions. As a result, the best model is still the multiple linear regression model of Model 3.

The top three predictors that have the most influence on the predicting a car's CO2 emissions based on their respective t-values are:

1. Miles per gallon (RND_ADJ_FE), with a t-value of -174.25, suggesting that as the number of miles per gallon a car is able to achieve increases, its CO2 emissions decreases.
2. Total hydrocarbon emissions (THC..g.mi.), with a t-value of 35.12
3. Electric Dynamometer Coefficient/mph (Target.Coef.B..lbf.mph.), with a t-value of 21.90. This predictor is the measure of force, speed, and power required to operate the car being measured.

As the total hydrocarbon emissions and the electric dynamometer coefficient increase, the CO2 emissions produced by a car will also increase.

VI. Conclusion

The results highlighted in this report could provide helpful insights for those debating which vehicle to purchase, specifically if they do not have the ability to purchase an electric vehicle. Most importantly, vehicles with a higher fuel economy (miles per gallon) produce fewer emissions on average. Additionally, vehicles that accept fuel types such as Tier 2 Certified gasoline and Carbon Phase II also produce fewer carbon dioxide emissions on average compared to Federal Certified Diesel fuel for example. Surprisingly, manual vehicles produced significantly fewer mean carbon dioxide emissions over the five year time span compared to vehicles with automatic transmission. Finally, a clear relationship was found between the manufacturer and the mean carbon dioxide emissions; thus, some manufacturers that appear to be more environmentally-friendly include Mazda, Toyota, Honda, and Mitsubishi.

In conclusion, statistical analysis including hypothesis testing and regression analysis yields useful results in analyzing the relationships between various aspects of vehicle design and greenhouse gas emissions. Due to electric vehicles still being relatively inaccessible and expensive for the general public, it will be important in the coming years for vehicle manufacturers to continuously improve their designs for fuel-based vehicles to produce fewer emissions.

VII. References

- Cage, Feilding. “The Long Road to Electric Cars in the U.S.” Reuters, Thomson Reuters, https://www.reuters.com/graphics/AUTOS-ELECTRIC/USA/mopanyqxwva/ (https://www.reuters.com/graphics/AUTOS-ELECTRIC/USA/mopanyqxwva/).
- “Data on Cars Used for Testing Fuel Economy.” EPA, Environmental Protection Agency, https://www.epa.gov/compliance-and-fuel-economy-data/data-cars-used-testing-fuel-economy (https://www.epa.gov/compliance-and-fuel-economy-data/data-cars-used-testing-fuel-economy).
- “GGPLOT2 Barplots : QUICK START GUIDE - R Software and Data Visualization.” STHDA, http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization (http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization).
- “GGPLOT2 Box Plot : Quick Start Guide - R Software and Data Visualization.” STHDA, http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization (http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization).
- “GGPLOT2 Legend : Easy Steps to Change the Position and the Appearance of a Graph Legend in R Software.” STHDA, http://www.sthda.com/english/wiki/ggplot2-legend-easy-steps-to-change-the-position-and-the-appearance-of-a-graph-legend-in-r-software (http://www.sthda.com/english/wiki/ggplot2-legend-easy-steps-to-change-the-position-and-the-appearance-of-a-graph-legend-in-r-software).
- Giannini, Edward H. “Design, Measurement, and Analysis of Clinical Investigations.” Textbook of Pediatric Rheumatology, 2005, pp. 142–173., https://doi.org/10.1016/b978-1-4160-0246-8.50012-7 (https://doi.org/10.1016/b978-1-4160-0246-8.50012-7).
- Leon, A. C. (1998). Descriptive and inferential statistics. Comprehensive Clinical Psychology, 243–285. https://doi.org/10.1016/b0080-4270(73)00264-9 (https://doi.org/10.1016/b0080-4270(73)00264-9)

- “MANOVA Test in R: Multivariate Analysis of Variance”, http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance (http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance)
- “Plot a QQ Chart.” R, https://braverock.com/brian/R/PerformanceAnalytics/html/chart.QQPlot.html (https://braverock.com/brian/R/PerformanceAnalytics/html/chart.QQPlot.html).
- “RGB Color Codes Chart.” RGB Color Codes Chart ,https://www.rapidtables.com/web/color/RGB_Color.html (https://www.rapidtables.com/web/color/RGB_Color.html).
- “Sources of Greenhouse Gas Emissions.” EPA, Environmental Protection Agency, https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions# (https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions#):~:text=Human%20activities%20are%20responsible%20for,over%20the%20last%20150%20years.&text=The%20largest%20source%20of%20greenhouse,electricity%2

VIII. Appendix

In this section, we provide the code that we used to conduct our analysis. Each section above references a portion of this Appendix, forming a connection between the analysis and the code used to conduct it. The sections can be referenced as follows:

- Appendix A: Data Cleaning
- Appendix B: MANOVA
- Chi-Squared Tests for Independence
- T-Tests
- Linear Regression

In order to include our various analyses in this Appendix, we opt to append their individual PDF outputs below.

Data Cleaning

ANLY-511-04 Group 3

12/06/2022

```
# load libraries for this assignment
library(tidyverse)
library(ggplot2)
library(readxl)
library(knitr)
library(kableExtra)
```

Functions for Use

```
calculate_proportions <- function(input_df, threshold) {

  # create data frame for printing proportions of of NA values
  df <- data.frame(colnames(input_df), round(colMeans(is.na(input_df)), 5), row.names = NULL)
  colnames(df) <- c('Column Name', 'Proportion of Values NA')
  df <- df[order(df$`Proportion of Values NA`, decreasing = TRUE),]
  df_table <- df[df$`Proportion of Values NA` > threshold,]

  return (df_table)
}
```

```
print_table <- function(input_table) {
  knitr::kable(input_table, row.names = FALSE) %>% kable_styling(bootstrap_options = "striped", full_width = F, p
osition = "center")
}
```

```
print_shape <- function(input_df) {
  print(paste('Our dataset shape is now: (', nrow(input_df), ', ', ncol(input_df), ')', sep = ''))
}
```

Read in the Data

Let's first read in the data. We have five datasets to read in, all by the naming convention of `data/cardataXX.xlsx`, where `XX` $\in$ `{18, 19, 20, 21, 22}`. We can also observe the *shape* of each dataset, indicated with the following syntax: `(number of rows, number of columns)`.

```
# load in the data
cardata2018 <- read_excel('../data/cardata2018.xlsx')
cardata2019 <- read_excel('../data/cardata2019.xlsx')
cardata2020 <- read_excel('../data/cardata2020.xlsx')
cardata2021 <- read_excel('../data/cardata2021.xlsx')
cardata2022 <- read_excel('../data/cardata2022.xlsx')
```

Dataset Name	Shape
data/cardata2018.xlsx	(4727, 67)
data/cardata2019.xlsx	(4719, 67)
data/cardata2020.xlsx	(4450, 67)
data/cardata2021.xlsx	(4265, 67)
data/cardata2022.xlsx	(4455, 67)

The five datasets come from the same source, but we want to ensure that they are compatible with one-another. The best way to check this is to ensure that the columns are the same across all datasets. If they are, we can easily append the five datasets together, by row, in order to create one, larger dataset.

```
# extract column names for each dataset
colnames2018 <- colnames(cardata2018)
colnames2019 <- colnames(cardata2019)
colnames2020 <- colnames(cardata2020)
colnames2021 <- colnames(cardata2021)
colnames2022 <- colnames(cardata2022)

# ensure that the column names are all the same across all datasets
if ( mean(colnames2018 == colnames2019) == 1 &
     mean(colnames2018 == colnames2020) == 1 &
     mean(colnames2018 == colnames2021) == 1 &
     mean(colnames2018 == colnames2022) == 1 ) {
  print('The column names are the same and in the same order across all five datasets.')
} else {
  print('The column names are NOT the same and in the same order across all five datasets.')
}
```

```
## [1] "The column names are the same and in the same order across all five datasets."
```

Let's append these datasets together since we know now that they have the same structure and column names.

```
# bind datasets together by row
cardata <- rbind(cardata2018, cardata2019, cardata2020, cardata2021, cardata2022)
print_shape(cardata)
```

```
## [1] "Our dataset shape is now: (22616, 67)"
```

There are also some columns that we are not concerned with for the purpose of our analysis, so we will drop those before proceeding with the cleaning phase. These columns are:

- Test Vehicle ID
- Test Veh Configuration #
- Actual Tested Testgroup
- Engine Code
- Transmission Overdrive Code
- Transmission Overdrive Desc
- Shift Indicator Light Use Cd
- Shift Indicator Light Use Desc
- Test Number
- Test Originator
- Analytically Derived FE?
- ADFE Test Number
- Test Procedure Cd
- Test Procedure Description
- Test Fuel Type Cd
- Test Category
- FE_UNIT
- FE Bag 1
- FE Bag 2
- FE Bag 3
- FE Bag 4

```
# define columns to remove
remove <- c("Test Vehicle ID", "Test Veh Configuration #", "Actual Tested Testgroup",
            "Engine Code", "Transmission Overdrive Code", "Transmission Overdrive Desc",
            "Shift Indicator Light Use Cd", "Shift Indicator Light Use Desc", "Test Number",
            "Test Originator", "Analytically Derived FE?", "ADFE Test Number",
            "Test Procedure Cd", "Test Procedure Description", "Test Fuel Type Cd",
            "Test Category", "FE_UNIT", "FE Bag 1",
            "FE Bag 2", "FE Bag 3", "FE Bag 4")

# remove columns and report new shape
cardata <- cardata[, !( colnames(cardata) %in% remove)]
print_shape(cardata)
```

```
## [1] "Our dataset shape is now: (22616, 46)"
```

Split Data into Two

Since one large are of focus of ours is comparing electric vehicles to non-electric vehicles, we will split our dataset into two groups: one containing only electric vehicles and one containing only non-electric vehicles. This will allow us to make easy comparisons both between datasets and within datasets.

```
# look at the unique fuel types
unique(cardata$`Test Fuel Type Description`)
```

```
##  [1] "Tier 2 Cert Gasoline"
##  [2] "Cold CO Premium (Tier 2)"
##  [3] "Federal Cert Diesel 7-15 PPM Sulfur"
##  [4] "Electricity"
##  [5] "Cold CO Regular (Tier 2)"
##  [6] "E85 (85% Ethanol 15% EPA Unleaded Gasoline)"
##  [7] "Tier 3 E10 Regular Gasoline (9 RVP @Low Alt.)"
##  [8] "Hydrogen 5"
##  [9] "Cold CO E10 Premium Gasoline (Tier 3)"
## [10] "Tier 3 E10 Premium Gasoline (9 RVP @Low Alt.)"
## [11] "CARB Phase II Gasoline"
## [12] "Cold CO Diesel 7-15 ppm Sulfur"
## [13] "CARB LEV3 E10 Regular Gasoline"
```

Observing the unique fuel types above, we will split the dataset on cars with `Test Fuel Type Description` labeled as `Electricity` or `Hydrogen 5`, leaving the remaining fuel types for the other dataset.

```
# split the dataset in two
cardata_electric <- cardata[cardata$`Test Fuel Type Description` %in% c('Electricity', 'Hydrogen 5'),]
cardata_nonelectric <- cardata[!(cardata$`Test Fuel Type Description` %in% c('Electricity', 'Hydrogen 5')),]
```

Address Missing Values by Column

Now, with the columns that we do have, we can observe how many `NA` values exist in each column. We will set a threshold of **10%**; a column with more than 10 percent of its values being `NA` expresses a large number of missing values in our eyes and we will remove it from our analysis. Below, we can see all columns expressing more than 10% of their values as `NA`.

```
# print the proportions of NA values in each column using a threshold of 10%
electric_table <- calculate_proportions(cardata_electric, 0.10)
nonelectric_table <- calculate_proportions(cardata_nonelectric, 0.10)
print_table(electric_table)
```

Column Name	Proportion of Values NA
ADFE Total Road Load HP	1.00000
ADFE Equiv. Test Weight (lbs.)	1.00000
ADFE N/V Ratio	1.00000
THC (g/mi)	1.00000
CO2 (g/mi)	1.00000
PM (g/mi)	1.00000
CH4 (g/mi)	1.00000
N2O (g/mi)	1.00000
Averaging Group ID	0.96811
CO (g/mi)	0.93850
Averaging Weighting Factor	0.92711
NOx (g/mi)	0.90774

Column Name	Proportion of Values NA
# of Cylinders and Rotors	0.86105
Aftertreatment Device Cd	0.86105
Aftertreatment Device Desc	0.86105
DT-Inertia Work Ratio Rating	0.85991
DT-Absolute Speed Change Ratg	0.85991
DT-Energy Economy Rating	0.85991
RND_ADJ_FE	0.15034

```
print_table(nonelectric_table)
```

Column Name	Proportion of Values NA
Averaging Group ID	0.98114
Averaging Weighting Factor	0.97249
ADFE Total Road Load HP	0.88683
ADFE Equiv. Test Weight (lbs.)	0.88683
ADFE N/V Ratio	0.88683
PM (g/mi)	0.81990
N2O (g/mi)	0.47796
CH4 (g/mi)	0.15457
NOx (g/mi)	0.10907

We can then remove these columns from each respective dataset. Note that the `RND_ADJ_FE` and `CO2 (g/mi)` variables are ones that we'd like to work with, however, so we will keep those regardless of their missing value counts.

```
# remove high-probability NA columns from data frame
remove_electric <- electric_table$`Column Name`
remove_electric <- remove_electric[!(remove_electric %in% c('RND_ADJ_FE', 'CO2 (g/mi)'))]
remove_nonelectric <- nonelectric_table$`Column Name`

# remove columns and report new shape
cardata_electric <- cardata_electric[, !( colnames(cardata_electric) %in% remove_electric)]
cardata_nonelectric <- cardata_nonelectric[, !( colnames(cardata_nonelectric) %in% remove_nonelectric)]
print_shape(cardata_electric)
```

```
## [1] "Our dataset shape is now: (878, 29)"
```

```
print_shape(cardata_nonelectric)
```

```
## [1] "Our dataset shape is now: (21738, 37)"
```

Electric Vehicle Data

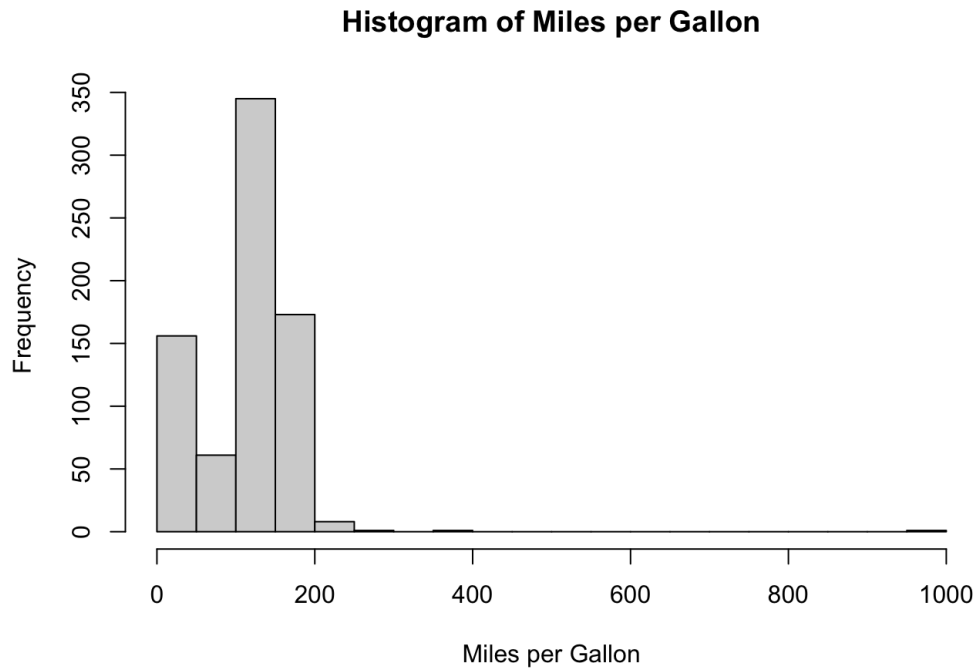
We still have columns expressing `NA` values, however. These columns, of course, have less than 10% of their values being `NA`, after the removal performed above. These columns can be seen below.

```
# print the proportions of NA values in each column using a threshold of 0%
electric_table <- calculate_proportions(cardata_electric, 0)
print_table(electric_table)
```

Column Name	Proportion of Values NA
CO2 (g/mi)	1.00000
RND_ADJ_FE	0.15034

We can see that the electric vehicle data only has two columns with missing values: `RND_ADJ_FE` and `CO2 (g/mi)`. We've decided to keep the `CO2 (g/mi)` in order to provide comparisons between the electric and non-electric vehicle data, but we want to address the specific missing values of the `RND_ADJ_FE` variable. This variable is an expression of miles per gallon. Let's take a look at a histogram of this variable.

```
# plot histogram of the electric car MPG values
hist(cardata_electric$RND_ADJ_FE, xlab = 'Miles per Gallon', main = 'Histogram of Miles per Gallon', breaks = 20)
```



We can see that there is at least one outlier on the high end of the distribution, expressing a value of 1000. We will replace these value(s) with missing values in order to obtain results that have more validity. We will then replace the missing values with the *median* value of the distribution, as this is a safer way to do so without making any bold assumptions.

```
# replace high outliers with missing value
cardata_electric$RND_ADJ_FE[cardata_electric$RND_ADJ_FE > 500] <- NA
# replace missing values with the median value
cardata_electric$RND_ADJ_FE[is.na(cardata_electric$RND_ADJ_FE)] <- median(cardata_electric$RND_ADJ_FE, na.rm = TR
UE)
```

We can now observe the columns with missing values in the electric vehicle dataset and notice that we have cleaned it to our liking. Note that the `CO2 (g/mi)` variable still remains and will likely be interpreted as **0** emission from these electric vehicles.

```
# print the proportions of NA values in each column using a threshold of 0%
electric_table <- calculate_proportions(cardata_electric, 0)
print_table(electric_table)
```

Column Name	Proportion of Values NA
CO2 (g/mi)	1

Non-Electric Vehicle Data

We still have columns expressing `NA` values for the non-electric vehicle dataset. These columns, of course, have less than 10% of their values being `NA`, after the removal performed above. These columns can be seen below.

```
# print the proportions of NA values in each column using a threshold of 0%
nonelectric_table <- calculate_proportions(cardata_nonelectric, 0)
print_table(nonelectric_table)
```

Column Name	Proportion of Values NA
THC (g/mi)	0.09803
CO (g/mi)	0.09789
DT-Inertia Work Ratio Rating	0.04835
DT-Absolute Speed Change Ratg	0.04835
DT-Energy Economy Rating	0.04835
CO2 (g/mi)	0.01702
Aftertreatment Device Cd	0.00888
Aftertreatment Device Desc	0.00888
RND_ADJ_FE	0.00823
# of Cylinders and Rotors	0.00745

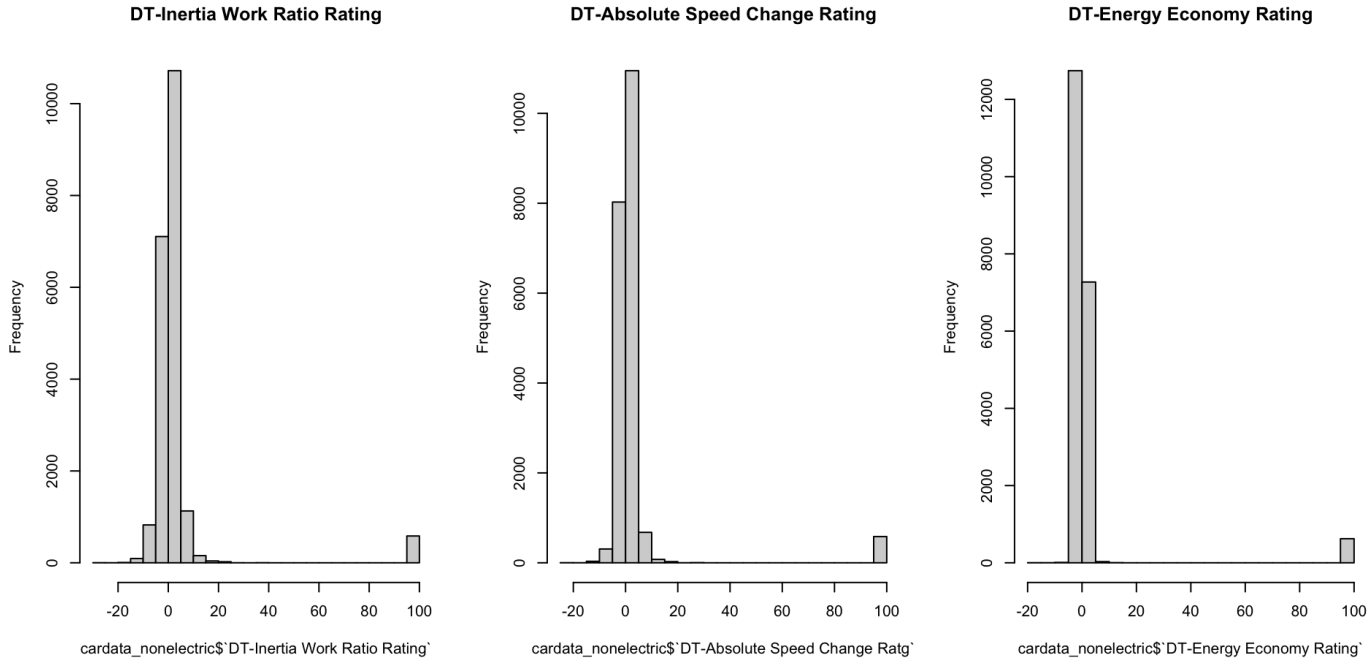
`DT-Inertia Work Ratio Rating`, `DT-Absolute Speed Change Ratg`, and `DT-Energy Economy Rating`

Let's start by observing these three variables:

- `DT-Inertia Work Ratio Rating`
- `DT-Absolute Speed Change Ratg`
- `DT-Energy Economy Rating`

We can make histograms to see what their distributions are like.

```
par(mfrow = c(1, 3))
hist(cardata_nonelectric$`DT-Inertia Work Ratio Rating`, main = 'DT-Inertia Work Ratio Rating', breaks = 20)
hist(cardata_nonelectric$`DT-Absolute Speed Change Ratg`, main = 'DT-Absolute Speed Change Rating', breaks = 20)
hist(cardata_nonelectric$`DT-Energy Economy Rating`, main = 'DT-Energy Economy Rating', breaks = 20)
```



It is evident that each of these columns has a set of observations marked with a rating around 100. This is likely comparable to an `NA` value, as the distribution of values for these attributes is almost entirely hovering around a value of 0. Let's replace these values with `NA` values so that we don't confuse ourselves with them being valid. Then, since the remaining distribution appears to be very symmetric, we can replace the missing values with the mean of the column for each variable.

```
# replace high outliers with NA
cardata_nonelectric$`DT-Inertia Work Ratio Rating`[cardata_nonelectric$`DT-Inertia Work Ratio Rating` > 50] <- NA
cardata_nonelectric$`DT-Absolute Speed Change Ratg`[cardata_nonelectric$`DT-Absolute Speed Change Ratg` > 50] <-
NA
cardata_nonelectric$`DT-Energy Economy Rating`[cardata_nonelectric$`DT-Energy Economy Rating` > 50] <- NA

# replace missing values with the mean value
cardata_nonelectric$`DT-Inertia Work Ratio Rating`[is.na(cardata_nonelectric$`DT-Inertia Work Ratio Rating`)] <-
 mean(cardata_nonelectric$`DT-Inertia Work Ratio Rating`, na.rm = TRUE)
cardata_nonelectric$`DT-Absolute Speed Change Ratg`[is.na(cardata_nonelectric$`DT-Absolute Speed Change Ratg`)] <
- mean(cardata_nonelectric$`DT-Absolute Speed Change Ratg`, na.rm = TRUE)
cardata_nonelectric$`DT-Energy Economy Rating`[is.na(cardata_nonelectric$`DT-Energy Economy Rating`)] <- mean(car
data_nonelectric$`DT-Energy Economy Rating`, na.rm = TRUE)
```

We can now observe our new, updated measurements of proportions of values in columns that are `NA`.

```
# create data frame for printing proportions of NA values
nonelectric_table <- calculate_proportions(cardata_nonelectric, 0)
print_table(nonelectric_table)
```

Column Name	Proportion of Values NA
THC (g/mi)	0.09803
CO (g/mi)	0.09789
CO2 (g/mi)	0.01702
Aftertreatment Device Cd	0.00888
Aftertreatment Device Desc	0.00888
RND_ADJ_FE	0.00823
# of Cylinders and Rotors	0.00745

`Aftertreatment Device Cd` and `Aftertreatment Device Desc`

The following two variables have the same number of missing values in the dataset. Let's see if they are from the same rows.

- `Aftertreatment Device Cd`
- `Aftertreatment Device Desc`

```
# extract row names where these three variables are missing
after_cd_null <- rownames(cardata_nonelectric[is.na(cardata_nonelectric$`Aftertreatment Device Cd`), ])
after_desc_null <- rownames(cardata_nonelectric[is.na(cardata_nonelectric$`Aftertreatment Device Desc`), ])

# ensure that the column names are all the same across all datasets
if ( mean(after_cd_null == after_desc_null) == 1 ) {
  print('These variables are missing in the same rows in the dataset.')
} else {
  print('These variables are missing in different rows in the dataset.')
}
```

```
## [1] "These variables are missing in the same rows in the dataset."
```

The missing values in the dataset for these two variables *do* come from the same rows, which makes sense given their association with each other; the `Aftertreatment Device Desc` variable is just a description for the label of the `Aftertreatment Device Cd` variable. Since these are categorical variables, and there are only a small number of `NA` values (less than 1%), we will leave the `NA` values as they are and likely remove them if we engage in an analysis on both columns. We opt not to remove the *rows* corresponding to these missing values because there is valuable data in the other 30 columns present.

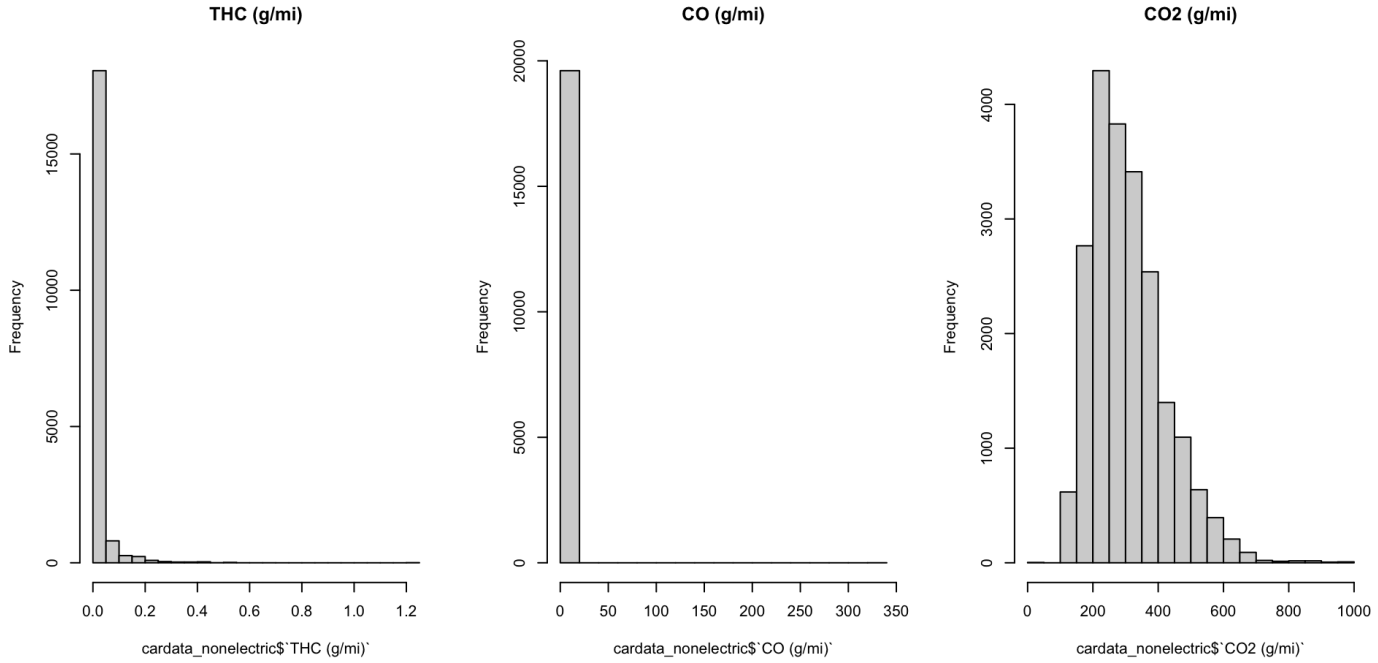
The last few variables containing missing values left to address are:

- `THC (g/mi)`
- `CO (g/mi)`
- `CO2 (g/mi)`
- `# of Cylinders and Rotors`
- `RND_ADJ_FE`

`THC (g/mi)`, `CO (g/mi)`, and `CO2 (g/mi)`

Let's start with these variables and observe their missing values.

```
par(mfrow = c(1, 3))
hist(cardata_nonelectric$`THC (g/mi)`, main = 'THC (g/mi)', breaks = 20)
hist(cardata_nonelectric$`CO (g/mi)`, main = 'CO (g/mi)', breaks = 20)
hist(cardata_nonelectric$`CO2 (g/mi)`, main = 'CO2 (g/mi)', breaks = 20)
```



Given the shapes of these distributions, we believe it is a safe choice to replace the missing values (which make up less than 10% of the data for each distribution) with the median value in that distribution. This is a conservative approach and allows us to refrain from making any bold assumptions about the true values that are missing.

```
# replace missing values with the mean value
cardata_nonelectric$`THC (g/mi)`[is.na(cardata_nonelectric$`THC (g/mi)`)] <- mean(cardata_nonelectric$`THC (g/mi)
`, na.rm = TRUE)
cardata_nonelectric$`CO (g/mi)`[is.na(cardata_nonelectric$`CO (g/mi)`)] <- mean(cardata_nonelectric$`CO (g/mi)`,
 na.rm = TRUE)
cardata_nonelectric$`CO2 (g/mi)`[is.na(cardata_nonelectric$`CO2 (g/mi)`)] <- mean(cardata_nonelectric$`CO2 (g/mi)
`, na.rm = TRUE)
```

`# of Cylinders and Rotors`

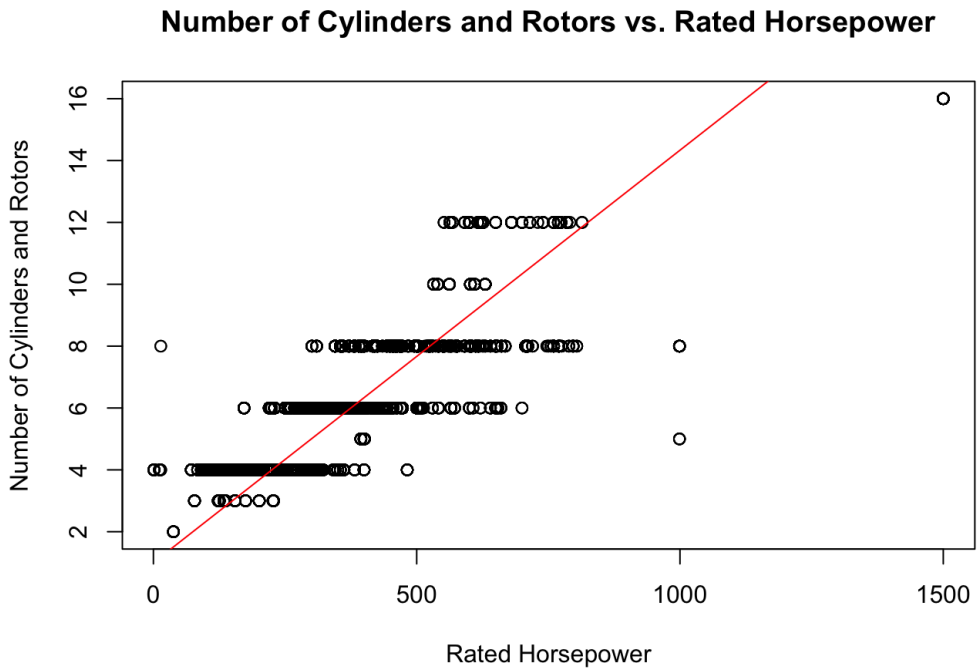
Let's move on to `# of Cylinders and Rotors`.

```
# table of number of cylinders
table(cardata_nonelectric$`# of Cylinders and Rotors`)
```

```
##
##     2     3     4     5     6     8    10    12    16
##    21   390 10709    45  6531  3333   166   350    31
```

We can see above that the number of cylinders and rotors is almost exclusively an even number, with values of 3 and 5 appearing some number of times.

```
# plot number of cylinders versus horsepower
plot(cardata_nonelectric$`Rated Horsepower`, cardata_nonelectric$`# of Cylinders and Rotors`, xlab = 'Rated Horse
power', ylab = 'Number of Cylinders and Rotors', main = 'Number of Cylinders and Rotors vs. Rated Horsepower')
abline(a = 1, b = 1/75, col = 'red')
```



Above, we can see an approximation of the number of cylinders (`# of Cylinders and Rotors`) and rotors predicted by the horsepower (`Rated Horsepower`). Using this approximation, we will fill in the missing values of `# of Cylinders and Rotors` using the nearest even value. Note that the even values make up the vast majority in this dataset, as values of 1, 3, and 5, for instance, are very rare.

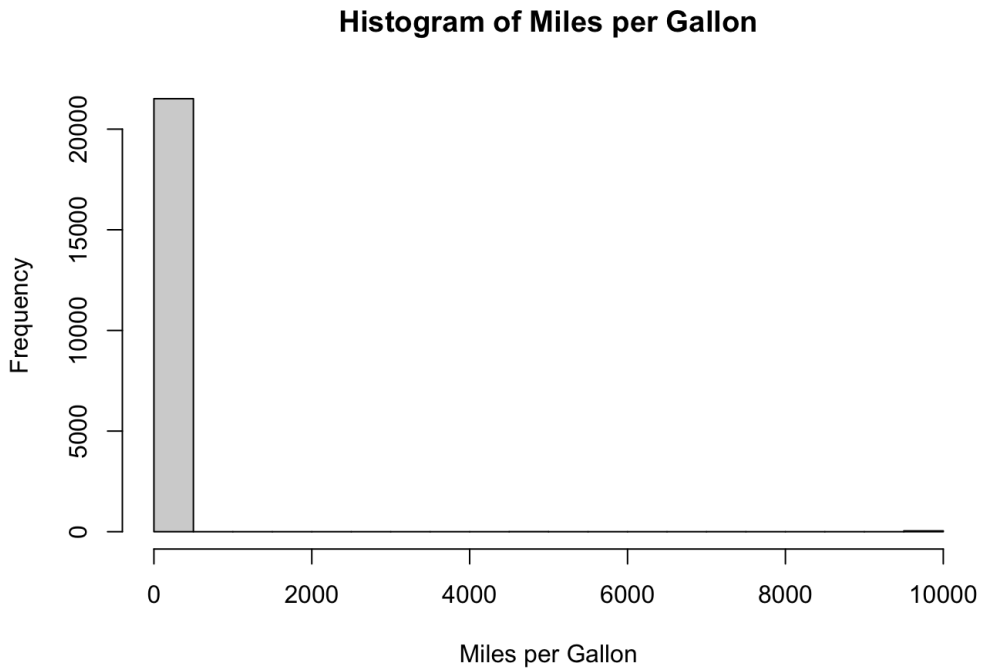
```
# replace missing cylinder values with the approximation above
for (i in 1:nrow(cardata_nonelectric)) {
  if (is.na(cardata_nonelectric$`# of Cylinders and Rotors`[i])) {
    val <- 1 + cardata_nonelectric$`Rated Horsepower`[i] / 75
    cardata_nonelectric$`# of Cylinders and Rotors`[i] <- round(val / 2) * 2
  }
}
```

`RND_ADJ_FE`

Finally, let's move on to `RND_ADJ_FE` .

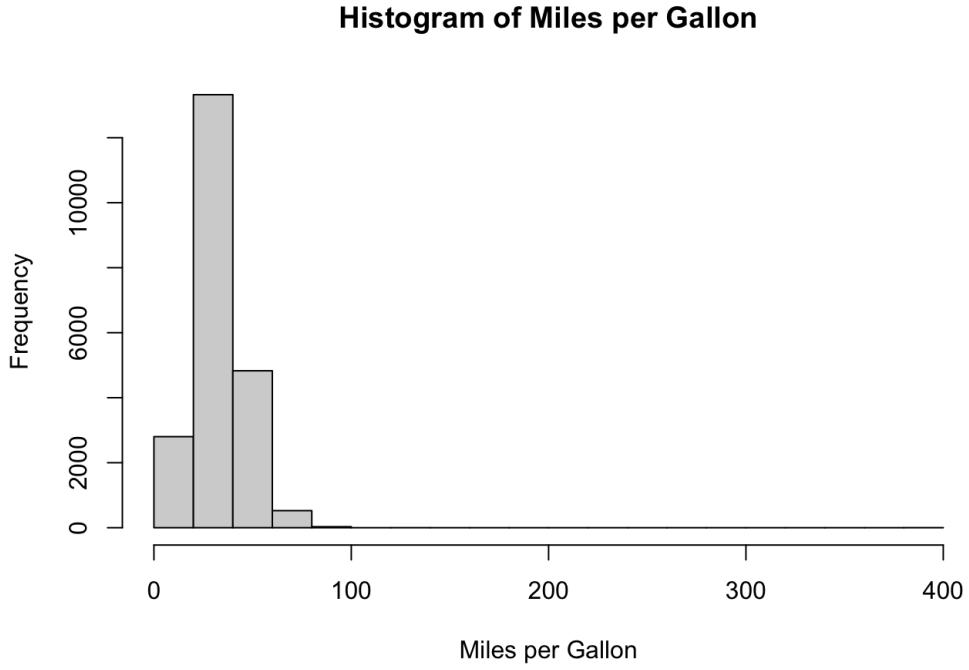
As we did above with the electric vehicles, let's observe a histogram of the `RND_ADJ_FE` variable for the non-electric vehicles. Remember, this variable expresses the number of miles per gallon for the vehicle.

```
# plot histogram of the electric car MPG values
hist(cardata_nonelectric$RND_ADJ_FE, xlab = 'Miles per Gallon', main = 'Histogram of Miles per Gallon', breaks =
20)
```



We can see that there is at least one outlier on the high end of the distribution, expressing a value of 1000. We will replace these value(s) with missing values in order to obtain results that have more validity. We can then try to view the true distribution of values by constructing another histogram.

```
# replace high outliers with missing value
cardata_nonelectric$RND_ADJ_FE[cardata_nonelectric$RND_ADJ_FE > 500] <- NA

# plot histogram of the electric car MPG values
hist(cardata_nonelectric$RND_ADJ_FE, xlab = 'Miles per Gallon', main = 'Histogram of Miles per Gallon', breaks =
20)
```



We will perform the same task as before with the electric vehicles by replacing the missing values in this column with the *median* value of the distribution, as this is a safer way to do so without making any bold assumptions.

```
# replace missing values with the median value
cardata_nonelectric$RND_ADJ_FE[is.na(cardata_nonelectric$RND_ADJ_FE)] <- median(cardata_nonelectric$RND_ADJ_FE, n
a.rm = TRUE)
```

We can now observe the columns with missing values in the non-electric vehicle dataset and notice that we have cleaned it to our liking. Remember that we opted to keep the missing values for the `Aftertreatment Device Cd` and `Aftertreatment Device Desc` columns, as mentioned above.

```
# print the proportions of NA values in each column using a threshold of 0%
nonelectric_table <- calculate_proportions(cardata_nonelectric, 0)
print_table(nonelectric_table)
```

Column Name	Proportion of Values NA
Aftertreatment Device Cd	0.00888
Aftertreatment Device Desc	0.00888

Final Look

Let's take a look at the proportions of missing values now, as we have addressed them all (and kept some).

```
# print the proportions of NA values in each column
electric_table <- calculate_proportions(cardata_electric, 0)
nonelectric_table <- calculate_proportions(cardata_nonelectric, 0)
print_table(electric_table)
```

Column Name	Proportion of Values NA
CO2 (g/mi)	1

```
print_table(nonelectric_table)
```

Column Name	Proportion of Values NA
Aftertreatment Device Cd	0.00888
Aftertreatment Device Desc	0.00888

After addressing the missing values in the data, we find that we only have approximately **0.15%** of our data missing. This is a great improvement from the **15.64%** that we had initially and something that we can take with us as we look to analyze the data further.

Export Clean Data

Finally, we need to export this cleaned data so that we can use it in our analyses.

```
write.csv(cardata_electric, '../data/cardata_electric_clean.csv')
write.csv(cardata_nonelectric, '../data/cardata_nonelectric_clean.csv')
```

ANLY 511 Final Project - MANOVA

LINLIN WANG

2022-12-02

MANOVA

MANOVA stands for multivariate analysis of variance. It's basically used to evaluate mean differences on two or more dependent variables simultaneously. That's the main difference compared with ANOVA.

We try to answer below research questions by performing MANOVA. And CO2, CO, and THC are three dependent variables.

Research Questions

1. We want to know if there is statistically significant difference in CO2, CO, and THC between the different types of vehicle.
2. We want to know if there is statistically significant difference in CO2, CO, and THC between vehicle manufacturers.
3. We want to know if there is statistically significant difference in CO2, CO, and THC between the different vehicle transmission types.

Assumptions of MANOVA

There are additional assumptions of MANOVA:

1. Homogeneity of the variances across the range of predictors:

Our data should have equal variance-covariance matrics for each combination formed by each group in the independent variable.

2. Multicollinearity:

Our data should be no multicollinearity among dependent variables.

Loading packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
```

```
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggpubr)
library(rstatix)
```

```
##
## Attaching package: 'rstatix'
##
## The following object is masked from 'package:stats':
##
##     filter
```

```
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(broom)
library(gplots)
```

```
##
## Attaching package: 'gplots'
##
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(mvnormalTest)
library(heplots)
```

Enter Data

```
# Read cleaned nonelectric data
nonelectric<-read.csv("/Users/linlinw/Desktop/ANLY511-Final-Project-main/data/cardata_nonelectric_clean
# Remove first column X
nonelectric<-nonelectric[,-1]
# View first couple rows of data
head(nonelectric)
```

```
##   Model.Year Vehicle.Manufacturer.Name Veh.Mfr.Code Represented.Test.Veh.Make
## 1       2018              aston martin          ASX              Aston Martin
## 2       2018              aston martin          ASX              Aston Martin
## 3       2018              aston martin          ASX              Aston Martin
## 4       2018              aston martin          ASX              Aston Martin
## 5       2018              aston martin          ASX              Aston Martin
## 6       2018              aston martin          ASX              Aston Martin
##   Represented.Test.Veh.Model Test.Veh.Displacement..L. Vehicle.Type
## 1                       DB11                       5.2          Car
## 2                       DB11                       5.2          Car
## 3                    DB11 V8                       4.0          Car
## 4                    DB11 V8                       4.0          Car
## 5                   Rapide S                       6.0          Car
## 6                   Rapide S                       6.0          Car
##   Rated.Horsepower X..of.Cylinders.and.Rotors Tested.Transmission.Type.Code
## 1              600                         12                            SA
## 2              600                         12                            SA
## 3              503                          8                            SA
## 4              503                          8                            SA
## 5              552                         12                            SA
## 6              552                         12                            SA
##   Tested.Transmission.Type X..of.Gears Transmission.Lockup. Drive.System.Code
## 1           Semi-Automatic           8                    Y                 R
## 2           Semi-Automatic           8                    Y                 R
## 3           Semi-Automatic           8                    Y                 R
## 4           Semi-Automatic           8                    Y                 R
## 5           Semi-Automatic           8                    Y                 R
## 6           Semi-Automatic           8                    Y                 R
##   Drive.System.Description Equivalent.Test.Weight..lbs.. Axle.Ratio N.V.Ratio
## 1      2-Wheel Drive, Rear                           4500       2.70      22.2
## 2      2-Wheel Drive, Rear                           4500       2.70      22.2
## 3      2-Wheel Drive, Rear                           4500       2.70      22.2
## 4      2-Wheel Drive, Rear                           4500       2.70      22.2
## 5      2-Wheel Drive, Rear                           4750       2.73      22.4
## 6      2-Wheel Drive, Rear                           4750       2.73      22.4
##   Test.Fuel.Type.Description THC..g.mi. CO..g.mi. CO2..g.mi. RND_ADJ_FE
## 1       Tier 2 Cert Gasoline   0.024700  0.418000     466.87       18.8
## 2       Tier 2 Cert Gasoline   0.001155  0.067334     285.00       30.9
## 3       Tier 2 Cert Gasoline   0.026500  0.070000     386.66       22.7
## 4       Tier 2 Cert Gasoline   0.000500  0.030000     259.74       33.8
## 5       Tier 2 Cert Gasoline   0.026900  0.500000     511.93       17.3
## 6       Tier 2 Cert Gasoline   0.000800  0.060000     296.63       29.9
##   DT.Inertia.Work.Ratio.Rating DT.Absolute.Speed.Change.Ratg
## 1                   -2.5300000                   -1.7300000
## 2                    1.3600000                    0.4400000
## 3                  -11.9900000                   -9.2600000
## 4                   -3.6400000                   -3.2100000
## 5                    0.5655838                    0.4420405
## 6                    0.5655838                    0.4420405
##   DT.Energy.Economy.Rating Target.Coef.A..lbf. Target.Coef.B..lbf.mph.
## 1               -1.7100000               40.94                  0.0169
## 2               -0.5900000               40.94                  0.0169
## 3               -7.7100000               40.94                  0.0169
## 4               -0.9600000               40.94                  0.0169
```

```
## 5               -0.2002973               32.66                  0.6085
## 6               -0.2002973               32.66                  0.6085
##   Target.Coef.C..lbf.mph..2. Set.Coef.A..lbf. Set.Coef.B..lbf.mph.
## 1                     0.0271           6.810               0.0807
## 2                     0.0271           6.810               0.0807
## 3                     0.0271          11.260               0.0919
## 4                     0.0271          11.260               0.0919
## 5                     0.0198           1.093               2.1980
## 6                     0.0198           1.093               2.1980
##   Set.Coef.C..lbf.mph..2. Aftertreatment.Device.Cd Aftertreatment.Device.Desc
## 1                  0.0245                      TWC         Three-way catalyst
## 2                  0.0245                      TWC         Three-way catalyst
## 3                  0.0251                      TWC         Three-way catalyst
## 4                  0.0251                      TWC         Three-way catalyst
## 5                  0.0280                      TWC         Three-way catalyst
## 6                  0.0280                      TWC         Three-way catalyst
##   Police...Emergency.Vehicle. Averaging.Method.Cd Averging.Method.Desc
## 1                           N                   N         No averaging
## 2                           N                   N         No averaging
## 3                           N                   N         No averaging
## 4                           N                   N         No averaging
## 5                           N                   N         No averaging
## 6                           N                   N         No averaging
```

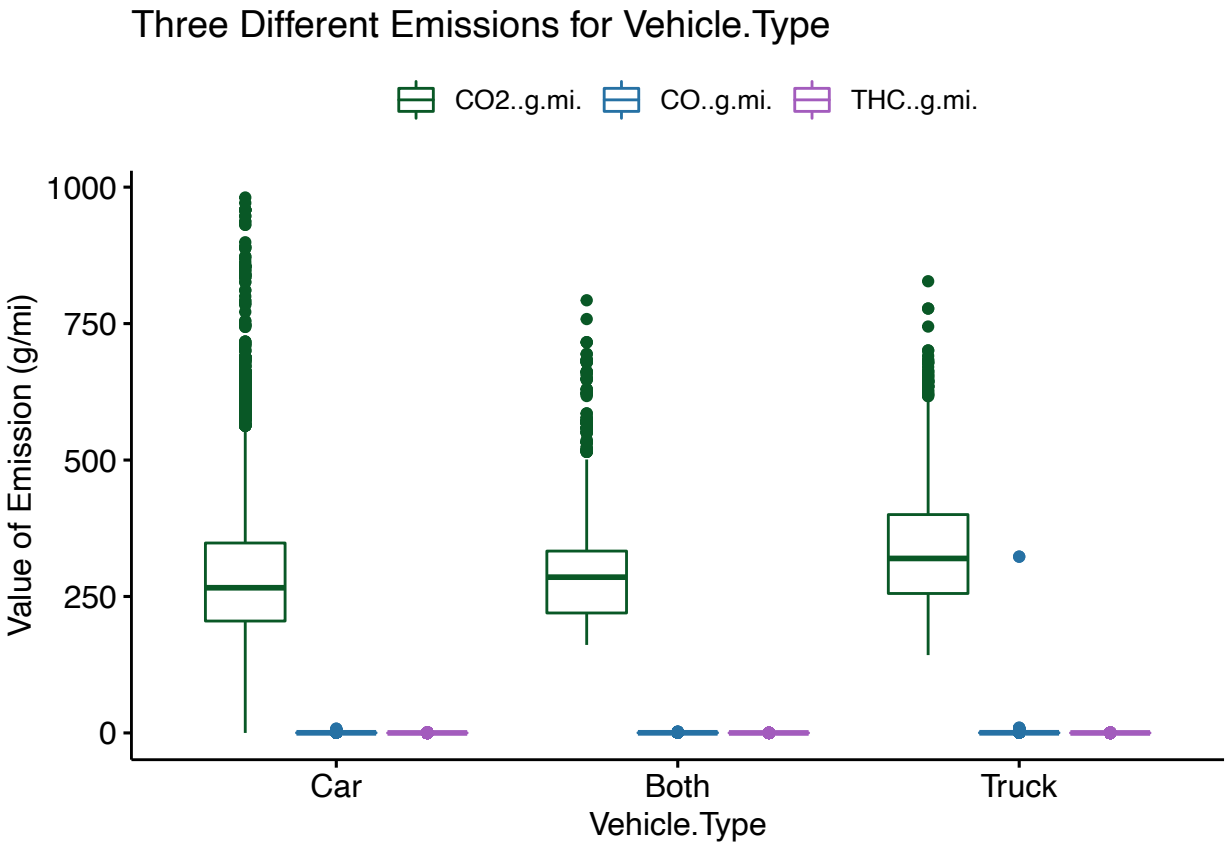
Research Question 1: Is there any important difference in CO2, CO, and THC between the different types of vehicle?

Exploratory Data Analysis:

```
# Create color palettes
Blues <- colorRampPalette(c("#0A146B", "#A9A3DA"))
Purples <- colorRampPalette(c("#3E1370", "#BDA3DA"))
GrBuPuPi <- c("#095826", "#0E7032", "#10913F", "#55A472", "#8CBF9E", "#8CBFB8",
              "#63B7AC", "#2D9A8B", "#137568", "#094E45", "#0B3C5C", "#17547C",
              "#2671A4", "#3C8CC1", "#72B1DB", "#96C3E1", "#B0CDE1", "#B0B3E1",
              "#858ACD", "#4F55AB", "#1923B3", "#0E1468", "#3C1075", "#5821A1",
              "#6B27C4", "#9455E5", "#A278D8", "#A990CA", "#ADA0BF", "#C1A5CB",
              "#B887CA", "#A35CBD", "#762594")
```

```
# Visaulize dataset
ggboxplot(
  nonelectric, x = "Vehicle.Type", y = c("CO2..g.mi.", "CO..g.mi.", "THC..g.mi."),
  merge = TRUE, palette = c("#095826",  "#2671A4","#A35CBD" ),
  title = "Three Different Emissions for Vehicle.Type",
  ylab = "Value of Emission (g/mi)"
  )
```

```
## Warning: `gather_()` was deprecated in tidyr 1.2.0.
## Please use `gather()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```



From the above boxplot, we can see there is significant difference in CO2 and CO between vehicle type, and significant difference in CO2 and THC between vehicle type. However, the difference in CO and THC between vehicle type is not obvious.

Test assumption

1. Check for Homogeneity

```
boxM(Y = nonelectric[, c("CO2..g.mi.", "CO..g.mi.", "THC..g.mi.")], group = nonelectric$Vehicle.Type)
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  nonelectric[, c("CO2..g.mi.", "CO..g.mi.", "THC..g.mi.")]
## Chi-Sq (approx.) = 58734, df = 12, p-value < 2.2e-16
```

Since the p-value is significant for Box's M test, we reject the null hypothesis at 5% significance level and conclude that variance-covariance matrices are not equal for each combination formed by each group in the independent variable. Thus, this assumption is satisfied.

2. Check Multicollinearity

```
cor_co2_co<-cor.test(x = nonelectric$CO2..g.mi., y = nonelectric$CO..g.mi., method = "pearson")$estimate
cor_co2_thc<-cor.test(x = nonelectric$CO2..g.mi., y = nonelectric$THC..g.mi., method = "pearson")$estima
cor_thc_co<-cor.test(x = nonelectric$THC..g.mi., y = nonelectric$CO..g.mi., method = "pearson")$estimate
result<-cbind(cor_co2_co, cor_co2_thc, cor_thc_co)
result
```

```
##     cor_co2_co cor_co2_thc cor_thc_co
## cor  0.0562401   0.3195182 0.07842236
```

As the correlation coefficient between each dependent variable is smaller than 0.9, there is no multicollinearity. So this assumption is hold.

Perform MANOVA

Hypotheses

H_0 : There is no significant difference in CO2, CO, and THC between the different types of vehicle.

H_a : There is significant difference in CO2, CO, and THC between the different types of vehicle.

```
# Fit the MANOVA model
fit1 = manova(cbind(CO2..g.mi., CO..g.mi., THC..g.mi.) ~ Vehicle.Type, data = nonelectric)
summary(fit1, intercept = TRUE)
```

```
##                 Df  Pillai approx F num Df den Df    Pr(>F)
## (Intercept)      1 0.88848    57714      3  21733 < 2.2e-16 ***
## Vehicle.Type     2 0.04618      171      6  43468 < 2.2e-16 ***
## Residuals    21735
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value for Vehicle.Type variable is smaller than the significance level 0.05, we can reject the null hypotheses at 5% significance level, and conclude that there is statistically significant difference in CO2, CO, and THC between the different types of vehicle.

However, we are unclear about which emissions are affected by vehicle type. We perform univariate ANOVAs to figure it out.

```
summary.aov(fit1)
```

```
##  Response CO2..g.mi. :
##                 Df    Sum Sq Mean Sq F value    Pr(>F)
## Vehicle.Type     2  10263186 5131593  421.74 < 2.2e-16 ***
## Residuals    21735 264465545   12168
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response CO..g.mi. :
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## Vehicle.Type     2    126  62.795  6.4026 0.00166 **
## Residuals    21735 213170   9.808
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response THC..g.mi. :
##                 Df Sum Sq  Mean Sq F value    Pr(>F)
## Vehicle.Type     2  0.147 0.073417  35.578 3.751e-16 ***
## Residuals    21735 44.852 0.002064
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

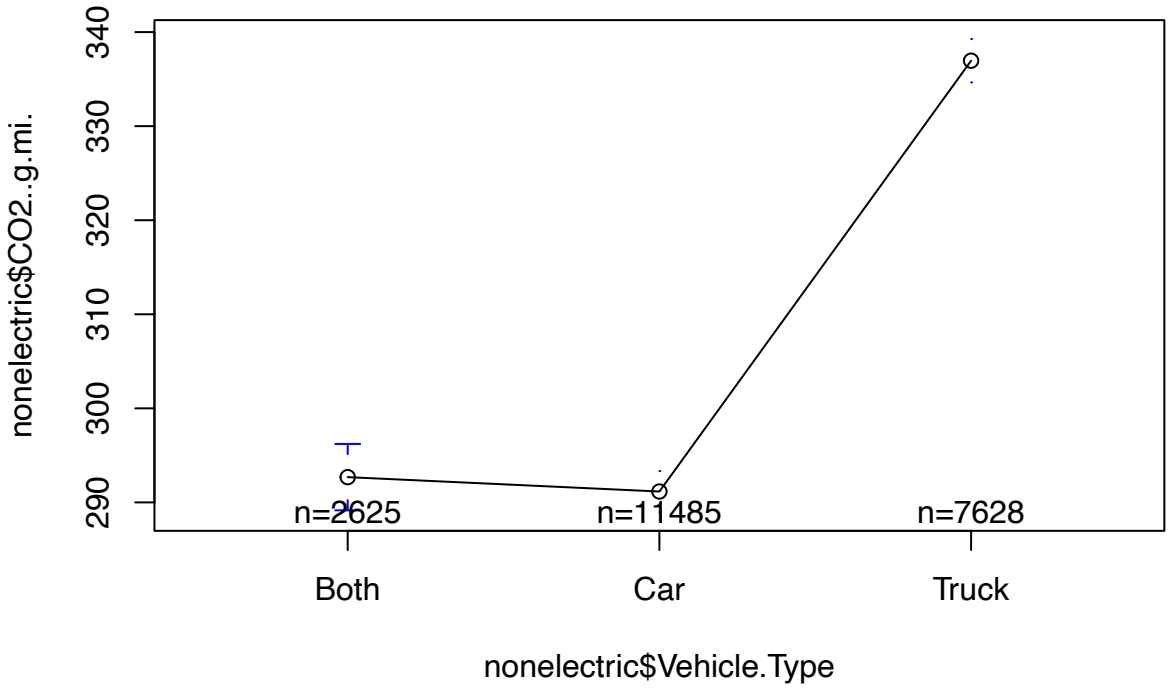
We can see from the output that the p-value for all univariate ANOVAs are smaller than significance level 0.05, which indicates that vehicle type has a statistically significant effect on CO2, CO, and THC emissions.

Visualizing Group Means

Visualizing the Group means for each level of our independent variable vehicle type is helpful to get a better understanding of our results.

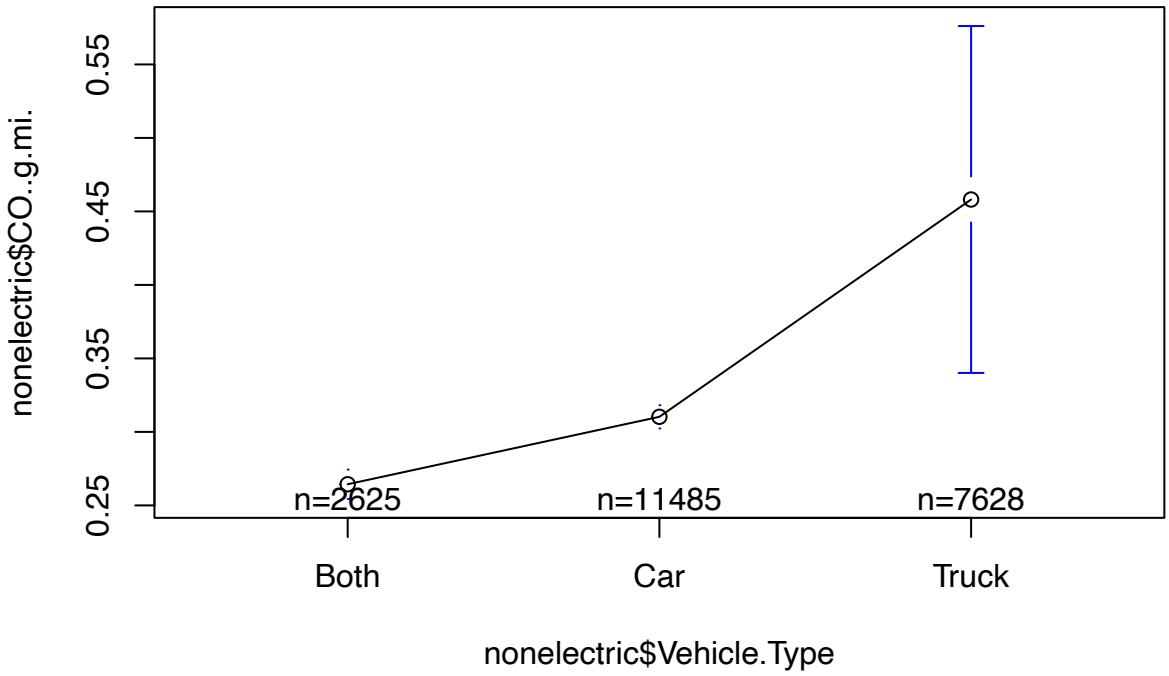
```
#visualize mean CO2 by vehicle type
plotmeans(nonelectric$CO2..g.mi. ~ nonelectric$Vehicle.Type)
```

```
## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```

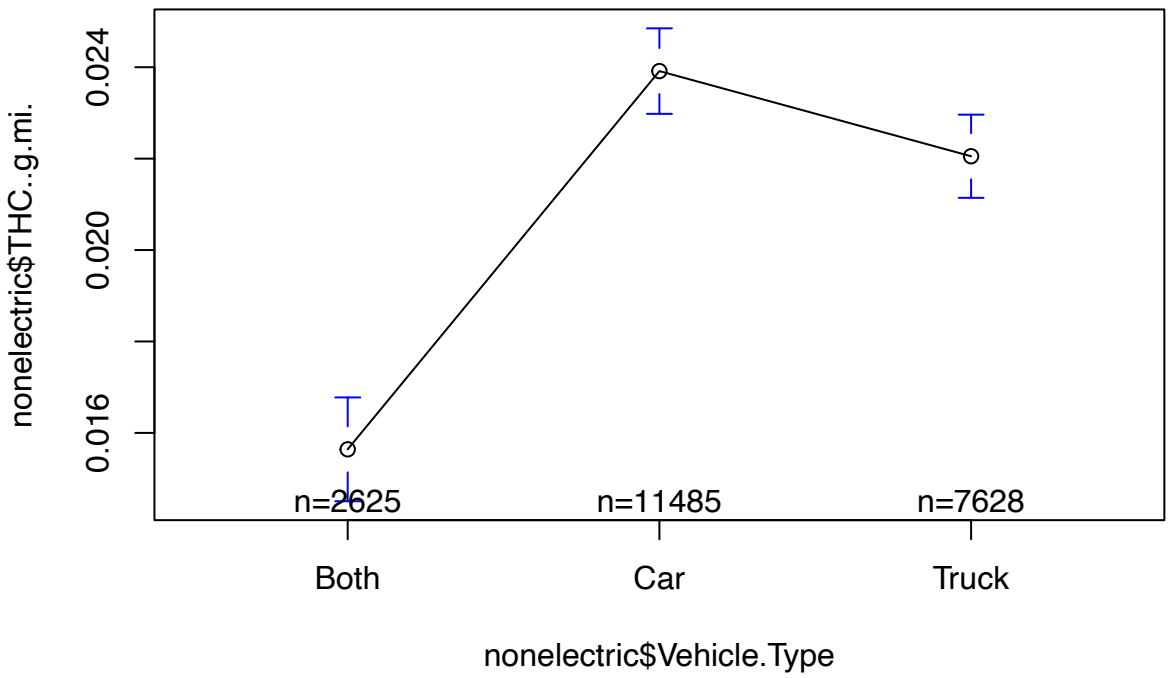



```
#visualize mean CO by vehicle type
plotmeans(nonelectric$CO..g.mi. ~ nonelectric$Vehicle.Type)
```

```
## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```



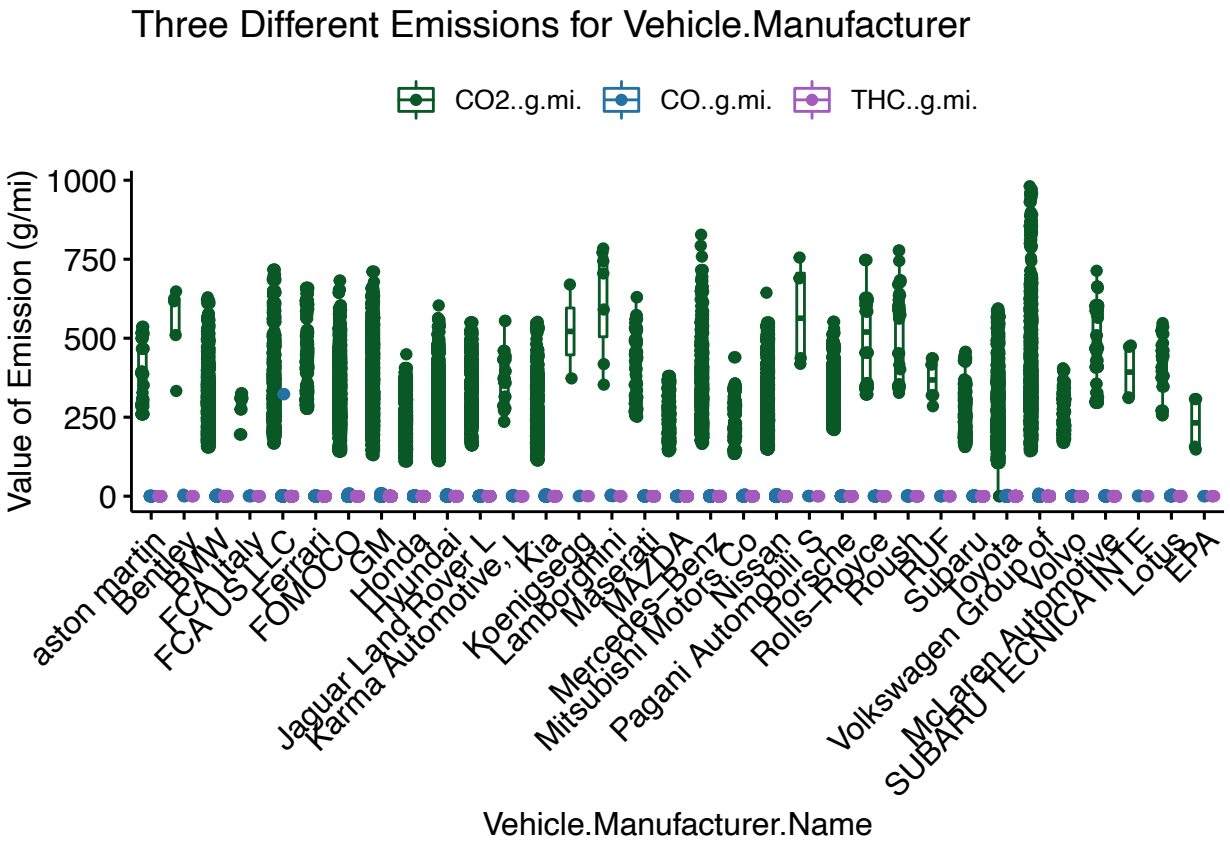

```
#visualize mean THC by vehicle type
plotmeans(nonelectric$THC..g.mi. ~ nonelectric$Vehicle.Type)
```



From the above three plots, we can see that the mean CO2, CO, and THC varies quite a bit by vehicle types. This matches the results from our MANOVA, which indicates that there is statistically significant difference in three emissions based on vehicle types.

Research Question 2: Is there any important difference in CO2, CO, and THC between the different vehicle manufacturers?

```
# Visaulize dataset
p2<-ggboxplot(
  nonelectric, x = "Vehicle.Manufacturer.Name", y = c("CO2..g.mi.", "CO..g.mi.", "THC..g.mi."),
  merge = TRUE, palette = c("#095826",  "#2671A4","#A35CBD" ),
  title = "Three Different Emissions for Vehicle.Manufacturer",
  ylab = "Value of Emission (g/mi)",
  add = "jitter"
)
p2 + rotate_x_text(45)
```



From this boxplot, we can get same conclusion with first plot. There is no important difference in CO and THC between vehicle manufacturer.

Hypotheses2

H_0 : There is no significant difference in CO2, CO, and THC between vehicle manufacturer.

H_a : There is significant difference in CO2, CO, and THC between vehicle manufacturer.

```
# Fit the MANOVA model
fit2 = manova(cbind(CO2..g.mi., CO..g.mi., THC..g.mi.) ~ Vehicle.Manufacturer.Name, data = nonelectric)
summary(fit2, intercept = TRUE)
```

```
##                              Df  Pillai approx F num Df den Df    Pr(>F)
## (Intercept)                   1 0.90421    68286      3  21703 < 2.2e-16 ***
## Vehicle.Manufacturer.Name    32 0.22328       55     96  65115 < 2.2e-16 ***
## Residuals                 21705
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value for vehicle manufacturer variable is smaller than the significance level 0.05, we can reject the null hypotheses at 5% significance level, and conclude that there is statistically significant difference in CO2, CO, and THC between the different types of vehicle.

Then we perform univariate ANOVAs to figure it out which emissions are affected by vehicle manufacturer

```
summary.aov(fit2)
```

```
##  Response CO2..g.mi. :
##                             Df    Sum Sq Mean Sq F value    Pr(>F)    
## Vehicle.Manufacturer.Name    32  51944440 1623264  158.15 < 2.2e-16 ***
## Residuals                 21705 222784291   10264                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##  Response CO..g.mi. :
##                             Df Sum Sq Mean Sq F value    Pr(>F)    
## Vehicle.Manufacturer.Name    32    701 21.8911   2.235 7.695e-05 ***
## Residuals                 21705 212595  9.7947                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##  Response THC..g.mi. :
##                             Df Sum Sq  Mean Sq F value    Pr(>F)    
## Vehicle.Manufacturer.Name    32  1.726 0.053943  27.057 < 2.2e-16 ***
## Residuals                 21705 43.273 0.001994                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from the output that the p-value for all univariate ANOVAs are smaller than significance level 0.05, which indicates that Vehicle Manufacturer has a statistically significant effect on CO2, CO, and THC emissions.

Visualizing Group Means

Visualizing the Group means for each level of our independent variable Vehicle.Manufacturer.Name is helpful to get a better understanding of our results.

```
#visualize mean CO2 by vehicle type
plotmeans(nonelectric$CO2..g.mi. ~ nonelectric$Vehicle.Manufacturer.Name) +rotate_x_text(45)
```

```
## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```

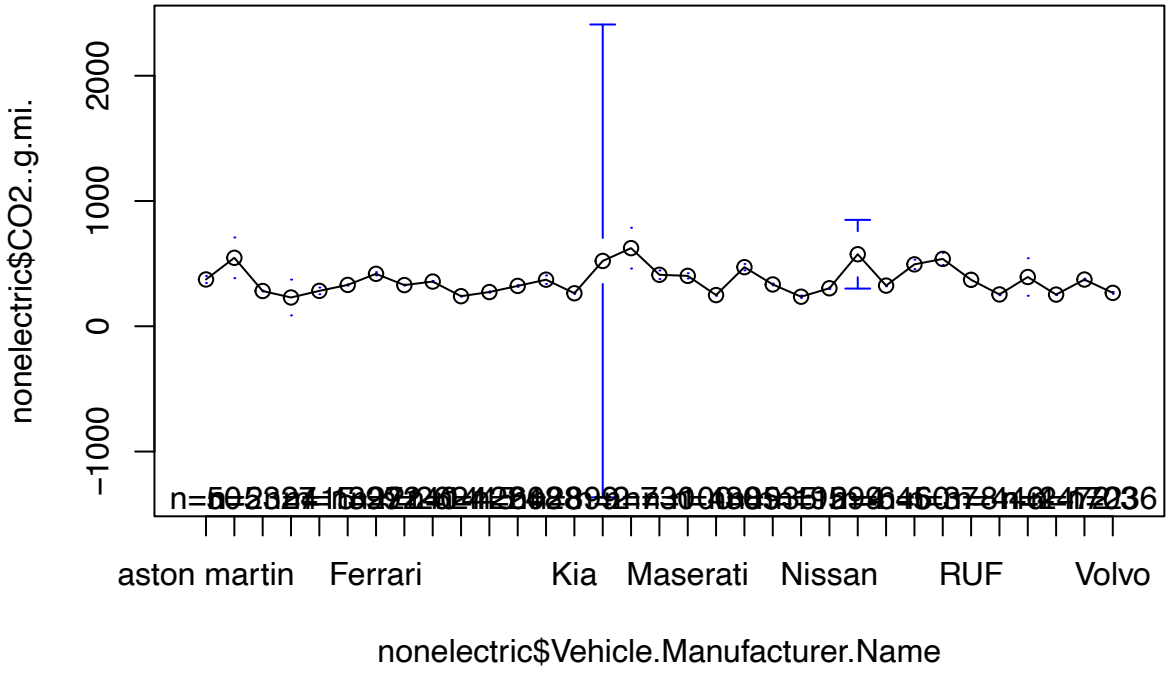
```
## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```

```
## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```

```
## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```

```
## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```

```
## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```




```
## NULL
```

```
#visualize mean CO by vehicle type
plotmeans(nonelectric$CO..g.mi. ~ nonelectric$Vehicle.Manufacturer.Name)
```

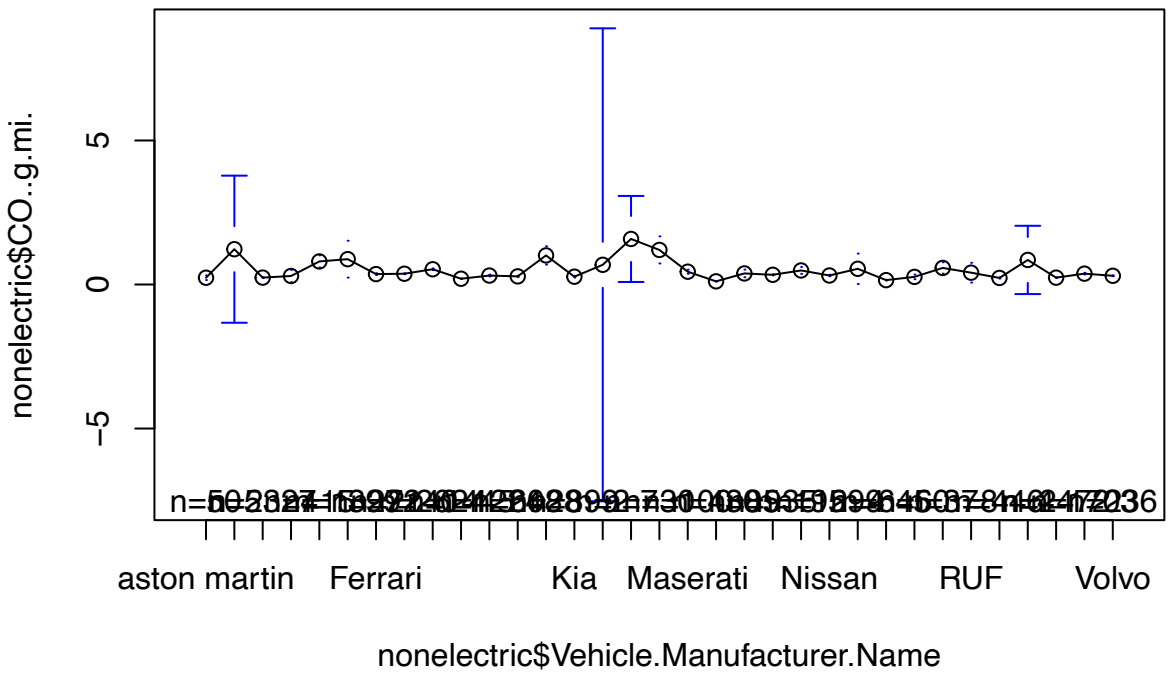
```
## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
```

```
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
```

```
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```

```
## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
```

```
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```



```
#visualize mean THC by vehicle type
plotmeans(nonelectric$THC..g.mi. ~ nonelectric$Vehicle.Manufacturer.Name)
```

```
## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```

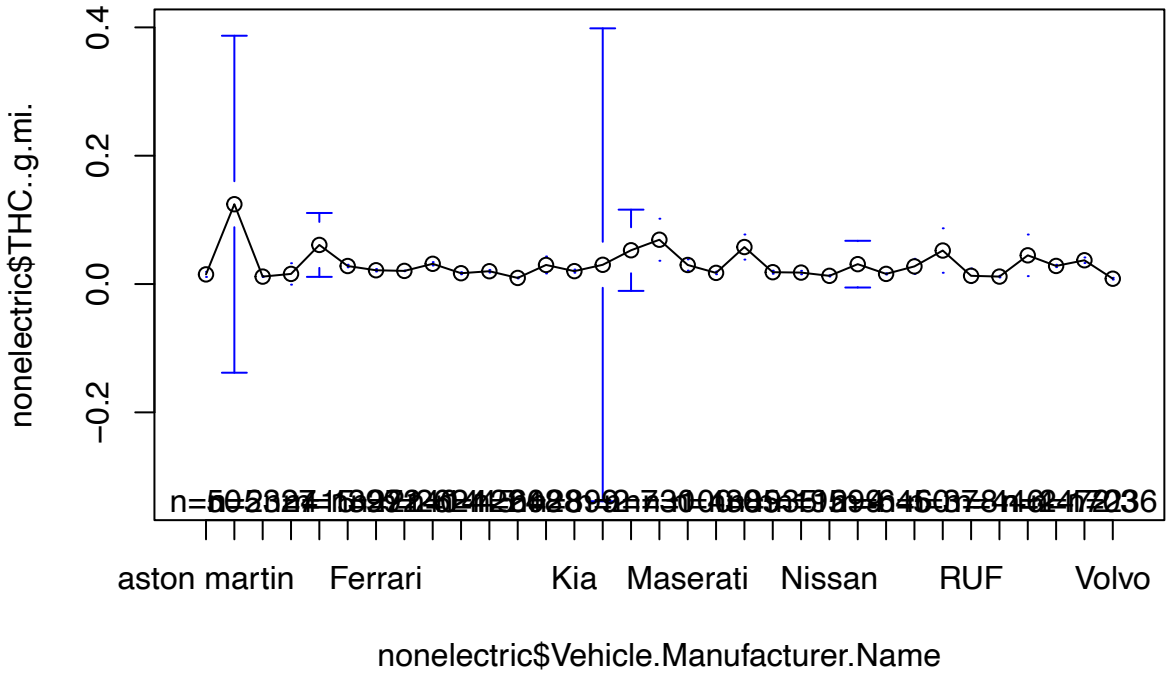
```
## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
```

```
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
```

```
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```

```
## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
```

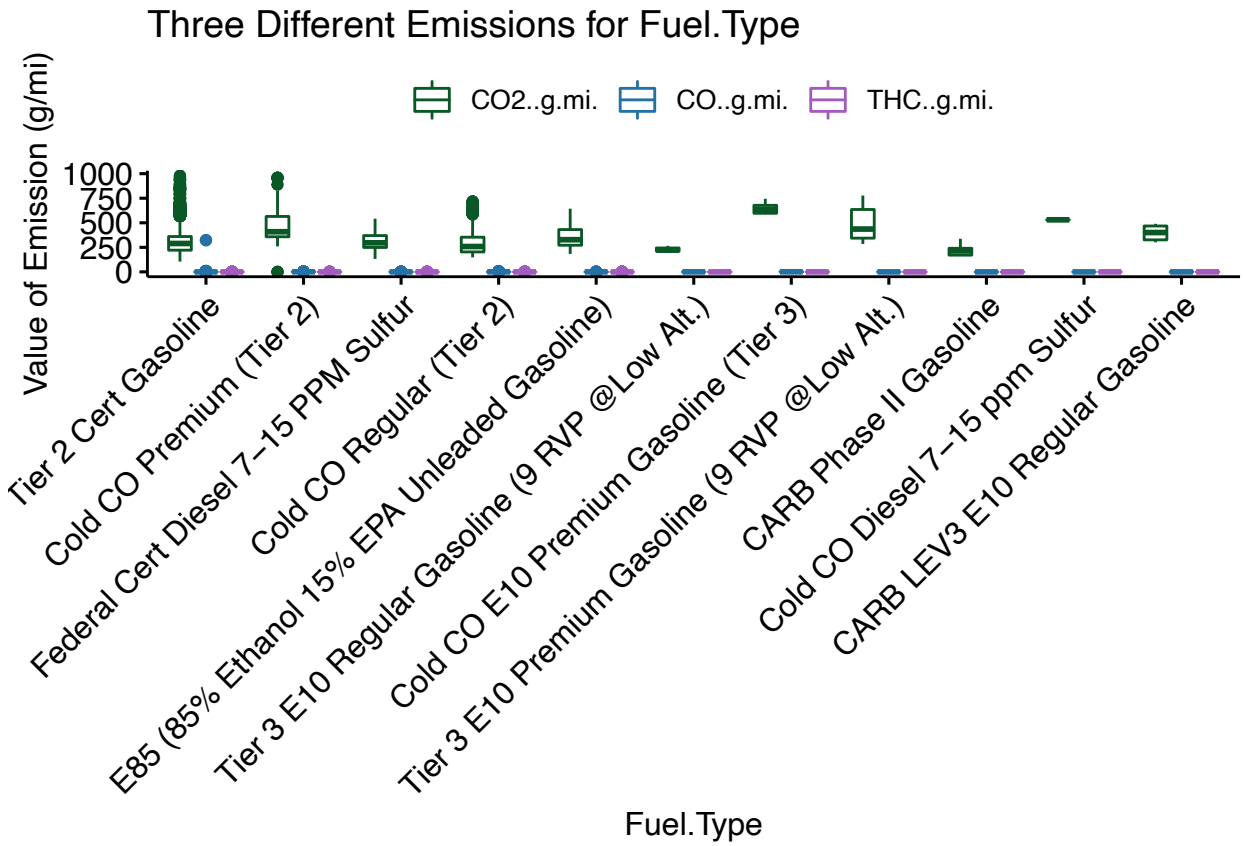
```
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```



From the above three plots, we can see that the mean CO2, CO, and THC do not vary a lot by vehicle manufacturer. This doesn't matches the results from our MANOVA.

Research Question 3: Is there any important difference in CO2, CO, and THC between the different vehicle transmission types?

```
# Visaulize dataset
p3<-ggboxplot(
  nonelectric, x = "Test.Fuel.Type.Description", y = c("CO2..g.mi.", "CO..g.mi.", "THC..g.mi."),
  merge = TRUE, palette = c("#095826",  "#2671A4","#A35CBD" ),
  title = "Three Different Emissions for Fuel.Type",
  ylab = "Value of Emission (g/mi)",
  xlab = "Fuel.Type"
)
p3 + rotate_x_text(45)
```



Hypotheses 3

H_0 : There is no significant difference in CO2, CO, and THC between the different types of fuel

H_a : There is significant difference in CO2, CO, and THC between the different types of fuel

```
# Fit the MANOVA model
fit3 = manova(cbind(CO2..g.mi., CO..g.mi., THC..g.mi.) ~ Test.Fuel.Type.Description, data = nonelectric
summary(fit3, intercept = TRUE)
```

```
##                               Df  Pillai approx F num Df den Df    Pr(>F)
## (Intercept)                    1 0.88787    57342      3  21725 < 2.2e-16 ***
## Test.Fuel.Type.Description    10 0.52439      460     30  65181 < 2.2e-16 ***
## Residuals                  21727
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value for Fuel.Type variable is smaller than the significance level 0.05, we can reject the null hypotheses at 5% significance level, and conclude that there is statistically significant difference in CO2, CO, and THC between the different types of fuel.

However, we are unclear about which emissions are affected by fuel type. We perform univariate ANOVAs to figure it out.

```
summary.aov(fit3)
```

```
##  Response CO2..g.mi. :
```

```
##                                Df    Sum Sq Mean Sq F value    Pr(>F)
## Test.Fuel.Type.Description    10  11752782 1175278  97.101 < 2.2e-16 ***
## Residuals                  21727 262975949   12104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response CO..g.mi. :
##                                Df Sum Sq Mean Sq F value    Pr(>F)
## Test.Fuel.Type.Description    10    439  43.911  4.4822 2.379e-06 ***
## Residuals                  21727 212856   9.797
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response THC..g.mi. :
##                                Df Sum Sq Mean Sq F value    Pr(>F)
## Test.Fuel.Type.Description    10 21.274 2.12738  1948.2 < 2.2e-16 ***
## Residuals                  21727 23.725 0.00109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from the output that the p-value for all univariate ANOVAs are smaller than significance level 0.05, which indicates that fuel type has a statistically significant effect on CO2, CO, and THC emissions.

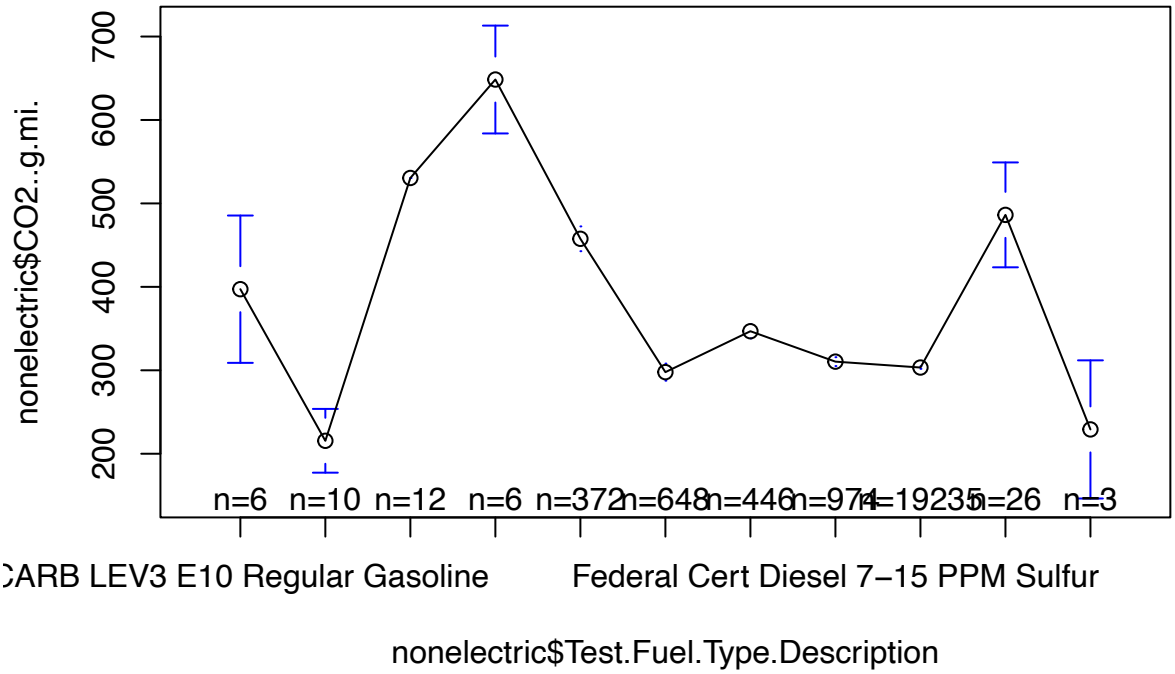
Visualizing Group Means

Visualizing the Group means for each level of our independent variable vehicle type is helpful to get a better understanding of our results.

```
#visualize mean CO2 by vehicle type
plotmeans(nonelectric$CO2..g.mi. ~ nonelectric$Test.Fuel.Type.Description)
```

```
## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```

```
## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```

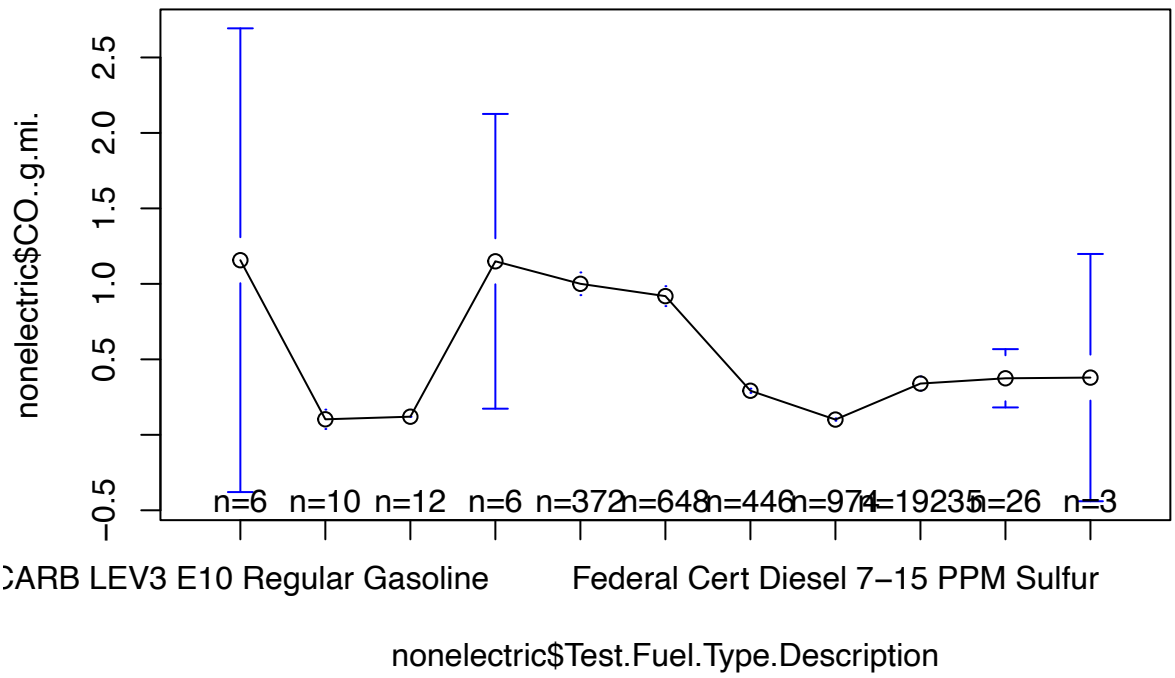



```
#visualize mean CO by vehicle type
plotmeans(nonelectric$CO..g.mi. ~ nonelectric$Test.Fuel.Type.Description)
```

```
## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```

```
## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```

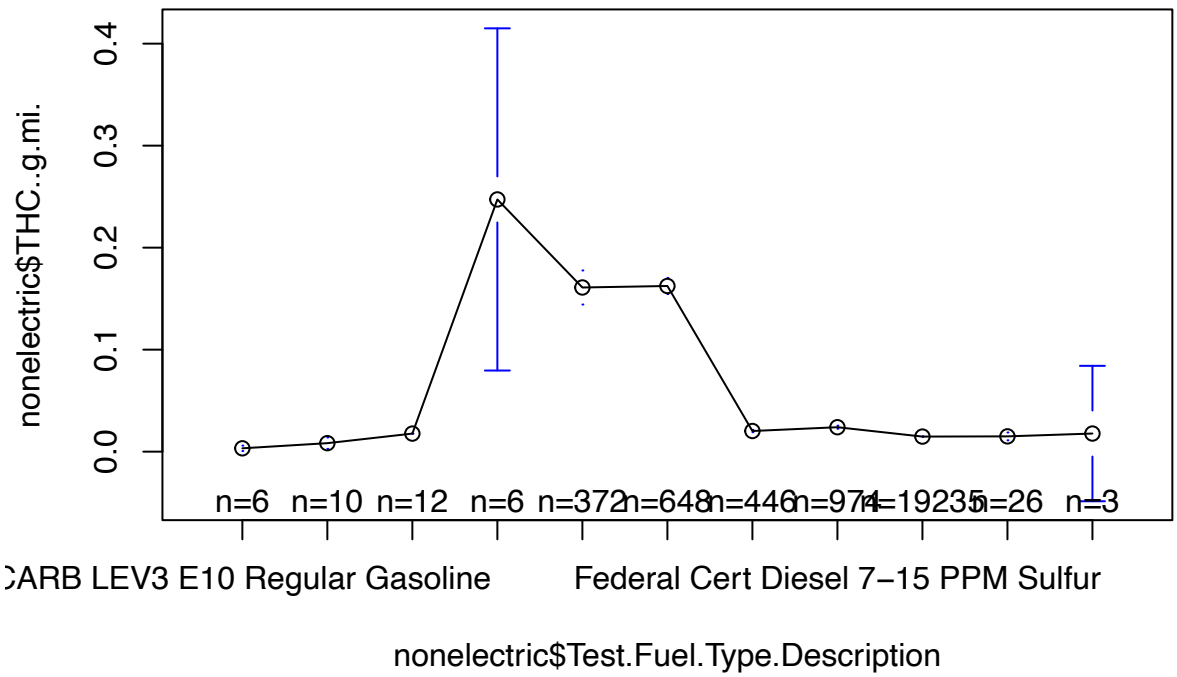
```
## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```



```r
#visualize mean THC by vehicle type
plotmeans(nonelectric$THC..g.mi. ~ nonelectric$Test.Fuel.Type.Description)
```

```
## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
```

```
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```



From the above three plots, we can see that the mean CO2, CO, and THC varies quite a bit by fuel types. This matches the results from our MANOVA, which indicates that there is statistically significant difference in three emissions based on fuel types.

Chi Squared Testing for Project

2022-12-04

Read in Data

```
# Read in electric car data
e_df <- read.csv('../data/cardata_electric_clean.csv')

# Read in non-electric car data
ne_df <- read.csv('../data/cardata_nonelectric_clean.csv')

(nrow(e_df))
```

```
## [1] 878
```

```
(nrow(ne_df))
```

```
## [1] 21738
```

```
head(e_df)
```

```
##   X Model.Year Vehicle.Manufacturer.Name Veh.Mfr.Code Represented.Test.Veh.Make
## 1 1       2018                      BMW          BMX                       BMW
## 2 2       2018                      BMW          BMX                       BMW
## 3 3       2018                      BMW          BMX                       BMW
## 4 4       2018                      BMW          BMX                       BMW
## 5 5       2018                      BMW          BMX                       BMW
## 6 6       2018                      BMW          BMX                       BMW
##   Represented.Test.Veh.Model Test.Veh.Displacement..L. Vehicle.Type
## 1                      330e                         2          Car
## 2                      330e                         2          Car
## 3                      330e                         2          Car
## 4                      330e                         2          Car
## 5                      530e                         2          Car
## 6                      530e                         2          Car
##   Rated.Horsepower Tested.Transmission.Type.Code Tested.Transmission.Type
## 1              180                           SA          Semi-Automatic
## 2              180                           SA          Semi-Automatic
## 3              180                           SA          Semi-Automatic
## 4              180                           SA          Semi-Automatic
## 5              180                           SA          Semi-Automatic
## 6              180                           SA          Semi-Automatic
##   X..of.Gears Transmission.Lockup. Drive.System.Code Drive.System.Description
```

```
## 1           8                    Y                 R      2-Wheel Drive, Rear
## 2           8                    Y                 R      2-Wheel Drive, Rear
## 3           8                    Y                 R      2-Wheel Drive, Rear
## 4           8                    Y                 R      2-Wheel Drive, Rear
## 5           8                    Y                 R      2-Wheel Drive, Rear
## 6           8                    Y                 R      2-Wheel Drive, Rear
##   Equivalent.Test.Weight..lbs.. Axle.Ratio N.V.Ratio Test.Fuel.Type.Description
## 1                          4250       2.93      26.0                Electricity
## 2                          4250       2.93      26.0                Electricity
## 3                          4250       2.93      26.0                Electricity
## 4                          4250       2.93      26.0                Electricity
## 5                          4500       3.23      26.6                Electricity
## 6                          4500       3.23      26.6                Electricity
##   CO2..g.mi. RND_ADJ_FE Target.Coef.A..lbf. Target.Coef.B..lbf.mph.
## 1         NA        0.0                52.9                  -0.113
## 2         NA        0.0                52.9                  -0.113
## 3         NA        0.0                44.9                  -0.063
## 4         NA        0.0                44.9                  -0.063
## 5         NA      122.8                51.1                  -0.114
## 6         NA      122.8                51.1                  -0.114
##   Target.Coef.C..lbf.mph..2. Set.Coef.A..lbf. Set.Coef.B..lbf.mph.
## 1                    0.01826             21.2                0.056
## 2                    0.01826             21.2                0.056
## 3                    0.01831             13.0                0.128
## 4                    0.01831             13.0                0.128
## 5                    0.02015             12.1                0.305
## 6                    0.02015             12.1                0.305
##   Set.Coef.C..lbf.mph..2. Police...Emergency.Vehicle. Averaging.Method.Cd
## 1                 0.01632                           N                   N
## 2                 0.01632                           N                   N
## 3                 0.01611                           N                   N
## 4                 0.01611                           N                   N
## 5                 0.01540                           N                   N
## 6                 0.01540                           N                   N
##   Averging.Method.Desc
## 1         No averaging
## 2         No averaging
## 3         No averaging
## 4         No averaging
## 5         No averaging
## 6         No averaging
```

```
head(ne_df)
```

```
##   X Model.Year Vehicle.Manufacturer.Name Veh.Mfr.Code Represented.Test.Veh.Make
## 1 1       2018              aston martin          ASX              Aston Martin
## 2 2       2018              aston martin          ASX              Aston Martin
## 3 3       2018              aston martin          ASX              Aston Martin
## 4 4       2018              aston martin          ASX              Aston Martin
## 5 5       2018              aston martin          ASX              Aston Martin
## 6 6       2018              aston martin          ASX              Aston Martin
##   Represented.Test.Veh.Model Test.Veh.Displacement..L. Vehicle.Type
## 1                       DB11                        5.2          Car
## 2                       DB11                        5.2          Car
```

```
## 3                   DB11 V8                      4.0          Car
## 4                   DB11 V8                      4.0          Car
## 5                  Rapide S                      6.0          Car
## 6                  Rapide S                      6.0          Car
##   Rated.Horsepower X..of.Cylinders.and.Rotors Tested.Transmission.Type.Code
## 1              600                        12                           SA
## 2              600                        12                           SA
## 3              503                         8                           SA
## 4              503                         8                           SA
## 5              552                        12                           SA
## 6              552                        12                           SA
##   Tested.Transmission.Type X..of.Gears Transmission.Lockup. Drive.System.Code
## 1          Semi-Automatic           8                    Y                 R
## 2          Semi-Automatic           8                    Y                 R
## 3          Semi-Automatic           8                    Y                 R
## 4          Semi-Automatic           8                    Y                 R
## 5          Semi-Automatic           8                    Y                 R
## 6          Semi-Automatic           8                    Y                 R
##   Drive.System.Description Equivalent.Test.Weight..lbs.. Axle.Ratio N.V.Ratio
## 1     2-Wheel Drive, Rear                          4500       2.70      22.2
## 2     2-Wheel Drive, Rear                          4500       2.70      22.2
## 3     2-Wheel Drive, Rear                          4500       2.70      22.2
## 4     2-Wheel Drive, Rear                          4500       2.70      22.2
## 5     2-Wheel Drive, Rear                          4750       2.73      22.4
## 6     2-Wheel Drive, Rear                          4750       2.73      22.4
##   Test.Fuel.Type.Description THC..g.mi. CO..g.mi. CO2..g.mi. RND_ADJ_FE
## 1      Tier 2 Cert Gasoline   0.024700  0.418000     466.87       18.8
## 2      Tier 2 Cert Gasoline   0.001155  0.067334     285.00       30.9
## 3      Tier 2 Cert Gasoline   0.026500  0.070000     386.66       22.7
## 4      Tier 2 Cert Gasoline   0.000500  0.030000     259.74       33.8
## 5      Tier 2 Cert Gasoline   0.026900  0.500000     511.93       17.3
## 6      Tier 2 Cert Gasoline   0.000800  0.060000     296.63       29.9
##   DT.Inertia.Work.Ratio.Rating DT.Absolute.Speed.Change.Ratg
## 1                   -2.5300000                    -1.7300000
## 2                    1.3600000                     0.4400000
## 3                  -11.9900000                    -9.2600000
## 4                   -3.6400000                    -3.2100000
## 5                    0.5655838                     0.4420405
## 6                    0.5655838                     0.4420405
##   DT.Energy.Economy.Rating Target.Coef.A..lbf. Target.Coef.B..lbf.mph.
## 1               -1.7100000               40.94                  0.0169
## 2               -0.5900000               40.94                  0.0169
## 3               -7.7100000               40.94                  0.0169
## 4               -0.9600000               40.94                  0.0169
## 5               -0.2002973               32.66                  0.6085
## 6               -0.2002973               32.66                  0.6085
##   Target.Coef.C..lbf.mph..2. Set.Coef.A..lbf. Set.Coef.B..lbf.mph.
## 1                     0.0271           6.810               0.0807
## 2                     0.0271           6.810               0.0807
## 3                     0.0271          11.260               0.0919
## 4                     0.0271          11.260               0.0919
## 5                     0.0198           1.093               2.1980
## 6                     0.0198           1.093               2.1980
##   Set.Coef.C..lbf.mph..2. Aftertreatment.Device.Cd Aftertreatment.Device.Desc
```

```
## 1                  0.0245                       TWC         Three-way catalyst
## 2                  0.0245                       TWC         Three-way catalyst
## 3                  0.0251                       TWC         Three-way catalyst
## 4                  0.0251                       TWC         Three-way catalyst
## 5                  0.0280                       TWC         Three-way catalyst
## 6                  0.0280                       TWC         Three-way catalyst
##   Police...Emergency.Vehicle. Averaging.Method.Cd Averging.Method.Desc
## 1                           N                   N         No averaging
## 2                           N                   N         No averaging
## 3                           N                   N         No averaging
## 4                           N                   N         No averaging
## 5                           N                   N         No averaging
## 6                           N                   N         No averaging
```

```
# Create color palettes
Blues <- colorRampPalette(c("#0A146B", "#A9A3DA"))
Purples <- colorRampPalette(c("#3E1370", "#BDA3DA"))
GrBuPuPi <- c("#095826", "#0E7032", "#10913F", "#55A472", "#8CBF9E", "#8CBFB8",
              "#63B7AC", "#2D9A8B", "#137568", "#094E45", "#0B3C5C", "#17547C",
              "#2671A4", "#3C8CC1", "#72B1DB", "#96C3E1", "#B0CDE1", "#B0B3E1",
              "#858ACD", "#4F55AB", "#1923B3", "#0E1468", "#3C1075", "#5821A1",
              "#6B27C4", "#9455E5", "#A278D8", "#A990CA", "#ADA0BF", "#C1A5CB",
              "#B887CA", "#A35CBD", "#762594")
```

EDA

To formulate our hypotheses, we first perform EDA on the dataset

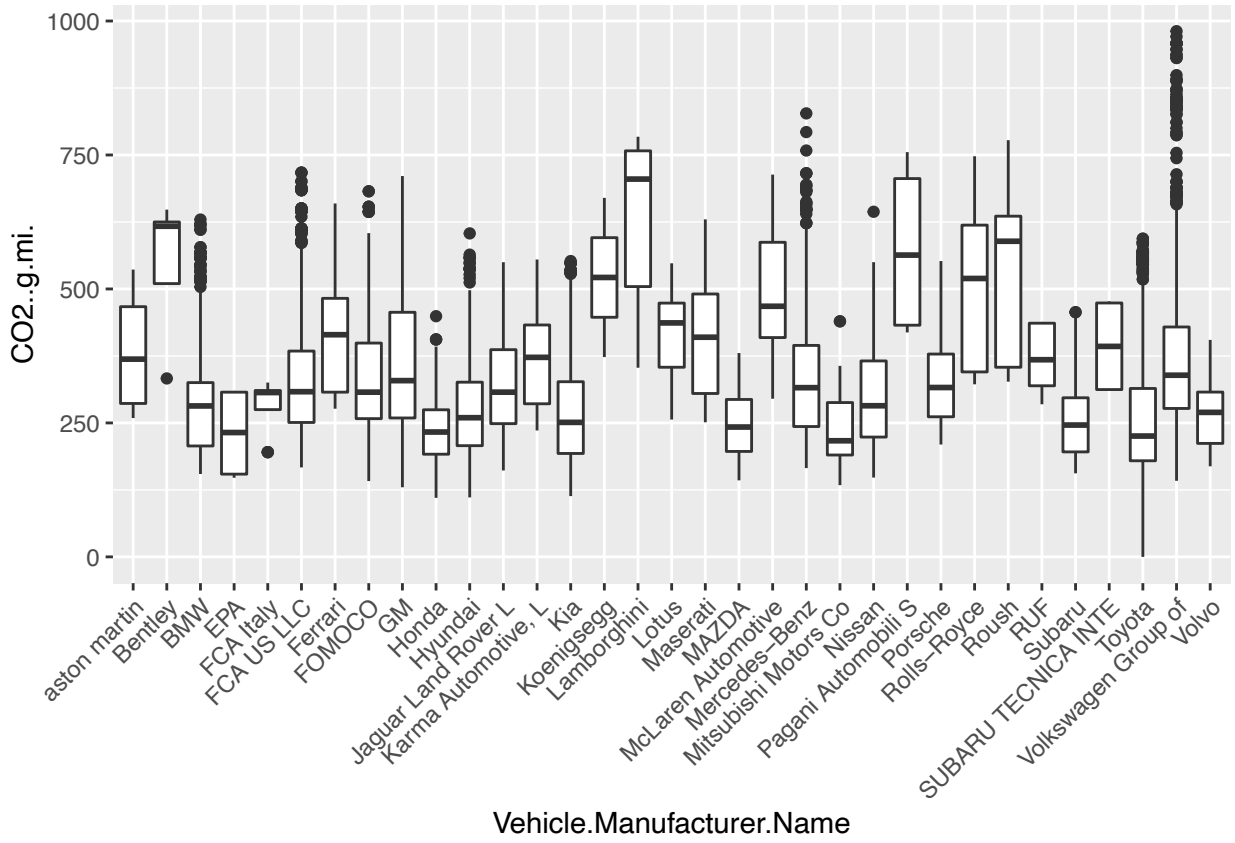
```
#Vehicle.Manufacturer.Name
#CO2..g.mi.
library(ggplot2)
library(tidyr)

# Drop NAs for the emissions column
ne_df <- ne_df %>% drop_na(CO2..g.mi.)

ne_df$emission_cat[ne_df$CO2..g.mi. < 250] <- "low"
ne_df$emission_cat[ne_df$CO2..g.mi. >= 250 & ne_df$CO2..g.mi. < 500] <- "medium"
ne_df$emission_cat[ne_df$CO2..g.mi. >= 500] <- "high"

head(ne_df)
```

```
##   X Model.Year Vehicle.Manufacturer.Name Veh.Mfr.Code Represented.Test.Veh.Make
## 1 1       2018              aston martin          ASX              Aston Martin
## 2 2       2018              aston martin          ASX              Aston Martin
## 3 3       2018              aston martin          ASX              Aston Martin
## 4 4       2018              aston martin          ASX              Aston Martin
## 5 5       2018              aston martin          ASX              Aston Martin
## 6 6       2018              aston martin          ASX              Aston Martin
##   Represented.Test.Veh.Model Test.Veh.Displacement..L. Vehicle.Type
## 1                       DB11                       5.2          Car
## 2                       DB11                       5.2          Car
```

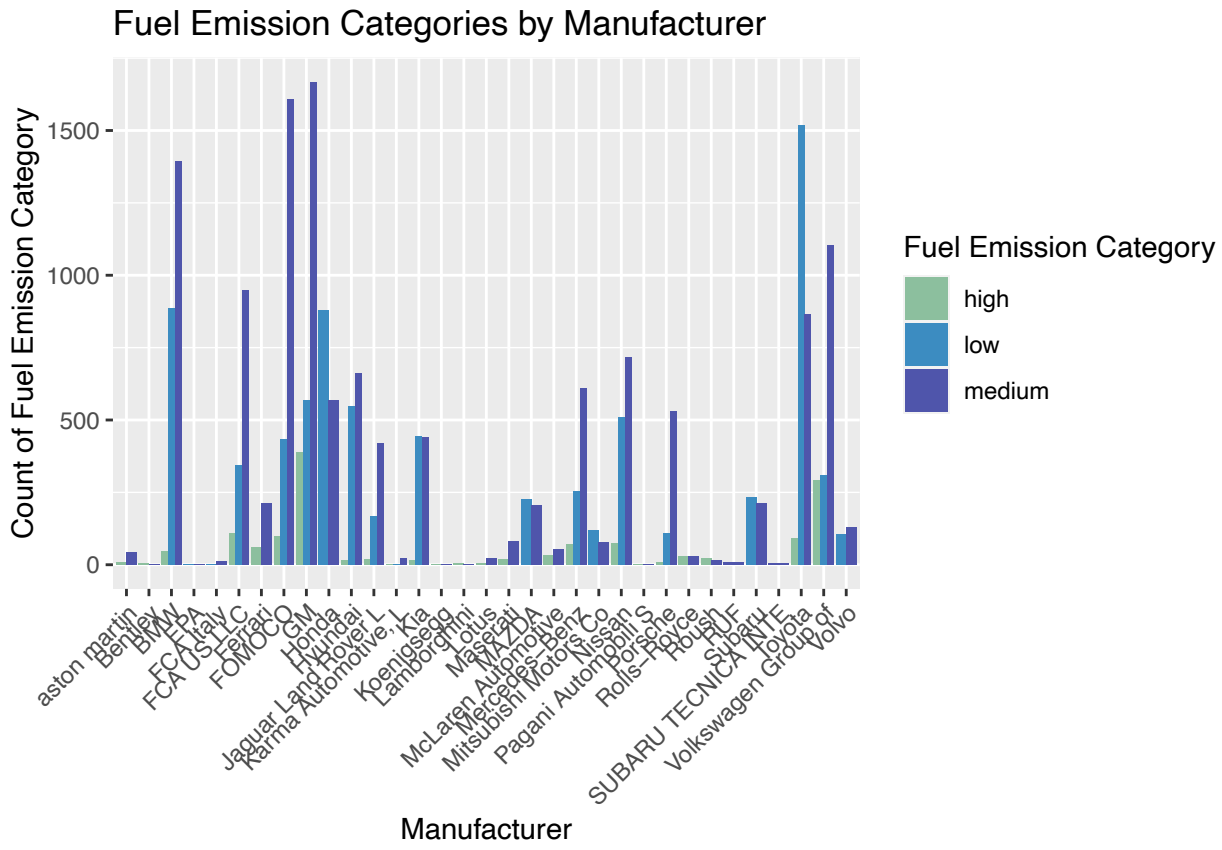
```
## 3                   DB11 V8                      4.0          Car
## 4                   DB11 V8                      4.0          Car
## 5                  Rapide S                      6.0          Car
## 6                  Rapide S                      6.0          Car
##   Rated.Horsepower X..of.Cylinders.and.Rotors Tested.Transmission.Type.Code
## 1              600                        12                           SA
## 2              600                        12                           SA
## 3              503                         8                           SA
## 4              503                         8                           SA
## 5              552                        12                           SA
## 6              552                        12                           SA
##   Tested.Transmission.Type X..of.Gears Transmission.Lockup. Drive.System.Code
## 1          Semi-Automatic           8                    Y                 R
## 2          Semi-Automatic           8                    Y                 R
## 3          Semi-Automatic           8                    Y                 R
## 4          Semi-Automatic           8                    Y                 R
## 5          Semi-Automatic           8                    Y                 R
## 6          Semi-Automatic           8                    Y                 R
##   Drive.System.Description Equivalent.Test.Weight..lbs.. Axle.Ratio N.V.Ratio
## 1     2-Wheel Drive, Rear                          4500       2.70      22.2
## 2     2-Wheel Drive, Rear                          4500       2.70      22.2
## 3     2-Wheel Drive, Rear                          4500       2.70      22.2
## 4     2-Wheel Drive, Rear                          4500       2.70      22.2
## 5     2-Wheel Drive, Rear                          4750       2.73      22.4
## 6     2-Wheel Drive, Rear                          4750       2.73      22.4
##   Test.Fuel.Type.Description THC..g.mi. CO..g.mi. CO2..g.mi. RND_ADJ_FE
## 1       Tier 2 Cert Gasoline   0.024700  0.418000     466.87       18.8
## 2       Tier 2 Cert Gasoline   0.001155  0.067334     285.00       30.9
## 3       Tier 2 Cert Gasoline   0.026500  0.070000     386.66       22.7
## 4       Tier 2 Cert Gasoline   0.000500  0.030000     259.74       33.8
## 5       Tier 2 Cert Gasoline   0.026900  0.500000     511.93       17.3
## 6       Tier 2 Cert Gasoline   0.000800  0.060000     296.63       29.9
##   DT.Inertia.Work.Ratio.Rating DT.Absolute.Speed.Change.Ratg
## 1                   -2.5300000                    -1.7300000
## 2                    1.3600000                     0.4400000
## 3                  -11.9900000                    -9.2600000
## 4                   -3.6400000                    -3.2100000
## 5                    0.5655838                     0.4420405
## 6                    0.5655838                     0.4420405
##   DT.Energy.Economy.Rating Target.Coef.A..lbf. Target.Coef.B..lbf.mph.
## 1               -1.7100000               40.94                  0.0169
## 2               -0.5900000               40.94                  0.0169
## 3               -7.7100000               40.94                  0.0169
## 4               -0.9600000               40.94                  0.0169
## 5               -0.2002973               32.66                  0.6085
## 6               -0.2002973               32.66                  0.6085
##   Target.Coef.C..lbf.mph..2. Set.Coef.A..lbf. Set.Coef.B..lbf.mph.
## 1                     0.0271           6.810               0.0807
## 2                     0.0271           6.810               0.0807
## 3                     0.0271          11.260               0.0919
## 4                     0.0271          11.260               0.0919
## 5                     0.0198           1.093               2.1980
## 6                     0.0198           1.093               2.1980
##   Set.Coef.C..lbf.mph..2. Aftertreatment.Device.Cd Aftertreatment.Device.Desc
```

```
## 1                   0.0245                       TWC         Three-way catalyst
## 2                   0.0245                       TWC         Three-way catalyst
## 3                   0.0251                       TWC         Three-way catalyst
## 4                   0.0251                       TWC         Three-way catalyst
## 5                   0.0280                       TWC         Three-way catalyst
## 6                   0.0280                       TWC         Three-way catalyst
##   Police...Emergency.Vehicle. Averaging.Method.Cd Averging.Method.Desc
## 1                           N                    N         No averaging
## 2                           N                    N         No averaging
## 3                           N                    N         No averaging
## 4                           N                    N         No averaging
## 5                           N                    N         No averaging
## 6                           N                    N         No averaging
##   emission_cat
## 1       medium
## 2       medium
## 3       medium
## 4       medium
## 5         high
## 6       medium
```

```
# Make bar plot of transmission type and CO2 emissions
ggplot(data=ne_df, aes(x=Vehicle.Manufacturer.Name,y=CO2..g.mi.))+geom_boxplot()+ scale_x_discrete(guid
```

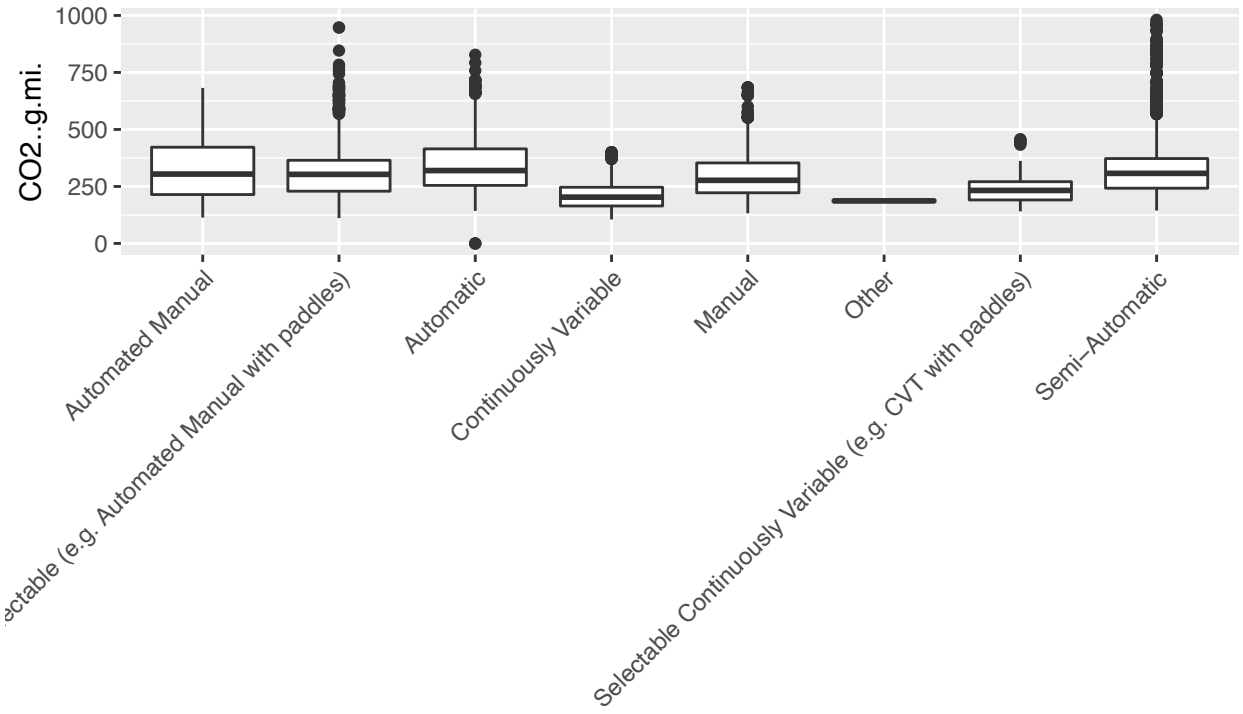


```
# make barplot of emissions categories

ggplot(ne_df, aes(x=Vehicle.Manufacturer.Name, fill=emission_cat)) + geom_bar(position="dodge")+ scale_
```

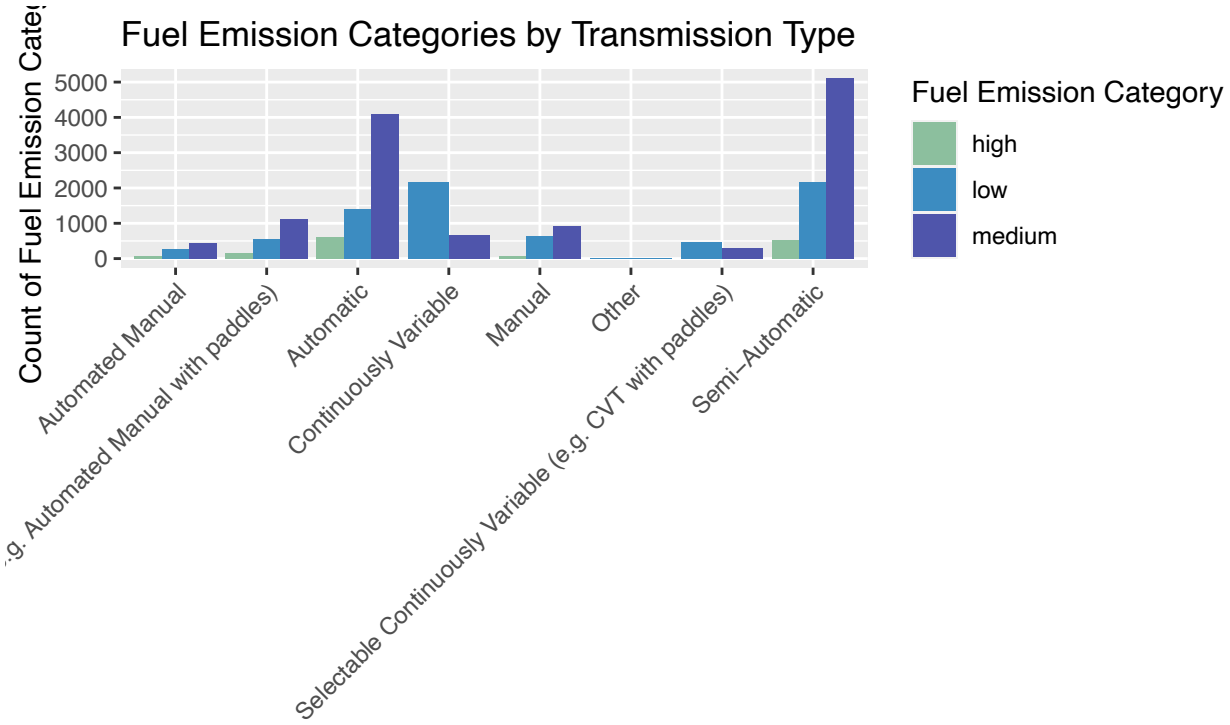


We definitely see some car manufacturers have higher average fuel emissions. For instance, Lamborghini, Bentley, and Rolls-Royce have higher emissions likely due to them being luxury brands. Honda and Mitsubishi, on the other hand, are more affordable brands and have lower average emissions.

```
# Make bar plot of transmission type and CO2 emissions
ggplot(data=ne_df, aes(x=Tested.Transmission.Type,y=CO2..g.mi., fill=CO2..g.mi.))+geom_boxplot()+ scale_
```



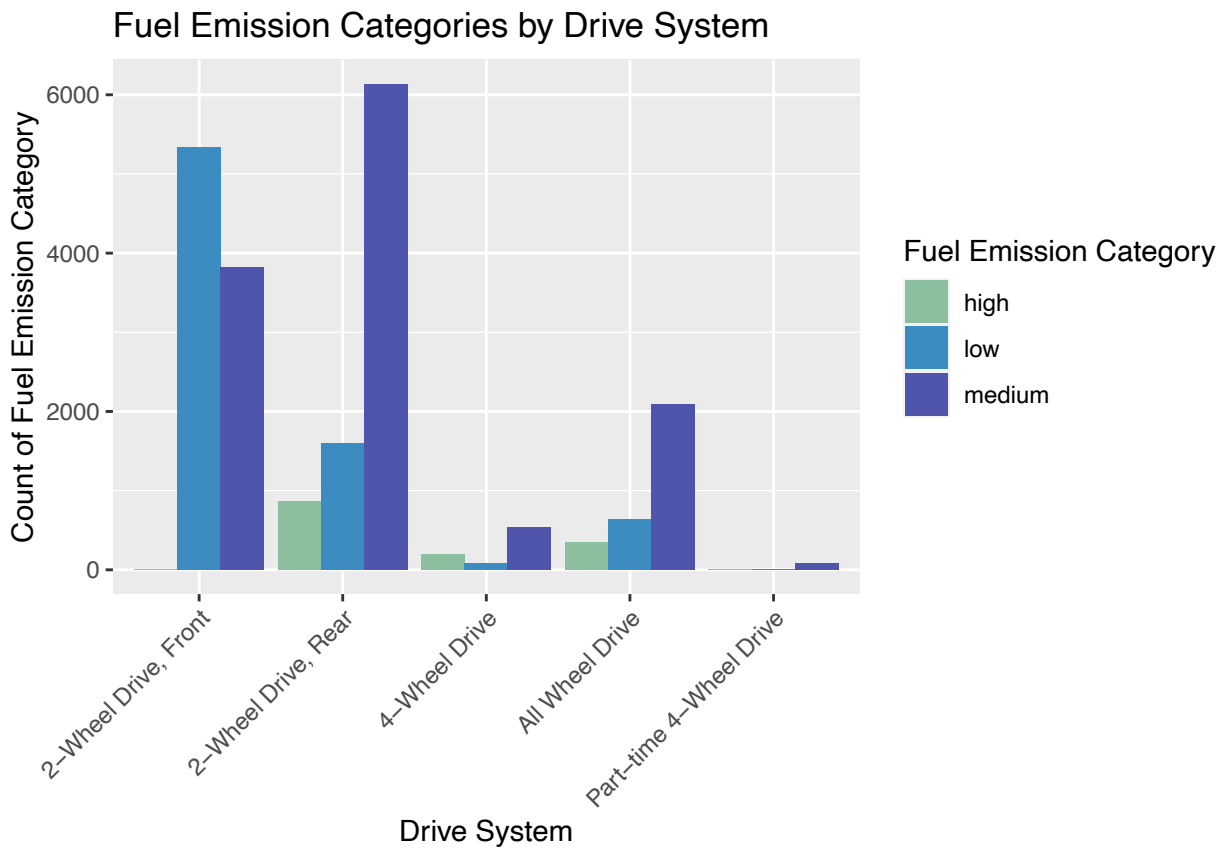

```
ggplot(ne_df, aes(x=Tested.Transmission.Type, fill=emission_cat)) + geom_bar(position="dodge")+ scale_x
```



Interestingly, it appears like the high emissions cars are mostly within the automatic and semi-automatic categories. Manual tends to have mostly low and medium with very few high emissions cars.

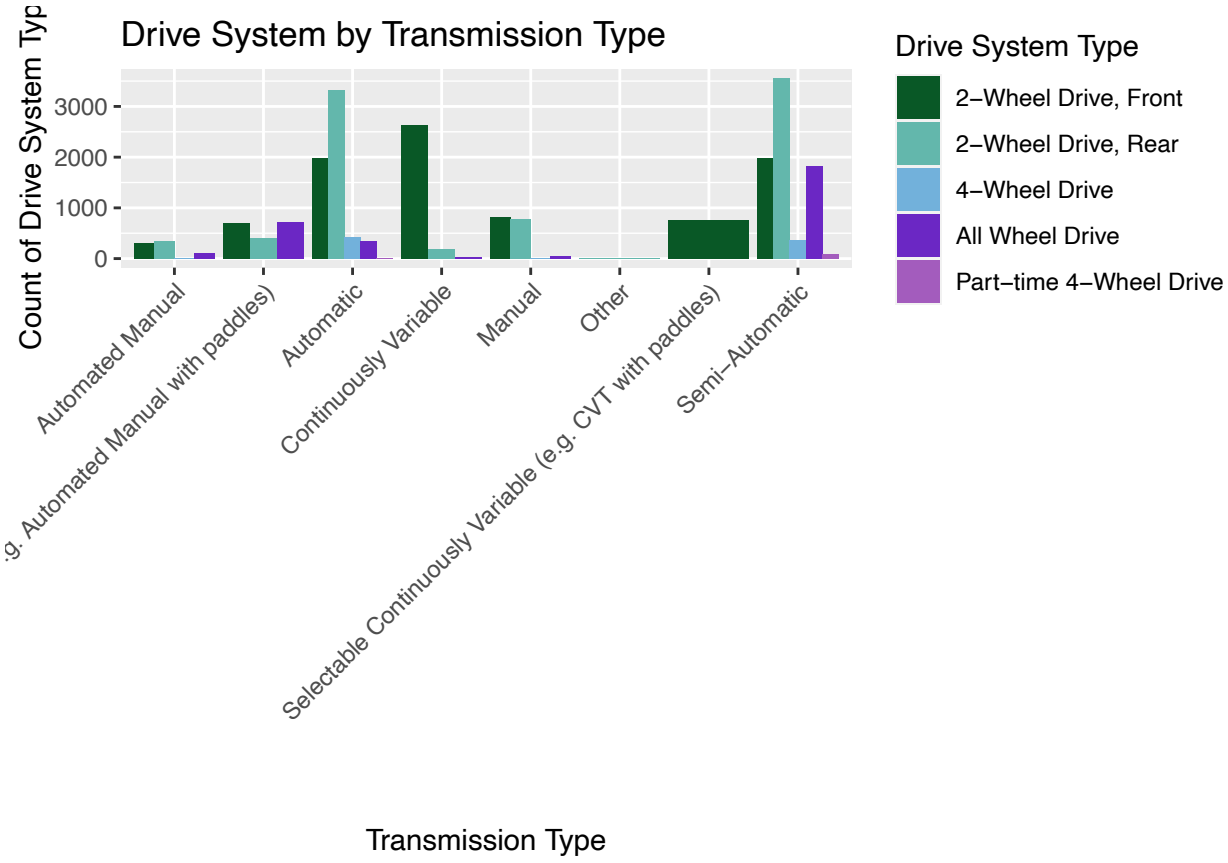
Drive system vs Emission category. Looks like 2 wheel drive front does not have high emissions compared to the 4 wheel drive and all wheel drive. Perhaps fuel emissions increase the more the wheel drive increases.

```
ggplot(ne_df, aes(x=Drive.System.Description, fill=emission_cat)) + geom_bar(position="dodge")+ scale_x
```



Next, check observe the transmission type vs the drive system type

```
ggplot(ne_df, aes(fill=Drive.System.Description, x=Tested.Transmission.Type)) + geom_bar(position="dodg
```



Define Hypotheses

Question 1: Is there a relationship between car brand and fuel emission level?

Null hypothesis: Car Brand and fuel emissions level are independent. The fuel emissions level does not depend on the car brand

Alternative Hypothesis: Car Brand and fuel emissions level are dependent. The fuel emissions level does depend on the car brand

Question 2: Is there a relationship between transmission type and fuel emission level?

Null hypothesis: Transmission type and fuel emissions level are independent. The fuel emissions level does not depend on the transmission type

Alternative Hypothesis: Transmission type and fuel emissions level are dependent. The fuel emissions level does depend on the Transmission type

Question 3: Is there a relationship between transmission type and the drive system type?

Null hypothesis: Transmission type and drive system type are independent. The transmission type does not depend on the drive system type

Alternative Hypothesis: Transmission type and drive system type are dependent. The transmission type does depend on the drive system type

Question 4: Is there a relationship between drive system and the fuel emission level type?

Null hypothesis: Drive system and fuel emissions level are independent. The fuel emissions level does not depend on the drive system

Alternative Hypothesis: Drive System and fuel emissions level are dependent. The fuel emissions level does depend on the drive system

Hypothesis Testing

```
head(ne_df)
```

```
##   X Model.Year Vehicle.Manufacturer.Name Veh.Mfr.Code Represented.Test.Veh.Make
## 1 1       2018              aston martin          ASX              Aston Martin
## 2 2       2018              aston martin          ASX              Aston Martin
## 3 3       2018              aston martin          ASX              Aston Martin
## 4 4       2018              aston martin          ASX              Aston Martin
## 5 5       2018              aston martin          ASX              Aston Martin
## 6 6       2018              aston martin          ASX              Aston Martin
##   Represented.Test.Veh.Model Test.Veh.Displacement..L. Vehicle.Type
## 1                      DB11                       5.2          Car
## 2                      DB11                       5.2          Car
## 3                   DB11 V8                       4.0          Car
## 4                   DB11 V8                       4.0          Car
## 5                  Rapide S                       6.0          Car
## 6                  Rapide S                       6.0          Car
##   Rated.Horsepower X..of.Cylinders.and.Rotors Tested.Transmission.Type.Code
## 1              600                         12                            SA
## 2              600                         12                            SA
## 3              503                          8                            SA
## 4              503                          8                            SA
## 5              552                         12                            SA
## 6              552                         12                            SA
##   Tested.Transmission.Type X..of.Gears Transmission.Lockup. Drive.System.Code
## 1           Semi-Automatic           8                    Y                 R
## 2           Semi-Automatic           8                    Y                 R
## 3           Semi-Automatic           8                    Y                 R
## 4           Semi-Automatic           8                    Y                 R
## 5           Semi-Automatic           8                    Y                 R
## 6           Semi-Automatic           8                    Y                 R
##   Drive.System.Description Equivalent.Test.Weight..lbs.. Axle.Ratio N.V.Ratio
## 1      2-Wheel Drive, Rear                           4500       2.70      22.2
## 2      2-Wheel Drive, Rear                           4500       2.70      22.2
## 3      2-Wheel Drive, Rear                           4500       2.70      22.2
## 4      2-Wheel Drive, Rear                           4500       2.70      22.2
## 5      2-Wheel Drive, Rear                           4750       2.73      22.4
## 6      2-Wheel Drive, Rear                           4750       2.73      22.4
##   Test.Fuel.Type.Description THC..g.mi. CO..g.mi. CO2..g.mi. RND_ADJ_FE
## 1       Tier 2 Cert Gasoline   0.024700  0.418000     466.87       18.8
## 2       Tier 2 Cert Gasoline   0.001155  0.067334     285.00       30.9
## 3       Tier 2 Cert Gasoline   0.026500  0.070000     386.66       22.7
## 4       Tier 2 Cert Gasoline   0.000500  0.030000     259.74       33.8
## 5       Tier 2 Cert Gasoline   0.026900  0.500000     511.93       17.3
## 6       Tier 2 Cert Gasoline   0.000800  0.060000     296.63       29.9
##   DT.Inertia.Work.Ratio.Rating DT.Absolute.Speed.Change.Ratg
## 1                   -2.5300000                   -1.7300000
```

```
## 2                    1.3600000                    0.4400000
## 3                  -11.9900000                   -9.2600000
## 4                   -3.6400000                   -3.2100000
## 5                    0.5655838                    0.4420405
## 6                    0.5655838                    0.4420405
##   DT.Energy.Economy.Rating Target.Coef.A..lbf. Target.Coef.B..lbf.mph.
## 1               -1.7100000               40.94                  0.0169
## 2               -0.5900000               40.94                  0.0169
## 3               -7.7100000               40.94                  0.0169
## 4               -0.9600000               40.94                  0.0169
## 5               -0.2002973               32.66                  0.6085
## 6               -0.2002973               32.66                  0.6085
##   Target.Coef.C..lbf.mph..2. Set.Coef.A..lbf. Set.Coef.B..lbf.mph.
## 1                     0.0271           6.810               0.0807
## 2                     0.0271           6.810               0.0807
## 3                     0.0271          11.260               0.0919
## 4                     0.0271          11.260               0.0919
## 5                     0.0198           1.093               2.1980
## 6                     0.0198           1.093               2.1980
##   Set.Coef.C..lbf.mph..2. Aftertreatment.Device.Cd Aftertreatment.Device.Desc
## 1                  0.0245                      TWC         Three-way catalyst
## 2                  0.0245                      TWC         Three-way catalyst
## 3                  0.0251                      TWC         Three-way catalyst
## 4                  0.0251                      TWC         Three-way catalyst
## 5                  0.0280                      TWC         Three-way catalyst
## 6                  0.0280                      TWC         Three-way catalyst
##   Police...Emergency.Vehicle. Averaging.Method.Cd Averging.Method.Desc
## 1                           N                   N         No averaging
## 2                           N                   N         No averaging
## 3                           N                   N         No averaging
## 4                           N                   N         No averaging
## 5                           N                   N         No averaging
## 6                           N                   N         No averaging
##   emission_cat
## 1       medium
## 2       medium
## 3       medium
## 4       medium
## 5         high
## 6       medium
```

Question 1

Question 1: Is there a relationship between car brand and fuel emission level?

Null hypothesis: Car Brand and fuel emissions level are independent. The fuel emissions level does not depend on the car brand

Alternative Hypothesis: Car Brand and fuel emissions level are dependent. The fuel emissions level does depend on the car brand

```
# Make contingency table
cont <- table(ne_df$Vehicle.Manufacturer.Name, ne_df$emission_cat)
cont
```

```
##
##                         high  low medium
##   aston martin             8    0     42
##   Bentley                  4    0      1
##   BMW                     46  886   1395
##   EPA                      0    2      2
##   FCA Italy                0    3     12
##   FCA US LLC             109  342    948
##   Ferrari                 61    0    211
##   FOMOCO                  98  434   1608
##   GM                     387  569   1668
##   Honda                    0  878    567
##   Hyundai                 15  548    661
##   Jaguar Land Rover L     20  167    421
##   Karma Automotive, L      2    2     24
##   Kia                     17  443    439
##   Koenigsegg               1    0      1
##   Lamborghini              5    0      2
##   Lotus                    6    0     24
##   Maserati                20    0     80
##   MAZDA                    0  226    204
##   McLaren Automotive      33    0     52
##   Mercedes-Benz           71  255    609
##   Mitsubishi Motors Co     0  118     77
##   Nissan                  73  510    716
##   Pagani Automobili S      2    0      2
##   Porsche                  7  109    529
##   Rolls-Royce             30    0     30
##   Roush                   23    0     14
##   RUF                      0    0      8
##   Subaru                   0  234    212
##   SUBARU TECNICA INTE      0    0      4
##   Toyota                  91 1517    864
##   Volkswagen Group of    292  309   1102
##   Volvo                    0  106    130
```

```
chisq.test(cont)
```

```
## Warning in chisq.test(cont): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont
## X-squared = 4126, df = 64, p-value < 2.2e-16
```

Repeat test with the Yates correction

```
chisq.test(cont, correct = TRUE)
```

```
## Warning in chisq.test(cont, correct = TRUE): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont
## X-squared = 4126, df = 64, p-value < 2.2e-16
```

Since the Yates correction was not enough, we can switch over to Fisher's Exact Test

```
fisher.test(cont, simulate.p.value=TRUE)
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  cont
## p-value = 0.0004998
## alternative hypothesis: two.sided
```

Question 2

Question 2: Is there a relationship between transmission type and fuel emission level?

Null hypothesis: Transmission type and fuel emissions level are independent. The fuel emissions level does not depend on the transmission type

Alternative Hypothesis: Transmission type and fuel emissions level are dependent. The fuel emissions level does depend on the Transmission type

```
# Make contingency table
cont <- table(ne_df$Tested.Transmission.Type, ne_df$emission_cat)
cont
```

```
##
##                                                                    high  low
##   Automated Manual                                                   79  255
##   Automated Manual- Selectable (e.g. Automated Manual with paddles)  151  553
##   Automatic                                                         603 1411
##   Continuously Variable                                               0 2168
##   Manual                                                             72  634
##   Other                                                               0    4
##   Selectable Continuously Variable (e.g. CVT with paddles)            0  454
##   Semi-Automatic                                                    516 2179
##
##                                                                    medium
##   Automated Manual                                                   430
##   Automated Manual- Selectable (e.g. Automated Manual with paddles)  1115
##   Automatic                                                         4081
##   Continuously Variable                                              671
##   Manual                                                             932
##   Other                                                                0
##   Selectable Continuously Variable (e.g. CVT with paddles)           308
##   Semi-Automatic                                                    5122
```

```
chisq.test(cont)
```

```
## Warning in chisq.test(cont): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont
## X-squared = 3047.7, df = 14, p-value < 2.2e-16
```

Repeat test with the Yates correction

```
chisq.test(cont, correct = TRUE)
```

```
## Warning in chisq.test(cont, correct = TRUE): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont
## X-squared = 3047.7, df = 14, p-value < 2.2e-16
```

Since the Yates correction was not enough, we can switch over to Fisher's Exact Test

```
fisher.test(cont, simulate.p.value=TRUE)
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  cont
## p-value = 0.0004998
## alternative hypothesis: two.sided
```

Question 3

Question 3: Is there a relationship between transmission type and the drive system type?

Null hypothesis: Transmission type and drive system type are independent. The transmission type does not depend on the drive system type

Alternative Hypothesis: Transmission type and drive system type are dependent. The transmission type does depend on the drive system type

```
# Make contingency table
cont <- table(ne_df$Tested.Transmission.Type, ne_df$Drive.System.Description)
cont
```

```
##
##                                                                                 2-Wheel Drive, Front
##   Automated Manual                                                                               315
##   Automated Manual- Selectable (e.g. Automated Manual with paddles)                              700
##   Automatic                                                                                     1974
##   Continuously Variable                                                                         2623
##   Manual                                                                                         809
##   Other                                                                                            0
##   Selectable Continuously Variable (e.g. CVT with paddles)                                       762
##   Semi-Automatic                                                                                1976
##
##                                                                                 2-Wheel Drive, Rear
##   Automated Manual                                                                               336
##   Automated Manual- Selectable (e.g. Automated Manual with paddles)                              404
##   Automatic                                                                                     3330
##   Continuously Variable                                                                          196
##   Manual                                                                                         770
##   Other                                                                                            4
##   Selectable Continuously Variable (e.g. CVT with paddles)                                         0
##   Semi-Automatic                                                                                3565
##
##                                                                                 4-Wheel Drive
##   Automated Manual                                                                                 2
##   Automated Manual- Selectable (e.g. Automated Manual with paddles)                                0
##   Automatic                                                                                      434
##   Continuously Variable                                                                            0
##   Manual                                                                                          12
##   Other                                                                                            0
##   Selectable Continuously Variable (e.g. CVT with paddles)                                         0
##   Semi-Automatic                                                                                 363
##
##                                                                                 All Wheel Drive
##   Automated Manual                                                                               111
##   Automated Manual- Selectable (e.g. Automated Manual with paddles)                              715
##   Automatic                                                                                      355
##   Continuously Variable                                                                           20
##   Manual                                                                                          47
##   Other                                                                                            0
##   Selectable Continuously Variable (e.g. CVT with paddles)                                         0
##   Semi-Automatic                                                                                1830
##
##                                                                                 Part-time 4-Wheel Drive
##   Automated Manual                                                                                  0
##   Automated Manual- Selectable (e.g. Automated Manual with paddles)                                 0
##   Automatic                                                                                         2
##   Continuously Variable                                                                             0
##   Manual                                                                                            0
##   Other                                                                                             0
##   Selectable Continuously Variable (e.g. CVT with paddles)                                          0
##   Semi-Automatic                                                                                   83
```

```
chisq.test(cont)
```

```
## Warning in chisq.test(cont): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont
## X-squared = 7472.1, df = 28, p-value < 2.2e-16
```

Repeat test with the Yates correction

```
chisq.test(cont, correct = TRUE)
```

```
## Warning in chisq.test(cont, correct = TRUE): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont
## X-squared = 7472.1, df = 28, p-value < 2.2e-16
```

Since the Yates correction was not enough, we can switch over to Fisher's Exact Test

```
fisher.test(cont, simulate.p.value=TRUE)
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  cont
## p-value = 0.0004998
## alternative hypothesis: two.sided
```

Question 4: Is there a relationship between drive system and the fuel emission level type?

Null hypothesis: Drive system and fuel emissions level are independent. The fuel emissions level does not depend on the drive system

Alternative Hypothesis: Drive System and fuel emissions level are dependent. The fuel emissions level does depend on the drive system

```
# Make contingency table
cont <- table(ne_df$Drive.System.Description, ne_df$emission_cat)
cont
```

```
##
##                           high  low medium
##   2-Wheel Drive, Front       7 5332   3820
##   2-Wheel Drive, Rear      866 1603   6136
##   4-Wheel Drive            197   78    536
##   All Wheel Drive          350  643   2085
##   Part-time 4-Wheel Drive    1    2     82
```

Repeat test with the Yates correction

```
chisq.test(cont, correct = TRUE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont
## X-squared = 4416.8, df = 8, p-value < 2.2e-16
```

Since the Yates correction was not enough, we can switch over to Fisher's Exact Test

```
fisher.test(cont, simulate.p.value=TRUE)
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  cont
## p-value = 0.0004998
## alternative hypothesis: two.sided
```

ANLY 511 Final Project - T Tests/ Mann-Whitney U Tests

Mia Mayerhofer

2022-11-26

Data Preparation

```
# Load libraries
library(dplyr)
library(tidyverse)
library(RColorBrewer)
library(car)
```

```
# Create color palettes
Blues <- colorRampPalette(c("#0A146B", "#A9A3DA"))
Purples <- colorRampPalette(c("#3E1370", "#BDA3DA"))
GrBuPuPi <- c("#095826", "#0E7032", "#10913F", "#55A472", "#8CBF9E", "#8CBFB8",
              "#63B7AC", "#2D9A8B", "#137568", "#094E45", "#0B3C5C", "#17547C",
              "#2671A4", "#3C8CC1", "#72B1DB", "#96C3E1", "#B0CDE1", "#B0B3E1",
              "#858ACD", "#4F55AB", "#1923B3", "#0E1468", "#3C1075", "#5821A1",
              "#6B27C4", "#9455E5", "#A278D8", "#A990CA", "#ADA0BF", "#C1A5CB",
              "#B887CA", "#A35CBD", "#762594")
```

```
# Load in the cleaned csv data for nonelectric vehicles
gas <- read.csv("../data/cardata_nonelectric_clean.csv")
# Remove index columns
gas <- gas[,-1]
# View the data
head(gas)
```

```
##   Model.Year Vehicle.Manufacturer.Name Veh.Mfr.Code Represented.Test.Veh.Make
## 1       2018              aston martin          ASX              Aston Martin
## 2       2018              aston martin          ASX              Aston Martin
## 3       2018              aston martin          ASX              Aston Martin
## 4       2018              aston martin          ASX              Aston Martin
## 5       2018              aston martin          ASX              Aston Martin
## 6       2018              aston martin          ASX              Aston Martin
##   Represented.Test.Veh.Model Test.Veh.Displacement..L. Vehicle.Type
## 1                       DB11                       5.2          Car
## 2                       DB11                       5.2          Car
## 3                    DB11 V8                       4.0          Car
## 4                    DB11 V8                       4.0          Car
## 5                   Rapide S                       6.0          Car
## 6                   Rapide S                       6.0          Car
```

```
##   Rated.Horsepower X..of.Cylinders.and.Rotors Tested.Transmission.Type.Code
## 1              600                         12                           SA
## 2              600                         12                           SA
## 3              503                          8                           SA
## 4              503                          8                           SA
## 5              552                         12                           SA
## 6              552                         12                           SA
##   Tested.Transmission.Type X..of.Gears Transmission.Lockup. Drive.System.Code
## 1           Semi-Automatic           8                    Y                R
## 2           Semi-Automatic           8                    Y                R
## 3           Semi-Automatic           8                    Y                R
## 4           Semi-Automatic           8                    Y                R
## 5           Semi-Automatic           8                    Y                R
## 6           Semi-Automatic           8                    Y                R
##   Drive.System.Description Equivalent.Test.Weight..lbs.. Axle.Ratio N.V.Ratio
## 1     2-Wheel Drive, Rear                           4500       2.70      22.2
## 2     2-Wheel Drive, Rear                           4500       2.70      22.2
## 3     2-Wheel Drive, Rear                           4500       2.70      22.2
## 4     2-Wheel Drive, Rear                           4500       2.70      22.2
## 5     2-Wheel Drive, Rear                           4750       2.73      22.4
## 6     2-Wheel Drive, Rear                           4750       2.73      22.4
##   Test.Fuel.Type.Description THC..g.mi. CO..g.mi. CO2..g.mi. RND_ADJ_FE
## 1      Tier 2 Cert Gasoline   0.024700  0.418000     466.87       18.8
## 2      Tier 2 Cert Gasoline   0.001155  0.067334     285.00       30.9
## 3      Tier 2 Cert Gasoline   0.026500  0.070000     386.66       22.7
## 4      Tier 2 Cert Gasoline   0.000500  0.030000     259.74       33.8
## 5      Tier 2 Cert Gasoline   0.026900  0.500000     511.93       17.3
## 6      Tier 2 Cert Gasoline   0.000800  0.060000     296.63       29.9
##   DT.Inertia.Work.Ratio.Rating DT.Absolute.Speed.Change.Ratg
## 1                   -2.5300000                   -1.7300000
## 2                    1.3600000                    0.4400000
## 3                  -11.9900000                   -9.2600000
## 4                   -3.6400000                   -3.2100000
## 5                    0.5655838                    0.4420405
## 6                    0.5655838                    0.4420405
##   DT.Energy.Economy.Rating Target.Coef.A..lbf. Target.Coef.B..lbf.mph.
## 1               -1.7100000               40.94                  0.0169
## 2               -0.5900000               40.94                  0.0169
## 3               -7.7100000               40.94                  0.0169
## 4               -0.9600000               40.94                  0.0169
## 5               -0.2002973               32.66                  0.6085
## 6               -0.2002973               32.66                  0.6085
##   Target.Coef.C..lbf.mph..2. Set.Coef.A..lbf. Set.Coef.B..lbf.mph.
## 1                     0.0271           6.810               0.0807
## 2                     0.0271           6.810               0.0807
## 3                     0.0271          11.260               0.0919
## 4                     0.0271          11.260               0.0919
## 5                     0.0198           1.093               2.1980
## 6                     0.0198           1.093               2.1980
##   Set.Coef.C..lbf.mph..2. Aftertreatment.Device.Cd Aftertreatment.Device.Desc
## 1                  0.0245                      TWC         Three-way catalyst
## 2                  0.0245                      TWC         Three-way catalyst
## 3                  0.0251                      TWC         Three-way catalyst
## 4                  0.0251                      TWC         Three-way catalyst
```

```
## 5                  0.0280                       TWC         Three-way catalyst
## 6                  0.0280                       TWC         Three-way catalyst
##   Police...Emergency.Vehicle. Averaging.Method.Cd Averging.Method.Desc
## 1                           N                    N         No averaging
## 2                           N                    N         No averaging
## 3                           N                    N         No averaging
## 4                           N                    N         No averaging
## 5                           N                    N         No averaging
## 6                           N                    N         No averaging
```

Exploratory Data Analysis (EDA) for Test 1

Research Question: Is there a significant difference in the amount of carbon dioxide emissions between types of fuel, specifically the two most common fuel types?

How many observations are their for each type of fuel?

```
# Create a frequency table
frequencies <- data.frame(cbind(table(gas$`Fuel Type`)))
frequencies$`Fuel Type` <- row.names(frequencies)
frequencies$`Frequency` <- frequencies$cbind.table.gas..Fuel.Type...
frequencies <- frequencies %>% dplyr::select("Fuel Type", "Frequency")
rownames(frequencies) <- NULL
# Print table ordered by frequency
frequencies[order(frequencies$Frequency, decreasing = TRUE),]
```

```
## [1] Fuel Type Frequency
## <0 rows> (or 0-length row.names)
```

From the frequency table above, Tier 2 Cert Gasoline and Federal Cert Diesel 7-15 PPM Sulfur are the two most common gasoline types in the data set with 19,235 and 974 observations respectively.

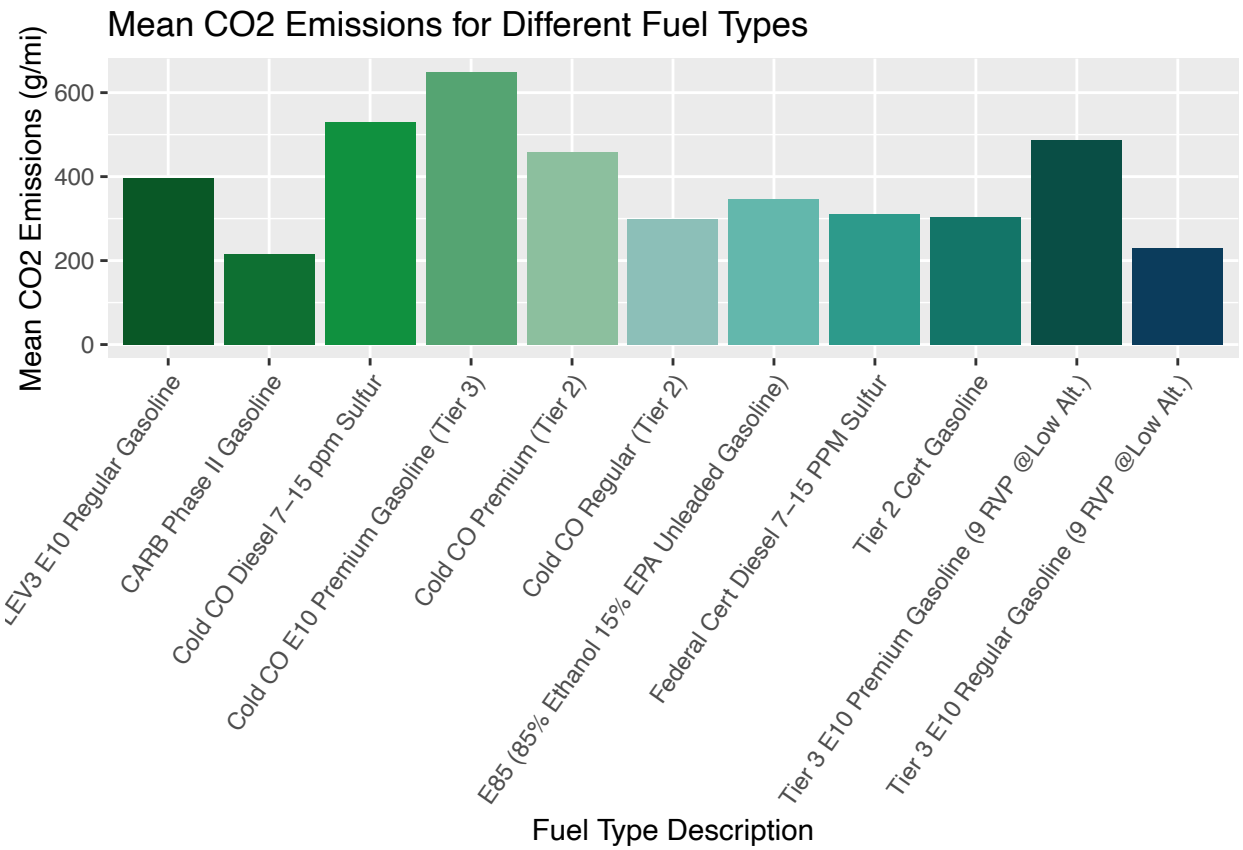
Which fuel type produces the most carbon dioxide emissions in this data set?

```
# Calculate the mean CO2 emissions for each fuel type
means_fuel <- gas %>% group_by(Test.Fuel.Type.Description) %>%
  summarise_at(vars(CO2..g.mi.), list(name = mean))
colnames(means_fuel) <- c("Fuel Type", "Mean CO2 Emissions")
# Print means ordered by mean
print(means_fuel[order(means_fuel$`Mean CO2 Emissions`, decreasing = TRUE),])
```

```
## # A tibble: 11 x 2
##    `Fuel Type`                                  `Mean CO2 Emissions`
##    <chr>                                                       <dbl>
##  1 Cold CO E10 Premium Gasoline (Tier 3)                        648.
##  2 Cold CO Diesel 7-15 ppm Sulfur                               530.
##  3 Tier 3 E10 Premium Gasoline (9 RVP @Low Alt.)                486.
##  4 Cold CO Premium (Tier 2)                                     458.
##  5 CARB LEV3 E10 Regular Gasoline                               397.
##  6 E85 (85% Ethanol 15% EPA Unleaded Gasoline)                  347.
##  7 Federal Cert Diesel 7-15 PPM Sulfur                          310.
##  8 Tier 2 Cert Gasoline                                         303.
##  9 Cold CO Regular (Tier 2)                                     298.
```

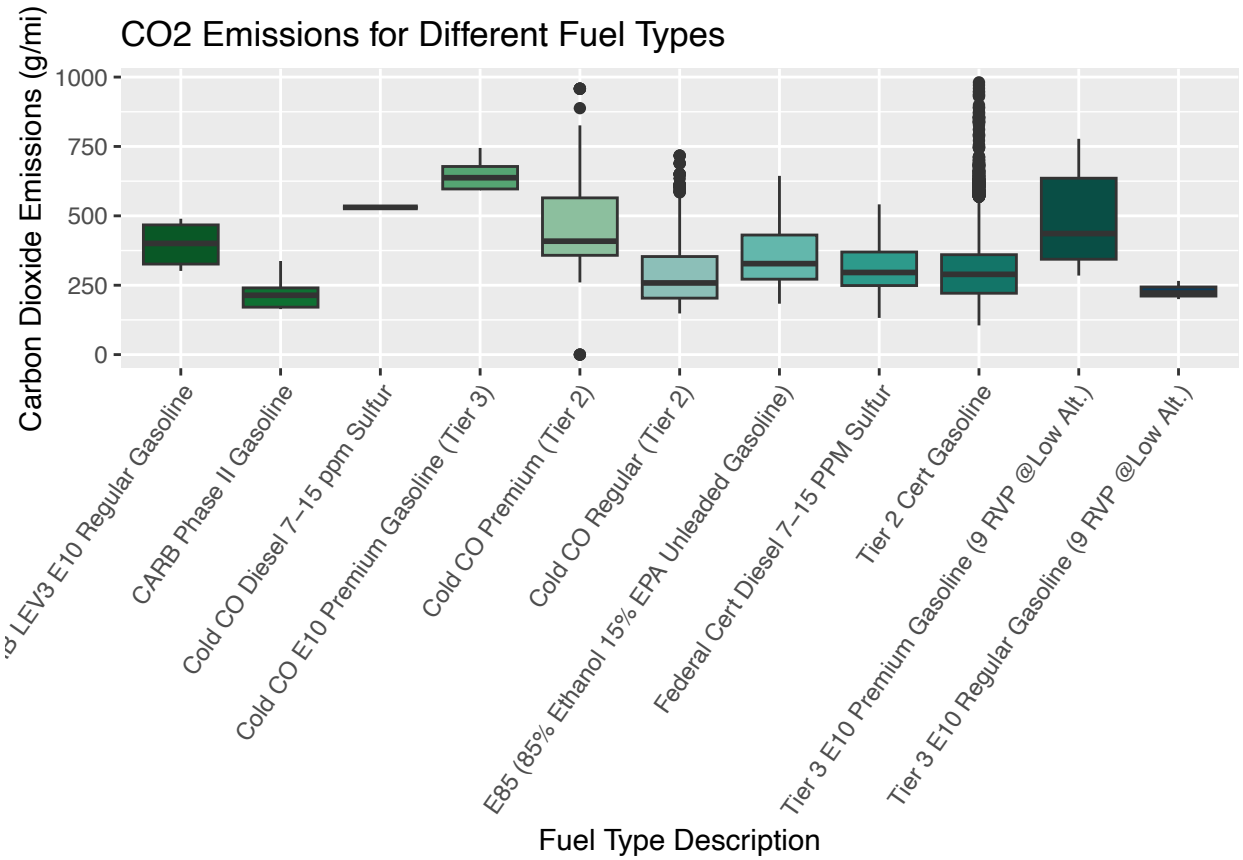
```
## 10 Tier 3 E10 Regular Gasoline (9 RVP @Low Alt.)                229.
## 11 CARB Phase II Gasoline                                       215.
```

```r
# Plot a barplot of the means
means_fuel %>% ggplot(aes(x = `Fuel Type`, y = `Mean CO2 Emissions`,
                          fill = `Fuel Type`)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  ggtitle("Mean CO2 Emissions for Different Fuel Types") +
  xlab("Fuel Type Description") +
  ylab("Mean CO2 Emissions (g/mi)") +
  theme(axis.text.x = element_text(angle = 55, vjust = 1, hjust=1)) +
  scale_fill_manual(values = GrBuPuPi)
```



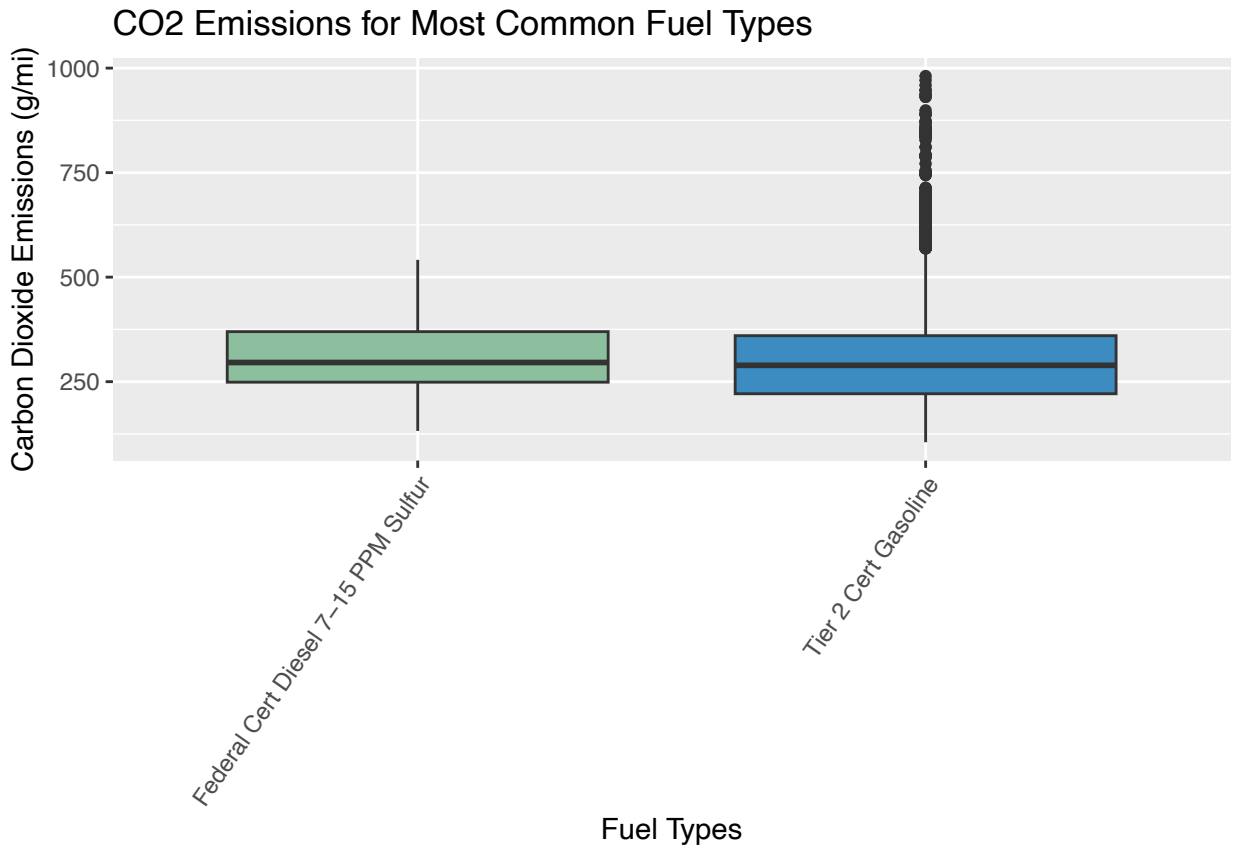
From the barplot and table above, it is clear that Cold CO E10 Premium Gasoline (Tier 3) produces the most carbon dioxide emissions out of all the different fuel types followed by Cold CO Diesel7-15 ppm Sulfur and Tier 3 E10 Premium Gasoline (9 RVP Low Alt.). The fuel types with the lowest mean carbon dioxide emissions are Cold CO Regular (Tier 2), Tier 3 E10 Regular Gasoline (9 RVP Low Alt.), and CARB Phase II Gasoline. Below, we will plot the boxplots of each fuel type to view the distributions and outliers.

```r
names(gas)[names(gas) == 'Test.Fuel.Type.Description'] <- 'Fuel Type'
gas %>% ggplot(aes(x = `Fuel Type`, y = CO2..g.mi., fill =`Fuel Type`)) +
  geom_boxplot(show.legend = FALSE) +
  ggtitle("CO2 Emissions for Different Fuel Types") +
  xlab("Fuel Type Description") + ylab("Carbon Dioxide Emissions (g/mi)") +
  theme(axis.text.x = element_text(angle = 55, vjust = 1, hjust=1)) +
  scale_fill_manual(values = GrBuPuPi)
```



From the boxplots above, we can see that the mean carbon dioxide emissions varies depending on the fuel type. It is clear that some gasolines' mean carbon dioxide emissions differ more significantly than others. Tier 2 Cert Gasoline has the most outliers out of the fuel types. Let's look closer at the boxplots of just the top two fuel types: Tier 2 Cert Gasoline and Federal Cert Diesel 7-15 PPM Sulfur.

```
top2fueltypes <- gas[gas$`Fuel Type` %in% c("Federal Cert Diesel 7-15 PPM Sulfur",
                                            "Tier 2 Cert Gasoline"), ]
top2fueltypes %>% ggplot(aes(x = `Fuel Type`,
                             y = CO2..g.mi.,
                             fill = `Fuel Type`)) +
  geom_boxplot(show.legend = FALSE) +
  ggtitle("CO2 Emissions for Most Common Fuel Types") +
  xlab("Fuel Types") +
  ylab("Carbon Dioxide Emissions (g/mi)") +
  theme(axis.text.x = element_text(angle = 55,
                                   vjust = 1,
                                   hjust=1)) +
  scale_fill_manual(values = GrBuPuPi[c(5, 14, 20)])
```



We can see that the means are relatively similar; thus, the test below will determine whether or not there is a significant difference.

Test 1

Let us compare the mean emissions between the two most common fuel types in the data set. Below we will test to see if there are statistically significant difference in the mean emissions between Tier 2 Cert Gasoline and Federal Cert Diesel 7-15 PPM Sulfur. We will define the following null and alternative hypotheses:

Declaring Hypotheses and Significance Level

H_0 : The mean carbon dioxide emissions is the same for Tier 2 Cert Gasoline and Federal Cert Diesel 7-15 PPM Sulfur.

H_a : The mean carbon dioxide emissions is greater for Federal Cert Diesel 7-15 PPM Sulfur than Tier 2 Cert Gasoline.

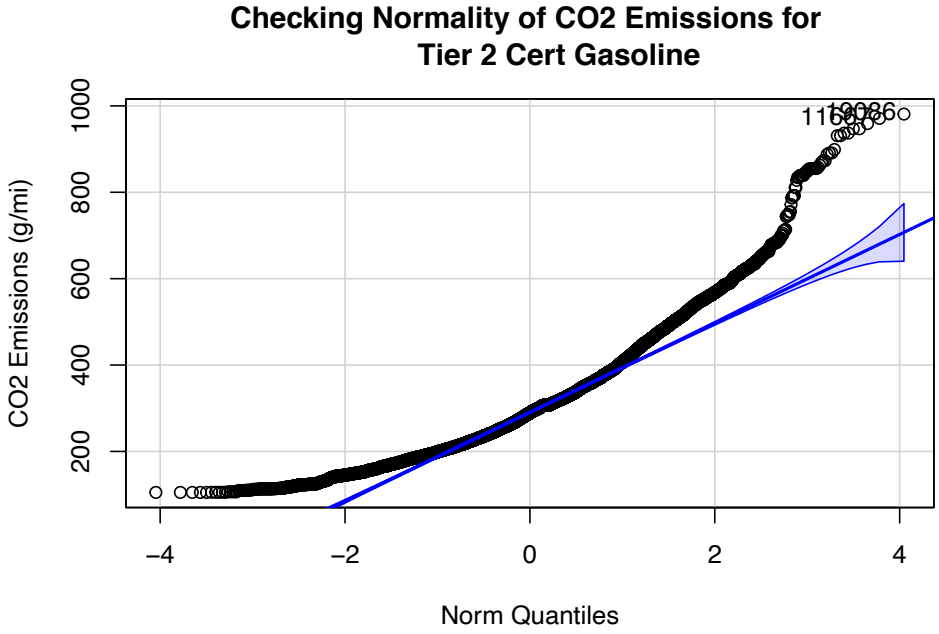
Significance Level: 1%

Checking Assumptions

```
# Separate into two data frames filtered by each type
tier2Cert <- gas %>%
  filter(`Fuel Type` == "Tier 2 Cert Gasoline")
```

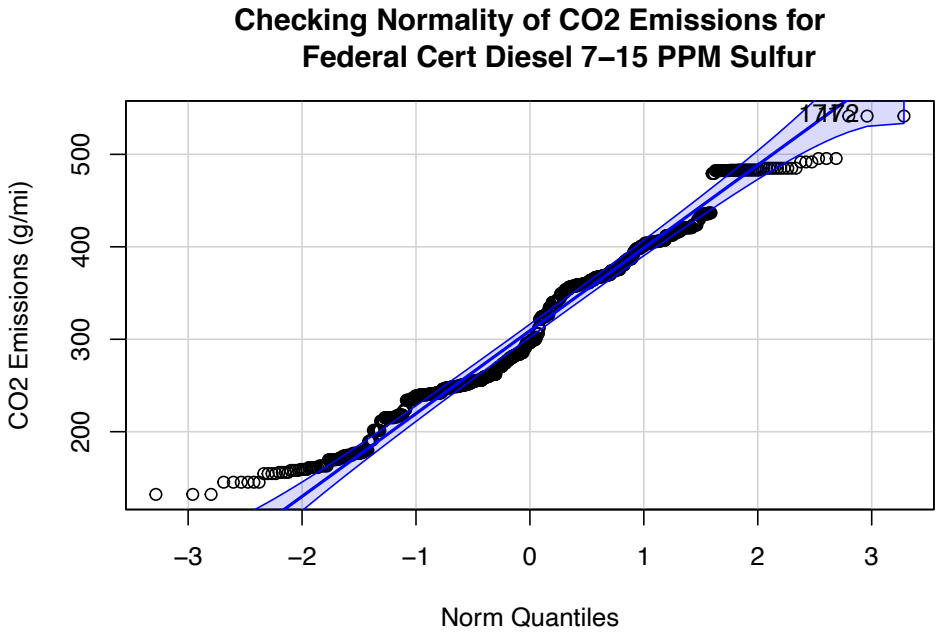
```r
fedCertDieselSulfur <- gas %>%
  filter(`Fuel Type` == "Federal Cert Diesel 7-15 PPM Sulfur")
```

```r
# Population 1: Tier 2 Cert Gasoline
qqPlot(tier2Cert$`CO2..g.mi.`,
       main = "Checking Normality of CO2 Emissions for
       Tier 2 Cert Gasoline",
       xlab = "Norm Quantiles",
       ylab = "CO2 Emissions (g/mi)")
```




```
## [1] 19086 11667
```

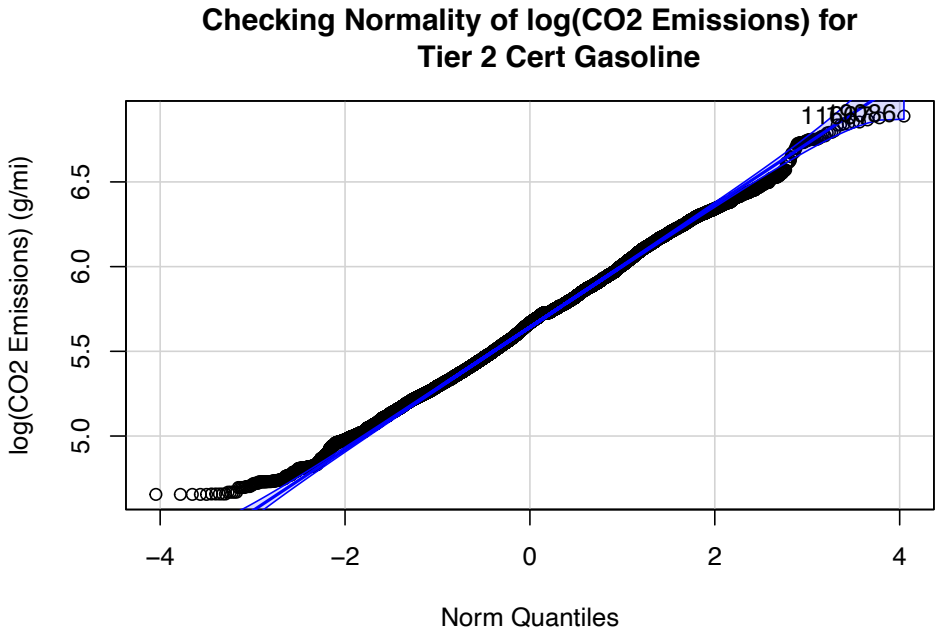
```r
# Population 2: Federal Cert Diesel 7-15 PPM Sulfur
qqPlot(fedCertDieselSulfur$`CO2..g.mi.`,
       main = "Checking Normality of CO2 Emissions for
       Federal Cert Diesel 7-15 PPM Sulfur",
       xlab = "Norm Quantiles",
       ylab = "CO2 Emissions (g/mi)")
```




```
## [1] 171 172
```

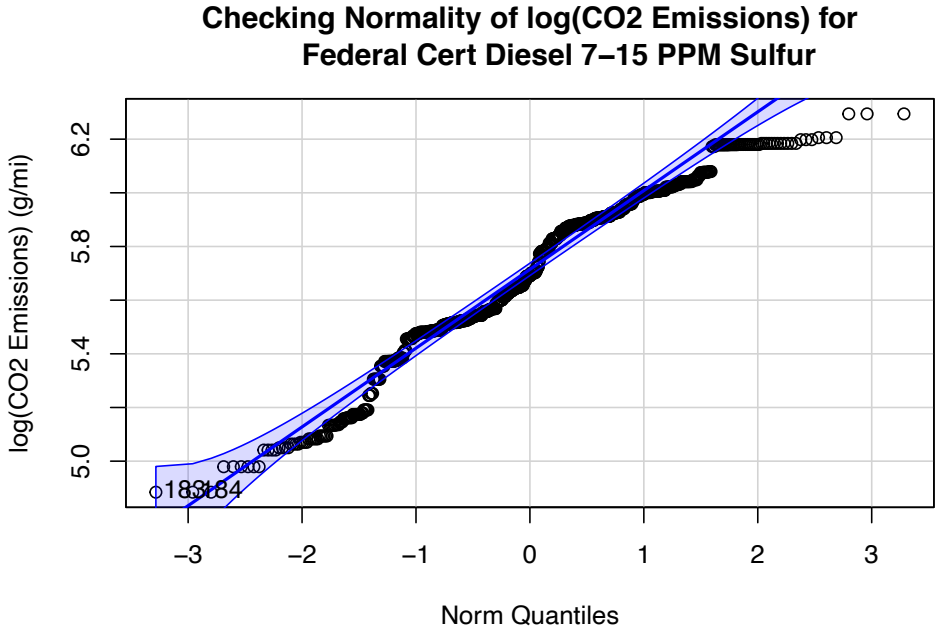
The QQ-plots show that the distribution of tier 2 cert gasoline carbon dioxide emissions is heavily skewed to the *right*, and the distribution of federal cert diesel carbon dioxide emissions is possibly *bimodal*. Thus, the normality assumption does not hold. Let us see if a log transformation normalizes the data:

```
# Population 1: Tier 2 Cert Gasoline
qqPlot(log(tier2Cert$`CO2..g.mi.`),
       main = "Checking Normality of log(CO2 Emissions) for
       Tier 2 Cert Gasoline",
       xlab = "Norm Quantiles",
       ylab = "log(CO2 Emissions) (g/mi)")
```




```
## [1] 19086 11667
```

```
# Population 2: Federal Cert Diesel 7-15 PPM Sulfur
qqPlot(log(fedCertDieselSulfur$`CO2..g.mi.`),
       main = "Checking Normality of log(CO2 Emissions) for
       Federal Cert Diesel 7-15 PPM Sulfur",
       xlab = "Norm Quantiles",
       ylab = "log(CO2 Emissions) (g/mi)")
```



```
## [1] 183 184
```

```
# Shapiro Test
shapiro.test(log(fedCertDieselSulfur$`CO2..g.mi.`))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(fedCertDieselSulfur$CO2..g.mi.)
## W = 0.96906, p-value = 1.548e-13
```

From the QQ-plots above, it is clear that the log transformation normalized the tier 2 cert gasoline data, but not the federal cert diesel data. The shapiro test result with a p-value of less than 0.01 confirms this result that the log of the federal cert diesel emissions is not normal. Thus, we will perform a Mann-Whitney U Test without the log transformation.

Mann-Whitney U Test

```
# Perform test
mw.test1 <- wilcox.test(fedCertDieselSulfur$`CO2..g.mi.`, tier2Cert$`CO2..g.mi.`,
                        na.rm = TRUE, paired = FALSE, exact = FALSE, conf.int = TRUE)
mw.test1
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  fedCertDieselSulfur$CO2..g.mi. and tier2Cert$CO2..g.mi.
## W = 10288865, p-value = 2.134e-07
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  10.61379 23.05800
## sample estimates:
## difference in location
##               16.96013
```

P-Value Analysis: Based on the test results above, the p-value $= 2.134e-07 < 0.01$ which is statistically significant. Thus, we reject the null hypothesis and conclude that the mean carbon dioxide emissions is greater for Federal Cert Diesel 7-15 PPM Sulfur than Tier 2 Cert Gasoline.

Confidence Interval Analysis: From the 95% confidence interval, we can conclude with 95% confidence that Federal Cert Diesel 7-15 PPM Sulfur, on average, produces between 10.61379 g/mi and 23.05800 g/mi more CO2 emissions than Tier 2 Cert Gasoline for the sample of vehicles in the data set.

Exploratory Data Analysis (EDA) for Test 2

Research Question: Is there a significant difference in the amount of carbon dioxide emissions between vehicle manufacturers, specifically the two most common manufacturers?

How many observations are their for each manufacturer?

```r
# Create a frequency table
frequencies <- data.frame(cbind(table(gas$Vehicle.Manufacturer.Name)))
frequencies$`Manufacturer` <- row.names(frequencies)
frequencies$`Frequency` <- frequencies$cbind.table.gas.Vehicle.Manufacturer.Name..
frequencies <- frequencies %>% dplyr::select("Manufacturer", "Frequency")
rownames(frequencies) <- NULL
# Print table ordered by frequency
frequencies[order(frequencies$Frequency, decreasing = TRUE),]
```

```
##            Manufacturer Frequency
## 9                    GM      2624
## 31               Toyota      2472
## 3                   BMW      2327
## 8                 FOMOCO      2140
## 32  Volkswagen Group of      1703
## 10                Honda      1445
## 6           FCA US LLC      1399
## 23               Nissan      1299
## 11              Hyundai      1224
## 21        Mercedes-Benz       935
## 14                  Kia       899
## 25              Porsche       645
## 12  Jaguar Land Rover L       608
## 29               Subaru       446
## 19                MAZDA       430
## 7               Ferrari       272
## 33                Volvo       236
## 22 Mitsubishi Motors Co       195
## 18             Maserati       100
## 20   McLaren Automotive        85
## 26          Rolls-Royce        60
## 1          aston martin        50
## 27                Roush        37
## 17                Lotus        30
## 13  Karma Automotive, L        28
## 5             FCA Italy        15
## 28                  RUF         8
## 16          Lamborghini         7
## 2               Bentley         5
## 4                   EPA         4
## 24  Pagani Automobili S         4
## 30  SUBARU TECNICA INTE         4
## 15           Koenigsegg         2
```

From the frequency table above, the two most common manufacturers are GM and Toyota.

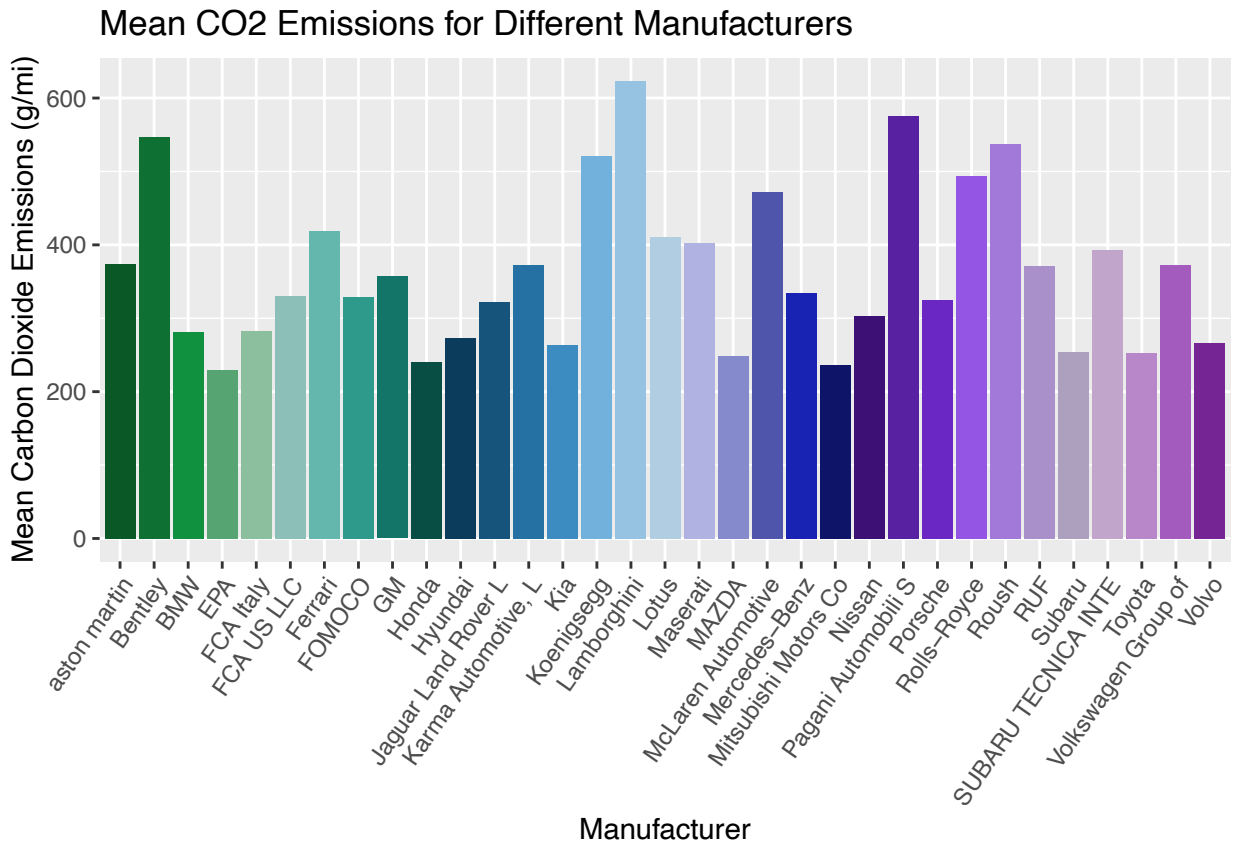
Which manufacturer produces the most carbon dioxide emissions in this data set?

```r
# Calculate the mean CO2 emissions for each fuel type
means_manufacturer <- gas %>%
  group_by(Vehicle.Manufacturer.Name) %>%
  summarise_at(vars(CO2..g.mi.), list(name = mean))
colnames(means_manufacturer) <- c("Manufacturer", "Mean CO2 Emissions")
```

```r
# Print means ordered by mean
print(means_manufacturer[order(means_manufacturer$`Mean CO2 Emissions`,
                               decreasing = TRUE),])
```

```
## # A tibble: 33 x 2
##    Manufacturer        `Mean CO2 Emissions`
##    <chr>                              <dbl>
##  1 Lamborghini                         624.
##  2 Pagani Automobili S                 575.
##  3 Bentley                             547.
##  4 Roush                               538.
##  5 Koenigsegg                          521.
##  6 Rolls-Royce                         494.
##  7 McLaren Automotive                  472.
##  8 Ferrari                             419.
##  9 Lotus                               411.
## 10 Maserati                            403.
## # ... with 23 more rows
```

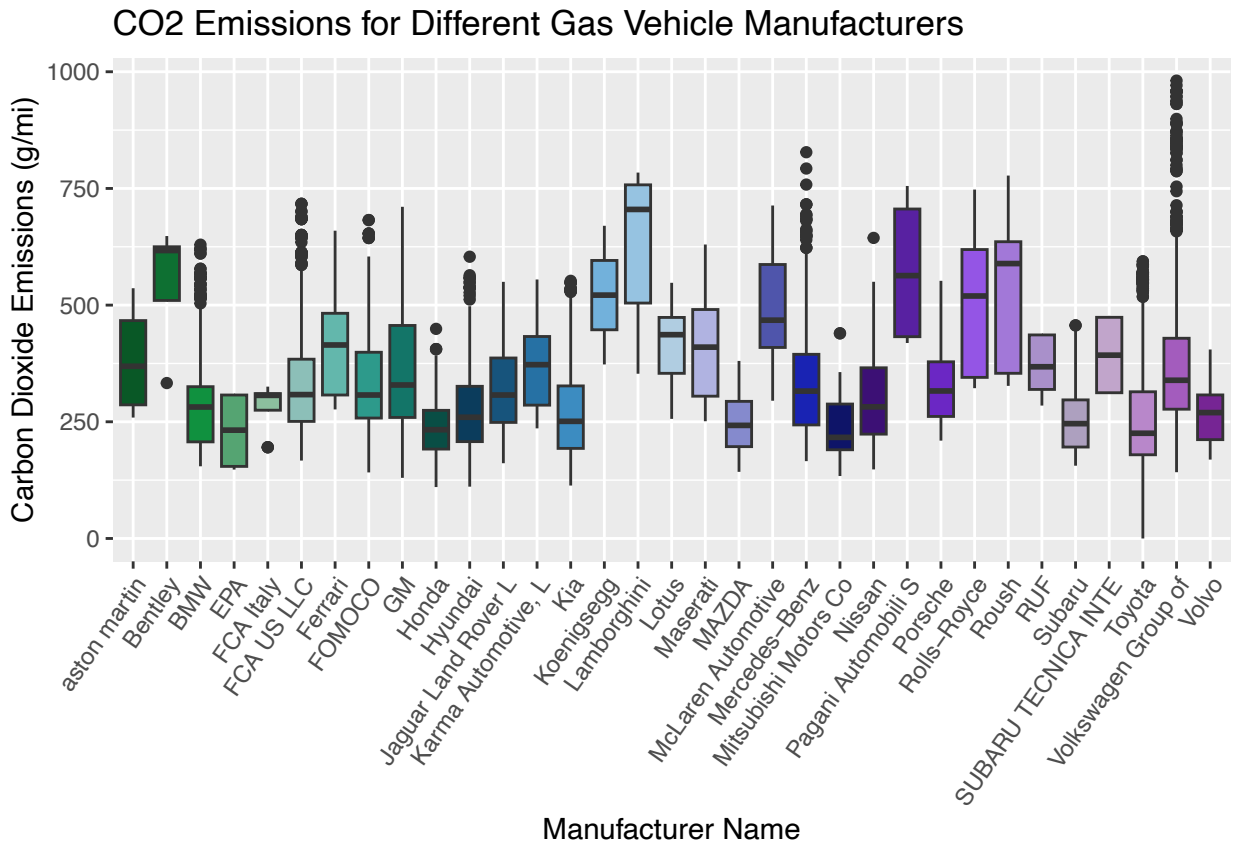
```r
# Plot a barplot of the means
means_manufacturer %>% ggplot(aes(x = Manufacturer,
                                  y = `Mean CO2 Emissions`,
                                  fill = Manufacturer)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  ggtitle("Mean CO2 Emissions for Different Manufacturers") +
  xlab("Manufacturer") +
  ylab("Mean Carbon Dioxide Emissions (g/mi)") +
  theme(axis.text.x = element_text(angle = 55,
                                   vjust = 1,
                                   hjust=1)) +
  scale_fill_manual(values = GrBuPuPi)
```



From the barplot and table above, the three manufacturers with the highest mean carbon dioxide emission in the data set are Lamborghini, Pagani Automobili S, and Bentley. The three manufacturers with the lowest mean carbon dioxide emission are Honda, Mitsubishi Motors Co, and EPA.

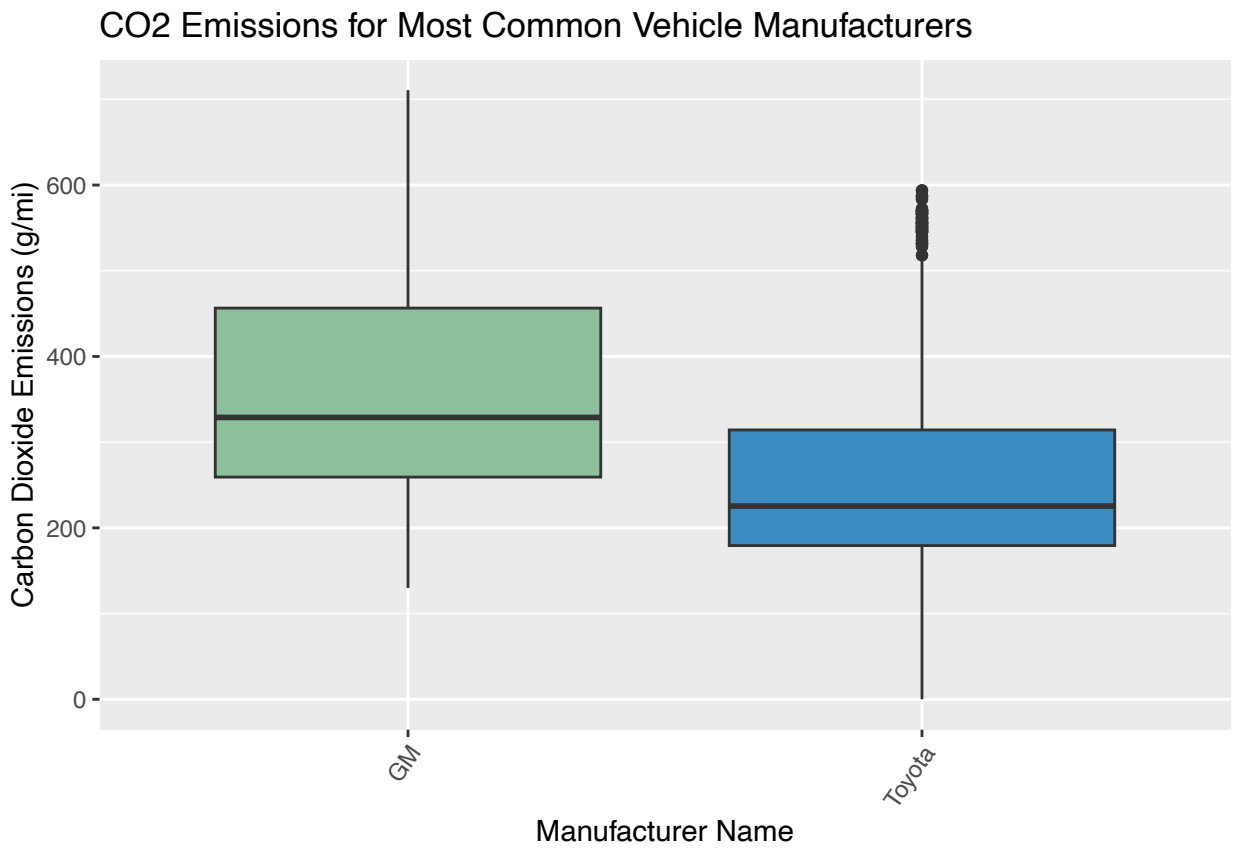
Below, we will plot the boxplots of carbon dioxide emissions for each manufacturer to view the distributions and outliers.

```
gas %>% ggplot(aes(x = Vehicle.Manufacturer.Name,
                   y = CO2..g.mi.,
                   fill = Vehicle.Manufacturer.Name)) +
  geom_boxplot(show.legend = FALSE) +
  ggtitle("CO2 Emissions for Different Gas Vehicle Manufacturers") +
  xlab("Manufacturer Name") +
  ylab("Carbon Dioxide Emissions (g/mi)") +
  theme(axis.text.x = element_text(angle = 55,
                                   vjust = 1,
                                   hjust=1)) +
  scale_fill_manual(values = GrBuPuPi)
```



From the boxplots above, we can see that the mean carbon dioxide emissions varies greatly between manufacturers. It is clear that some manufacturers' mean carbon dioxide emissions differ more significantly than others. FCA US LLC, Mercedes-Benz, and Volkswagen Group contain outliers with higher carbon dioxide emissions. Let's look closer at the boxplots of just the top two manufacturers: GM and Toyota.

```
top2manufacturers <- gas[gas$Vehicle.Manufacturer.Name %in% c("GM", "Toyota"), ]
top2manufacturers %>% ggplot(aes(x = Vehicle.Manufacturer.Name,
                                 y = CO2..g.mi.,
                                 fill = Vehicle.Manufacturer.Name)) +
  geom_boxplot(show.legend = FALSE) +
  ggtitle("CO2 Emissions for Most Common Vehicle Manufacturers") +
  xlab("Manufacturer Name") +
  ylab("Carbon Dioxide Emissions (g/mi)") +
  theme(axis.text.x = element_text(angle = 55,
                                   vjust = 1,
                                   hjust=1)) +
  scale_fill_manual(values = GrBuPuPi[c(5, 14, 20)])
```



Test 2

Let us compare the mean emissions between the two most common manufacturers in the data set: GM and Toyota. From the boxplot above, it appears that GM's mean carbon dioxide emission is higher than Toyota's, so we will test to see if this difference is significant below. We will define the following null and alternative hypotheses:

Declaring Hypotheses and Significance Level

H_0 : The mean carbon dioxide emissions is the same for GM and Toyota gasoline vehicles.

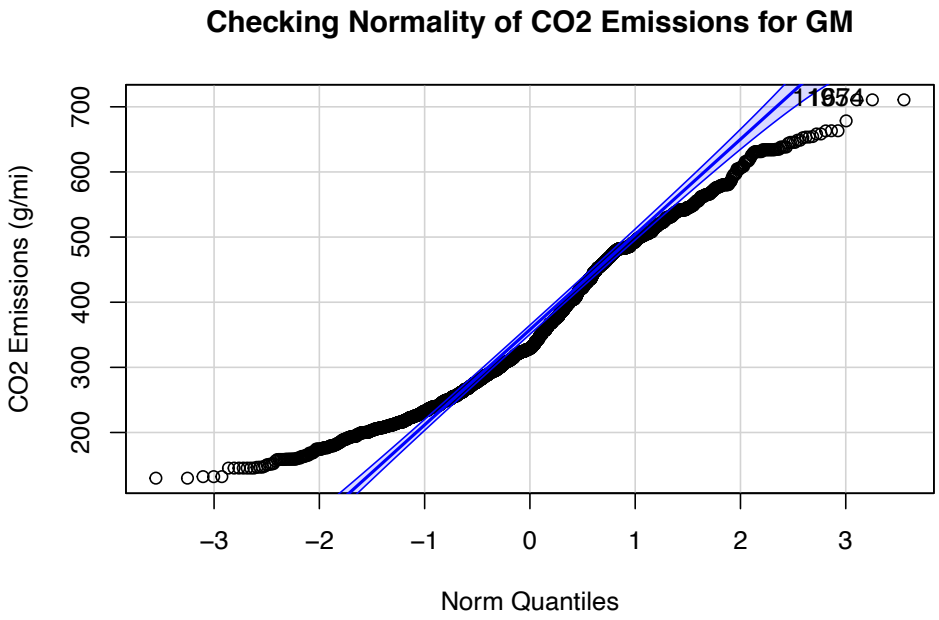
H_a : The mean carbon dioxide emissions is greater for GM gasoline vehicles than Toyota gasoline vehicles.

Significance Level: 1%

Checking Assumptions

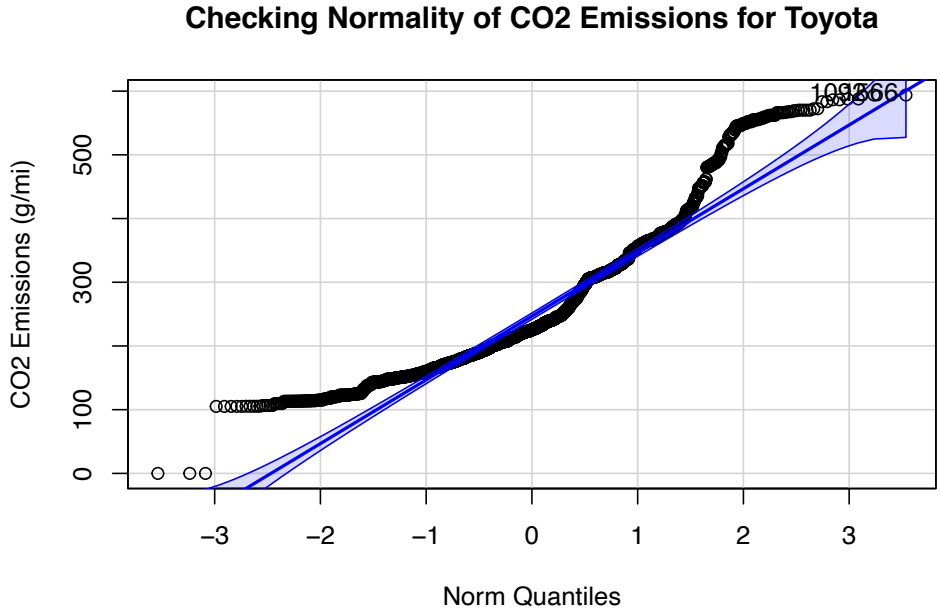
```
# Separate into two data frames filtered by each type
GM <- gas %>% filter(Vehicle.Manufacturer.Name == "GM")
Toyota <- gas %>% filter(Vehicle.Manufacturer.Name == "Toyota")
```

```r
# Population 1: GM
qqPlot(GM$`CO2..g.mi.`,
       main = "Checking Normality of CO2 Emissions for GM",
       xlab = "Norm Quantiles",
       ylab = "CO2 Emissions (g/mi)")
```




```
## [1] 1195 1674
```

```r
# Population 2: Toyota
qqPlot(Toyota$`CO2..g.mi.`,
       main = "Checking Normality of CO2 Emissions for Toyota",
       xlab = "Norm Quantiles",
       ylab = "CO2 Emissions (g/mi)")
```




```
## [1] 1092 1566
```

From the QQ-plots above, it is clear that both distributions are *not* normal. Thus, we will move forward with a Mann-Whitney U Test.

Mann-Whitney U Test

```
# Perform test
mw.test2 <- wilcox.test(GM$`CO2..g.mi.`, Toyota$`CO2..g.mi.`,
                        na.rm = TRUE, paired = FALSE,
                        exact = FALSE, conf.int = TRUE)
mw.test2
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  GM$CO2..g.mi. and Toyota$CO2..g.mi.
## W = 4902158, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##   94.60287 106.32853
## sample estimates:
## difference in location
##               100.4969
```

P-Value Analysis: Based on the test results above, the p-value $= 2.2e-16 < 0.01$ which is statistically significant. Thus, we reject the null hypothesis and conclude that the mean carbon dioxide emissions is greater for GM gasoline vehicles than Toyota gasoline vehicles.

Confidence Interval Analysis: From the 95% confidence interval, we can conclude with 95% confidence that GM gasoline vehicles, on average, produce between 99.16565 g/mi and 106.32853 g/mi more CO2 emissions than Toyota gasoline vehicles for the sample of vehicles in the data set.

Exploratory Data Analysis (EDA) for Test 3

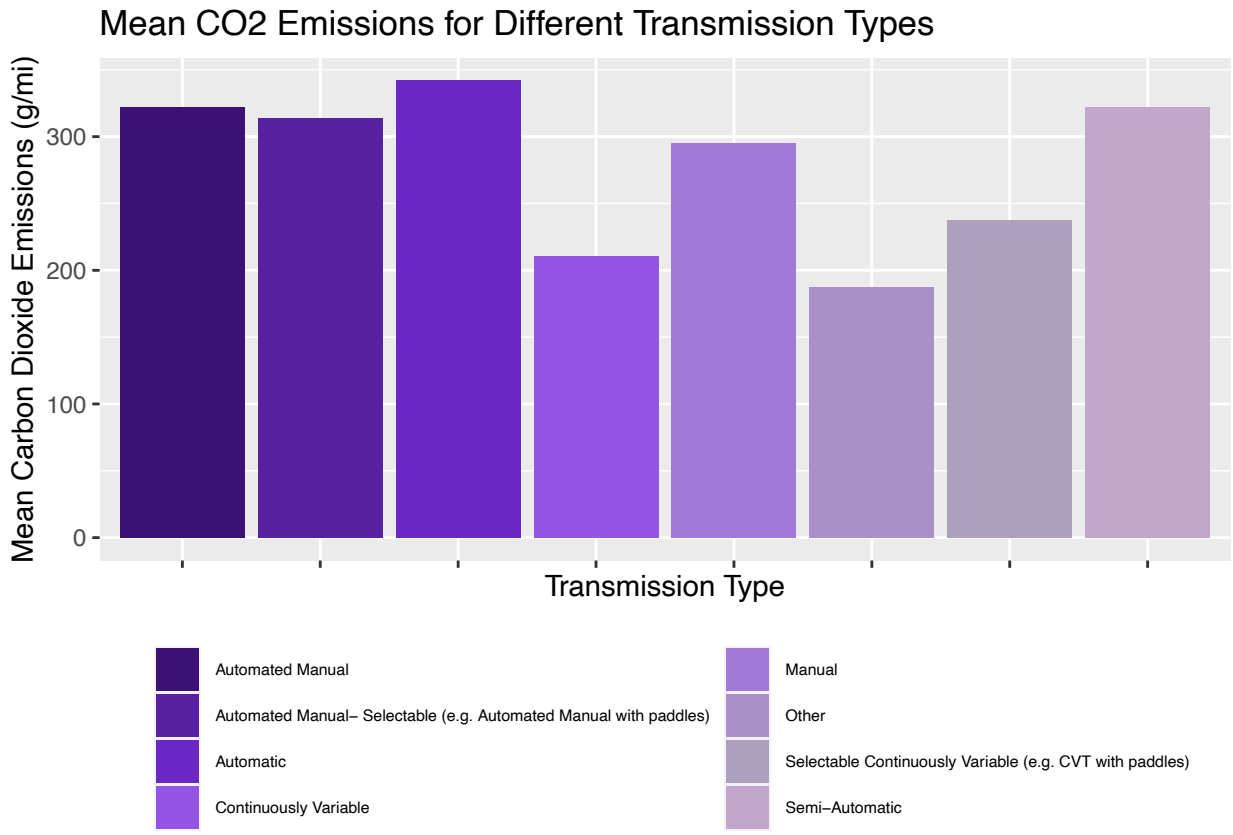
Research Question: Is there a significant difference in the amount of carbon dioxide emissions between vehicle transmission types, specifically manual and automatic vehicles?

Which transmission type produces the most carbon dioxide emissions in this data set?

```
# Calculate the mean CO2 emissions for each fuel type
means_transmissions <- gas %>%
  group_by(Tested.Transmission.Type) %>%
  summarise_at(vars(CO2..g.mi.), list(name = mean))
colnames(means_transmissions) <- c("Transmission Type", "Mean CO2 Emissions")
# Print means ordered by mean
print(means_transmissions[order(means_transmissions$`Mean CO2 Emissions`,
                                decreasing = TRUE),])
```

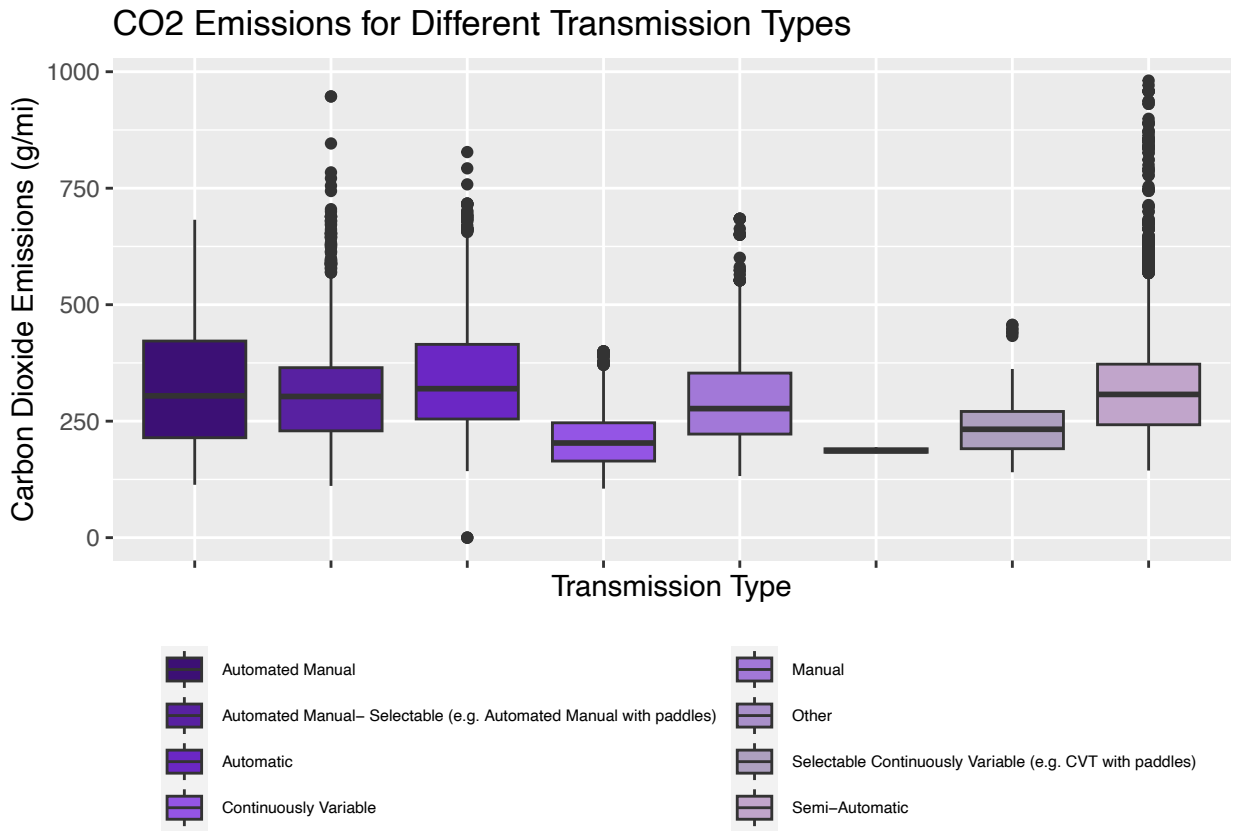
```
## # A tibble: 8 x 2
##   `Transmission Type`                                              Mean CO2 E~1
##   <chr>                                                                   <dbl>
## 1 Automatic                                                               342.
## 2 Automated Manual                                                        322.
## 3 Semi-Automatic                                                          322.
## 4 Automated Manual- Selectable (e.g. Automated Manual with paddles)       314.
## 5 Manual                                                                  295.
## 6 Selectable Continuously Variable (e.g. CVT with paddles)                238.
## 7 Continuously Variable                                                   211.
## 8 Other                                                                   187.
## # ... with abbreviated variable name 1: `Mean CO2 Emissions`
```

```
# Plot a barplot of the means
means_transmissions %>% ggplot(aes(x = `Transmission Type`,
                                   y = `Mean CO2 Emissions`,
                                   fill = `Transmission Type`)) +
  geom_bar(stat = "identity") +
  ggtitle("Mean CO2 Emissions for Different Transmission Types") +
  xlab("Transmission Type") +
  ylab("Mean Carbon Dioxide Emissions (g/mi)") +
  theme(axis.text.x = element_blank(),
        legend.position = "bottom",
        legend.text=element_text(size = 6)) +
  scale_fill_manual(values = GrBuPuPi[c(23,24,25,26,27,28,29,30)],
                    name = NULL) +
  guides(fill=guide_legend(ncol = 2))
```



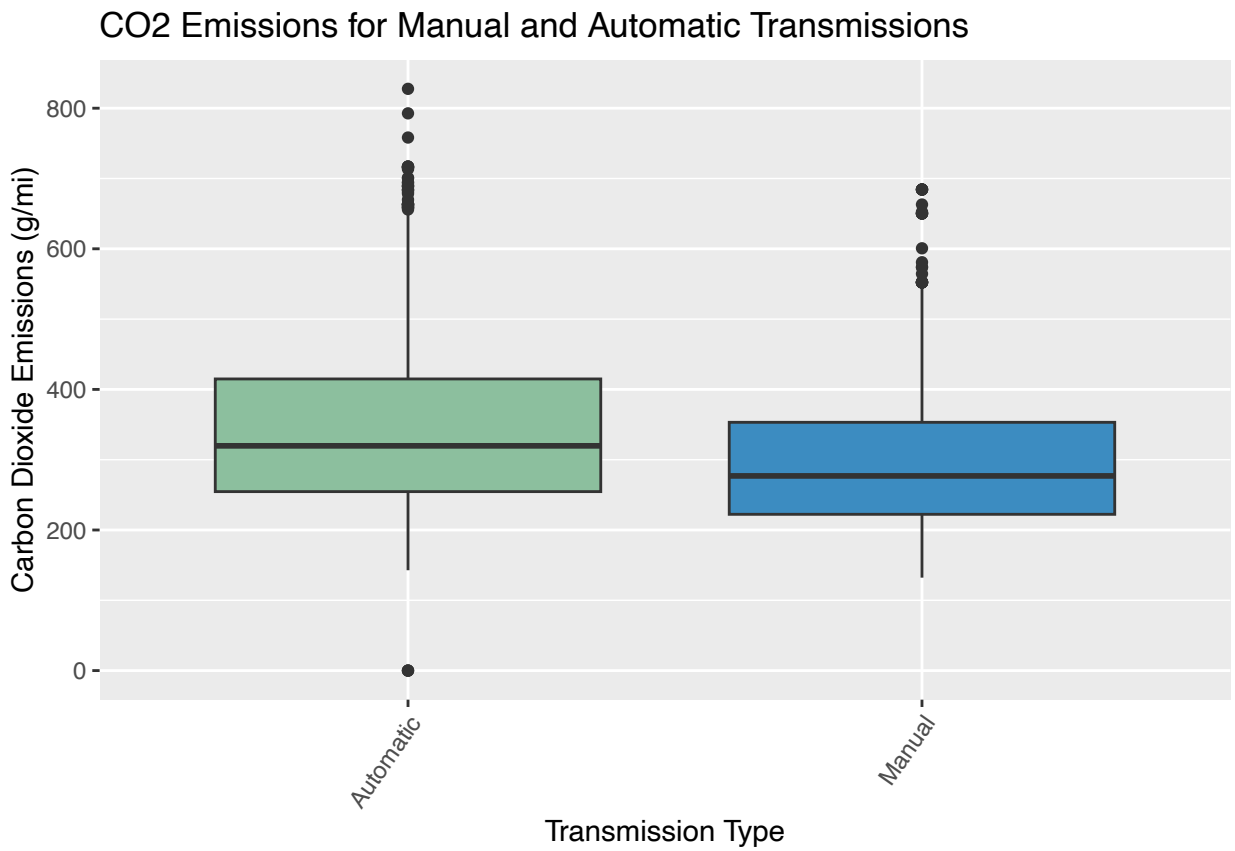
From the barplot and table above, we can see that the three transmission types with the highest mean carbon dioxide emissions are automatic, automated manual, and semi-automatic. The lowest three are selectable continuously variable, continuously variable, and other. Below, we will plot the boxplots of carbon dioxide emissions for each transmission type to view the distributions and outliers.

```
names(gas)[names(gas) == 'Tested.Transmission.Type'] <- 'Transmission Type'
gas %>% ggplot(aes(x = `Transmission Type`,
                   y = CO2..g.mi.,
                   fill = `Transmission Type`)) +
  geom_boxplot() +
  ggtitle("CO2 Emissions for Different Transmission Types") +
  xlab("Transmission Type") +
  ylab("Carbon Dioxide Emissions (g/mi)") +
  theme(axis.text.x = element_blank(),
        legend.position = "bottom",
        legend.text=element_text(size = 6)) +
  scale_fill_manual(values = GrBuPuPi[c(23,24,25,26,27,28,29,30)],
                    name = NULL) +
  guides(fill=guide_legend(ncol = 2))
```



From the boxplots above, we can see that the mean carbon dioxide emissions does not vary as much between transmission types as it did between manufacturer and fuel type. Many also contain several outliers that have higher carbon dioxide emissions. Let's look closer at the boxplots of just the manual and automatic vehicle CO2 emissions distributions.

```
df <- gas[gas$`Transmission Type` %in% c("Manual", "Automatic"), ]
df %>% ggplot(aes(x = `Transmission Type`,
                  y = CO2..g.mi.,
                  fill = `Transmission Type`)) +
  geom_boxplot(show.legend = FALSE) +
  ggtitle("CO2 Emissions for Manual and Automatic Transmissions") +
  xlab("Transmission Type") +
  ylab("Carbon Dioxide Emissions (g/mi)") +
  theme(axis.text.x = element_text(angle = 55,
                                   vjust = 1,
                                   hjust=1)) +
  scale_fill_manual(values = GrBuPuPi[c(5, 14, 20)])
```



From the boxplot above, it is clear that automatic transmissions have the slightly higher mean emissions; the test below will determine if this difference is significant.

T-Test 3

Let us compare the mean emissions between automatic cars and manual vehicles.

Declaring Hypotheses and Significance Level

H_0 : The mean carbon dioxide emissions is the same for automatic and manual vehicles.

H_a : The mean carbon dioxide emissions is greater for automatic vehicles is higher than for manual vehicles.

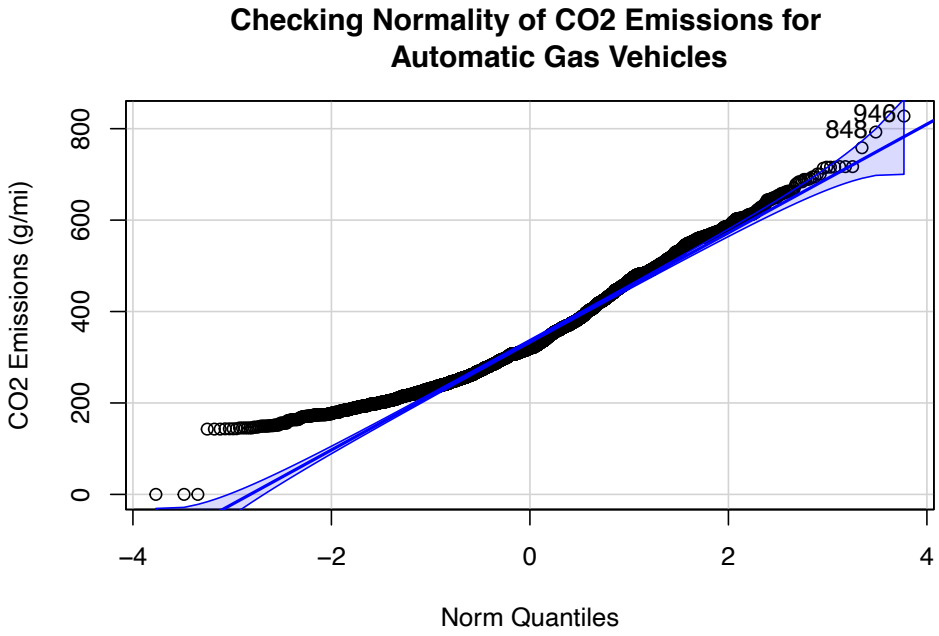
Significance Level: 1%

Checking Assumptions

```
# Separate into two data frames filtered by each type
automatic <- gas %>% filter(`Transmission Type` == "Automatic")
manual <- gas %>% filter(`Transmission Type` == "Manual")
```

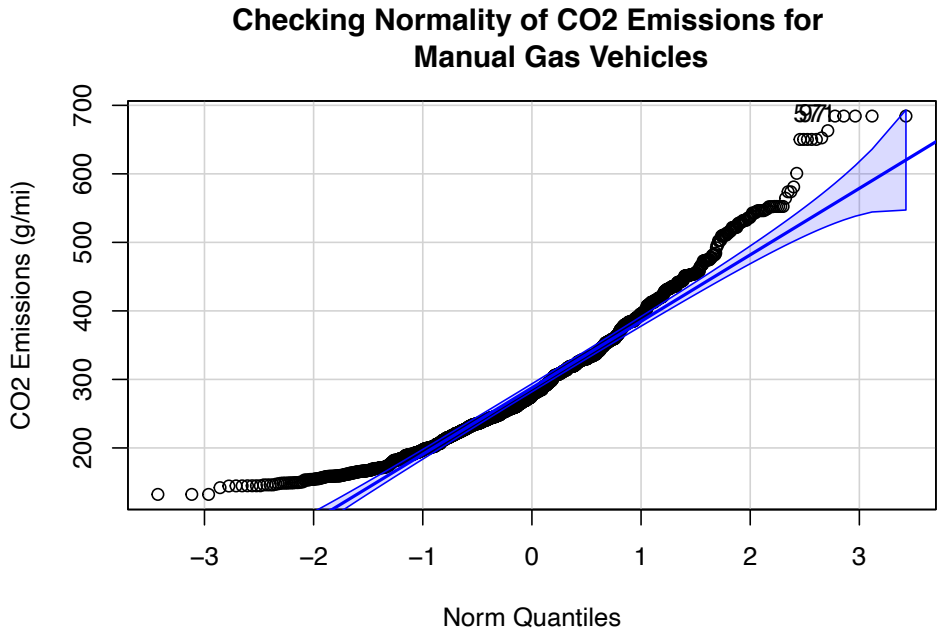
```
# Population 1: automatic
qqPlot(automatic$`CO2..g.mi.`,
       main = "Checking Normality of CO2 Emissions for
```

```
       Automatic Gas Vehicles",
       xlab = "Norm Quantiles",
       ylab = "CO2 Emissions (g/mi)")
```




```
## [1] 946 848
```

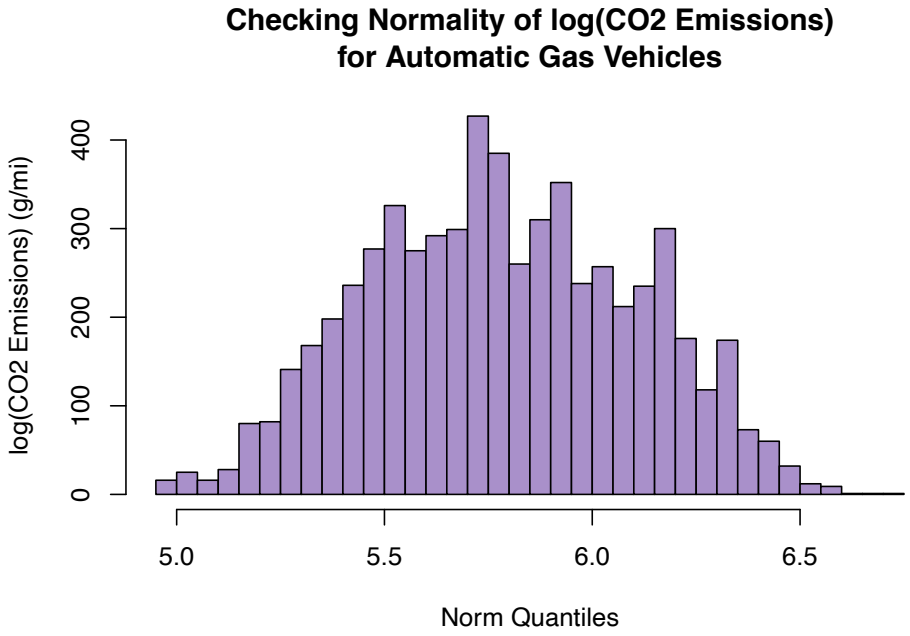
```
# Population 2: manual
qqPlot(manual$`CO2..g.mi.`,
       main = "Checking Normality of CO2 Emissions for
       Manual Gas Vehicles",
       xlab = "Norm Quantiles",
       ylab = "CO2 Emissions (g/mi)")
```




```
## [1]  97 571
```

Both distributions appear to be skewed to the right. Let us check if a log transformation is useful in normalizing the data.

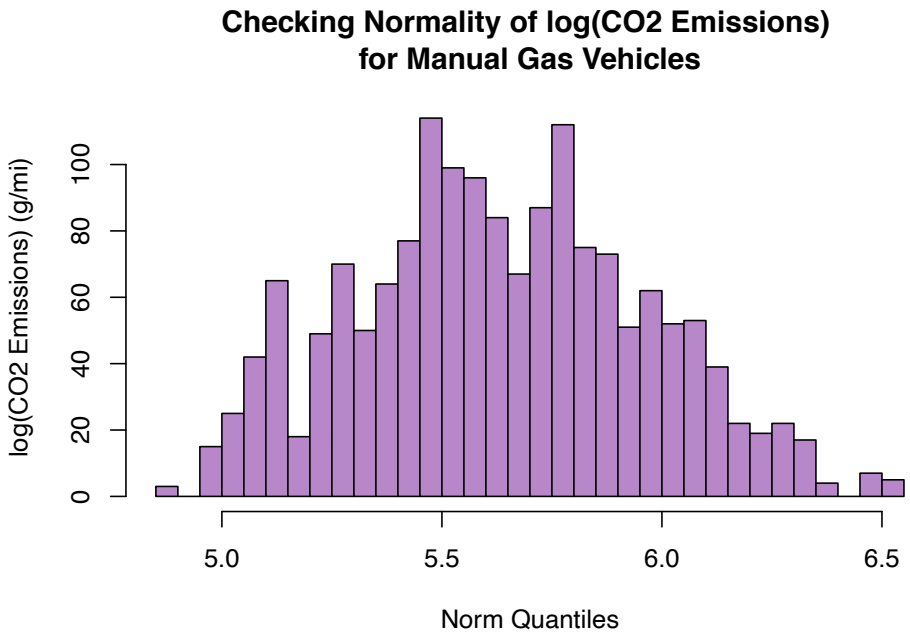
```
# Population 1: automatic
hist(log(automatic$`CO2..g.mi.`),
main = "Checking Normality of log(CO2 Emissions)
for Automatic Gas Vehicles",
xlab = "Norm Quantiles", ylab = "log(CO2 Emissions) (g/mi)",
col = GrBuPuPi[c(28)], breaks = 40)
```




```
data1 <- log(automatic$`CO2..g.mi.`)
data1[!is.finite(data1)] <- NA
shapiro.test(sample(data1, 5000))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sample(data1, 5000)
## W = 0.99112, p-value < 2.2e-16
```

```
# Population 2: manual
hist(log(manual$`CO2..g.mi.`),
main = "Checking Normality of log(CO2 Emissions)
for Manual Gas Vehicles",
xlab = "Norm Quantiles", ylab = "log(CO2 Emissions) (g/mi)",
col = GrBuPuPi[c(31)], breaks = 40)
```




```
data2 <- log(manual$`CO2..g.mi.`)
data2[!is.finite(data2)] <- NA
shapiro.test(data2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data2
## W = 0.99129, p-value = 2.697e-08
```

While the histograms show some improvement in normality from the log transformation, the Shapiro tests with very small p-values assert that the data still does not follow a normal distribution. Thus, we must move forward with a Mann-Whitney U Test.

Mann-Whitney U Test

```
# Perform test
mw.test3 <- wilcox.test(automatic$`CO2..g.mi.`, manual$`CO2..g.mi.`,
                        na.rm = TRUE, paired = FALSE,
                        exact = FALSE, conf.int = TRUE)
mw.test3
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  automatic$CO2..g.mi. and manual$CO2..g.mi.
## W = 6236277, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
## 95 percent confidence interval:
##  38.75607 49.62847
## sample estimates:
## difference in location
##               44.17741
```

P-Value Analysis: Based on the test results above, the p-value = $2.2e-16 < 0.01$ which is statistically significant. Thus, we reject the null hypothesis and conclude that the mean carbon dioxide emissions is greater for automatic gasoline vehicles than manual gasoline vehicles.

Confidence Interval Analysis: From the 95% confidence interval, we can conclude with 95% confidence that automatic gasoline vehicles, on average, produce between 38.75607 g/mi and 49.62847 g/mi more CO2 emissions than manual gasoline vehicles for the sample of vehicles in the data set.

This result shows that manual vehicles are more fuel efficient. This makes sense as manual vehicles are typically lighter and have a less complex engine set up.

References

1. R Color Codes:

https://www.rapidtables.com/web/color/RGB_Color.html

2. Barplots:

http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization

3. Boxplots:

http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization

4. Legend Customization:

http://www.sthda.com/english/wiki/ggplot2-legend-easy-steps-to-change-the-position-and-the-appearance-of-a-graph-legend-in-r-software

5. ANLY 511 Lecture 10 Slides

6. QQ-Plot Documentation:

https://braverock.com/brian/R/PerformanceAnalytics/html/chart.QQPlot.html

511-LM

Natalie Smith

2022-12-02

Linear Regression

The goal of linear regression, in short, is to predict the value of a chosen response variable based on the value of another variable or variables, which are called “predictors”. The equations of multiple linear regression models, which consider multiple predictor variables, take the following form:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$, where ϵ represents the error present in the model.

Research Question

1. Can the elements of a car’s design be used to predict its CO_2 output?
2. Which elements of a car’s design are best at predicting CO_2 output?

Assumptions of Linear Regression

1. Individual observations are independent from each other
2. A linear relationship exists between the independent predictor variables X_i and the dependent response variable Y
3. Homoscedasticity, or homogeneity of variance
4. The residuals of the model are normally distributed

In this section, we are going use multiple linear regression to determine which elements of a car’s design (predictor variables $X_1, X_2, \ldots, X_n$) are good predictors of a car’s CO_2 emissions (response variable Y).

Since electric cars do not give off CO_2 emissions, I am going to start by reading in our cleaned non-electric, or fuel-based cars, dataset.

```
setwd("..")
getwd()
```

```
## [1] "/Users/smithnatalie/ANLY511-Final-Project"
```

```
nonelectric = read.csv("data/cardata_nonelectric_clean.csv")
```

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(tidyverse)
```

```
## ── Attaching packages
## ─────────────────────────────────────
## tidyverse 1.3.2 ──
```

```
## ✔ tibble  3.1.8     ✔ dplyr   1.0.9
## ✔ tidyr   1.2.0     ✔ stringr 1.4.0
## ✔ readr   2.1.2     ✔ forcats 0.5.1
## ✔ purrr   0.3.4
## ── Conflicts ──────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ purrr::lift()   masks caret::lift()
```

```
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(ISLR2)
library(leaps)
```

Before splitting into training and testing, I want to ensure the predictor of Model Year can be considered as categorical for the regression model by adding a categorical column with the same data to the dataset.

```
dplyr::count(nonelectric, Model.Year, sort = TRUE)
```

```
##   Model.Year    n
## 1       2018 4607
## 2       2019 4559
## 3       2020 4298
## 4       2022 4191
## 5       2021 4083
```

```
nonelectric$Model.Year.Cat = as.character(nonelectric$Model.Year)
```

Next, the data will be split into training and testing datasets with an 80/20 split.

```
set.seed(101)

training.samples = nonelectric$CO2..g.mi. %>%
  createDataPartition(p = 0.8, list = FALSE)

training.data = nonelectric[training.samples,]
testing.data = nonelectric[-training.samples, ]
```

```
dim(training.data)
```

```
## [1] 17392    39
```

```
dim(testing.data)
```

```
## [1] 4346   39
```

```
head(training.data)
```

```
##   X Model.Year Vehicle.Manufacturer.Name Veh.Mfr.Code Represented.Test.Veh.Make
## 1 1       2018              aston martin          ASX              Aston Martin
## 2 2       2018              aston martin          ASX              Aston Martin
## 3 3       2018              aston martin          ASX              Aston Martin
## 4 4       2018              aston martin          ASX              Aston Martin
## 5 5       2018              aston martin          ASX              Aston Martin
## 6 6       2018              aston martin          ASX              Aston Martin
##   Represented.Test.Veh.Model Test.Veh.Displacement..L. Vehicle.Type
## 1                       DB11                       5.2          Car
## 2                       DB11                       5.2          Car
## 3                    DB11 V8                       4.0          Car
## 4                    DB11 V8                       4.0          Car
## 5                   Rapide S                       6.0          Car
## 6                   Rapide S                       6.0          Car
##   Rated.Horsepower X..of.Cylinders.and.Rotors Tested.Transmission.Type.Code
## 1              600                        12                           SA
## 2              600                        12                           SA
## 3              503                         8                           SA
## 4              503                         8                           SA
## 5              552                        12                           SA
## 6              552                        12                           SA
##   Tested.Transmission.Type X..of.Gears Transmission.Lockup. Drive.System.Code
## 1           Semi-Automatic           8                    Y                 R
## 2           Semi-Automatic           8                    Y                 R
## 3           Semi-Automatic           8                    Y                 R
## 4           Semi-Automatic           8                    Y                 R
## 5           Semi-Automatic           8                    Y                 R
## 6           Semi-Automatic           8                    Y                 R
##   Drive.System.Description Equivalent.Test.Weight..lbs.. Axle.Ratio N.V.Ratio
## 1      2-Wheel Drive, Rear                          4500       2.70      22.2
## 2      2-Wheel Drive, Rear                          4500       2.70      22.2
## 3      2-Wheel Drive, Rear                          4500       2.70      22.2
## 4      2-Wheel Drive, Rear                          4500       2.70      22.2
## 5      2-Wheel Drive, Rear                          4750       2.73      22.4
## 6      2-Wheel Drive, Rear                          4750       2.73      22.4
##   Test.Fuel.Type.Description THC..g.mi. CO..g.mi. CO2..g.mi. RND_ADJ_FE
## 1       Tier 2 Cert Gasoline   0.024700  0.418000     466.87       18.8
## 2       Tier 2 Cert Gasoline   0.001155  0.067334     285.00       30.9
## 3       Tier 2 Cert Gasoline   0.026500  0.070000     386.66       22.7
## 4       Tier 2 Cert Gasoline   0.000500  0.030000     259.74       33.8
## 5       Tier 2 Cert Gasoline   0.026900  0.500000     511.93       17.3
## 6       Tier 2 Cert Gasoline   0.000800  0.060000     296.63       29.9
##   DT.Inertia.Work.Ratio.Rating DT.Absolute.Speed.Change.Ratg
## 1                   -2.5300000                   -1.7300000
## 2                    1.3600000                    0.4400000
## 3                  -11.9900000                   -9.2600000
## 4                   -3.6400000                   -3.2100000
## 5                    0.5655838                    0.4420405
## 6                    0.5655838                    0.4420405
##   DT.Energy.Economy.Rating Target.Coef.A..lbf. Target.Coef.B..lbf.mph.
## 1               -1.7100000               40.94                  0.0169
## 2               -0.5900000               40.94                  0.0169
## 3               -7.7100000               40.94                  0.0169
## 4               -0.9600000               40.94                  0.0169
## 5               -0.2002973               32.66                  0.6085
## 6               -0.2002973               32.66                  0.6085
##   Target.Coef.C..lbf.mph..2. Set.Coef.A..lbf. Set.Coef.B..lbf.mph.
## 1                     0.0271           6.810               0.0807
## 2                     0.0271           6.810               0.0807
## 3                     0.0271          11.260               0.0919
## 4                     0.0271          11.260               0.0919
## 5                     0.0198           1.093               2.1980
## 6                     0.0198           1.093               2.1980
##   Set.Coef.C..lbf.mph..2. Aftertreatment.Device.Cd Aftertreatment.Device.Desc
## 1                  0.0245                      TWC         Three-way catalyst
## 2                  0.0245                      TWC         Three-way catalyst
## 3                  0.0251                      TWC         Three-way catalyst
## 4                  0.0251                      TWC         Three-way catalyst
```

```
## 5                  0.0280                      TWC         Three-way catalyst
## 6                  0.0280                      TWC         Three-way catalyst
##   Police...Emergency.Vehicle. Averaging.Method.Cd Averging.Method.Desc
## 1                           N                   N         No averaging
## 2                           N                   N         No averaging
## 3                           N                   N         No averaging
## 4                           N                   N         No averaging
## 5                           N                   N         No averaging
## 6                           N                   N         No averaging
##   Model.Year.Cat
## 1           2018
## 2           2018
## 3           2018
## 4           2018
## 5           2018
## 6           2018
```

```
head(testing.data)
```

```
##     X Model.Year Vehicle.Manufacturer.Name Veh.Mfr.Code
## 10 10       2018                  Bentley          BEX
## 26 26       2018                      BMW          BMX
## 29 29       2018                      BMW          BMX
## 30 30       2018                      BMW          BMX
## 49 49       2018                      BMW          BMX
## 51 51       2018                      BMW          BMX
##    Represented.Test.Veh.Make Represented.Test.Veh.Model
## 10                   BENTLEY             Continental GT
## 26                       BMW                 230i Coupe
## 29                       BMW                 230i Coupe
## 30                       BMW                 230i Coupe
## 49                       BMW                       320i
## 51                       BMW                       320i
##    Test.Veh.Displacement..L. Vehicle.Type Rated.Horsepower
## 10                     5.998          Car              616
## 26                     2.000          Car              248
## 29                     2.000          Car              248
## 30                     2.000          Car              248
## 49                     2.000         Both              181
## 51                     2.000         Both              180
##    X..of.Cylinders.and.Rotors Tested.Transmission.Type.Code
## 10                         12                            SA
## 26                          4                            SA
## 29                          4                            SA
## 30                          4                            SA
## 49                          4                             A
## 51                          4                             M
##    Tested.Transmission.Type X..of.Gears Transmission.Lockup. Drive.System.Code
## 10           Semi-Automatic           8                    Y                 F
## 26           Semi-Automatic           8                    Y                 R
## 29           Semi-Automatic           8                    Y                 R
## 30           Semi-Automatic           8                    Y                 R
## 49                Automatic           8                    Y                 R
## 51                   Manual           6                    N                 R
##    Drive.System.Description Equivalent.Test.Weight..lbs.. Axle.Ratio N.V.Ratio
## 10     2-Wheel Drive, Front                          6000       2.85      24.9
## 26      2-Wheel Drive, Rear                          3625       2.81      25.2
## 29      2-Wheel Drive, Rear                          3625       2.81      24.8
## 30      2-Wheel Drive, Rear                          3625       2.81      24.8
## 49      2-Wheel Drive, Rear                          3625       3.20      28.0
## 51      2-Wheel Drive, Rear                          3625       3.08      34.1
##    Test.Fuel.Type.Description THC..g.mi. CO..g.mi. CO2..g.mi. RND_ADJ_FE
## 10       Tier 2 Cert Gasoline  0.0711000 0.7680000   625.0000       14.2
## 26       Tier 2 Cert Gasoline  0.0049414 0.2191188   279.6962       31.8
## 29       Tier 2 Cert Gasoline  0.0010600 0.0609000   174.5200       50.5
## 30       Tier 2 Cert Gasoline  0.0054200 0.3940000   294.2200       30.0
## 49       Tier 2 Cert Gasoline  0.0173200 0.3219000   315.5700       28.2
## 51       Tier 2 Cert Gasoline  0.0095300 0.3500000   300.9900       29.5
##    DT.Inertia.Work.Ratio.Rating DT.Absolute.Speed.Change.Ratg
## 10                    0.5655838                    0.4420405
## 26                    2.8484370                    1.6135490
## 29                    2.1140000                    1.4250000
## 30                    0.0610000                   -0.1900000
## 49                    0.5655838                    0.4420405
## 51                    0.5655838                    0.4420405
##    DT.Energy.Economy.Rating Target.Coef.A..lbf. Target.Coef.B..lbf.mph.
## 10               -0.2002973              54.853                 0.04883
## 26                1.6329070              49.900                -0.20400
## 29               -0.6110000              46.300                -0.21500
## 30               -1.1230000              46.300                -0.21500
## 49               -0.2002973              46.200                -0.33400
## 51               -0.2002973              28.900                 0.11400
##    Target.Coef.C..lbf.mph..2. Set.Coef.A..lbf. Set.Coef.B..lbf.mph.
## 10                   0.022116           7.614              -0.0083
## 26                   0.020220          18.500               0.2390
## 29                   0.020230          15.000               0.2170
## 30                   0.020230          15.000               0.2170
```

```
## 49                  0.020270          21.000             -0.2680
## 51                  0.016280          16.400             -0.0940
##    Set.Coef.C..lbf.mph..2. Aftertreatment.Device.Cd Aftertreatment.Device.Desc
## 10                0.020935                      TWC         Three-way catalyst
## 26                0.014530                      TWC         Three-way catalyst
## 29                0.014640                      TWC         Three-way catalyst
## 30                0.014640                      TWC         Three-way catalyst
## 49                0.019040                      TWC         Three-way catalyst
## 51                0.015080                      TWC         Three-way catalyst
##    Police...Emergency.Vehicle. Averaging.Method.Cd Averging.Method.Desc
## 10                           N                   N         No averaging
## 26                           N                   N         No averaging
## 29                           N                   N         No averaging
## 30                           N                   N         No averaging
## 49                           N                   N         No averaging
## 51                           N                   N         No averaging
##    Model.Year.Cat
## 10           2018
## 26           2018
## 29           2018
## 30           2018
## 49           2018
## 51           2018
```

Next, it is important to check the data for any missing values before proceeding.

```
colSums(is.na(training.data))
```

```
##                              X                     Model.Year
##                              0                              0
##     Vehicle.Manufacturer.Name                  Veh.Mfr.Code
##                              0                              0
##     Represented.Test.Veh.Make    Represented.Test.Veh.Model
##                              0                              0
##     Test.Veh.Displacement..L.                  Vehicle.Type
##                              0                              0
##              Rated.Horsepower    X..of.Cylinders.and.Rotors
##                              0                              0
## Tested.Transmission.Type.Code      Tested.Transmission.Type
##                              0                              0
##                   X..of.Gears          Transmission.Lockup.
##                              0                              0
##             Drive.System.Code      Drive.System.Description
##                              0                              0
## Equivalent.Test.Weight..lbs..                    Axle.Ratio
##                              0                              0
##                      N.V.Ratio    Test.Fuel.Type.Description
##                              0                              0
##                    THC..g.mi.                     CO..g.mi.
##                              0                              0
##                    CO2..g.mi.                    RND_ADJ_FE
##                              0                              0
##  DT.Inertia.Work.Ratio.Rating DT.Absolute.Speed.Change.Ratg
##                              0                              0
##      DT.Energy.Economy.Rating           Target.Coef.A..lbf.
##                              0                              0
##       Target.Coef.B..lbf.mph.    Target.Coef.C..lbf.mph..2.
##                              0                              0
##              Set.Coef.A..lbf.          Set.Coef.B..lbf.mph.
##                              0                              0
##       Set.Coef.C..lbf.mph..2.      Aftertreatment.Device.Cd
##                              0                           158
##    Aftertreatment.Device.Desc   Police...Emergency.Vehicle.
##                           158                              0
##           Averaging.Method.Cd          Averging.Method.Desc
##                              0                              0
##                Model.Year.Cat
##                              0
```

```
colSums(is.na(testing.data))
```

```
##                              X                     Model.Year
##                              0                              0
##     Vehicle.Manufacturer.Name                  Veh.Mfr.Code
##                              0                              0
##     Represented.Test.Veh.Make    Represented.Test.Veh.Model
##                              0                              0
##     Test.Veh.Displacement..L.                  Vehicle.Type
##                              0                              0
##              Rated.Horsepower    X..of.Cylinders.and.Rotors
##                              0                              0
## Tested.Transmission.Type.Code      Tested.Transmission.Type
##                              0                              0
##                   X..of.Gears          Transmission.Lockup.
##                              0                              0
##             Drive.System.Code      Drive.System.Description
##                              0                              0
## Equivalent.Test.Weight..lbs..                    Axle.Ratio
##                              0                              0
##                      N.V.Ratio    Test.Fuel.Type.Description
##                              0                              0
##                    THC..g.mi.                     CO..g.mi.
##                              0                              0
##                    CO2..g.mi.                    RND_ADJ_FE
##                              0                              0
##  DT.Inertia.Work.Ratio.Rating DT.Absolute.Speed.Change.Ratg
##                              0                              0
##      DT.Energy.Economy.Rating           Target.Coef.A..lbf.
##                              0                              0
##       Target.Coef.B..lbf.mph.    Target.Coef.C..lbf.mph..2.
##                              0                              0
##              Set.Coef.A..lbf.          Set.Coef.B..lbf.mph.
##                              0                              0
##       Set.Coef.C..lbf.mph..2.      Aftertreatment.Device.Cd
##                              0                             35
##    Aftertreatment.Device.Desc   Police...Emergency.Vehicle.
##                             35                              0
##           Averaging.Method.Cd          Averging.Method.Desc
##                              0                              0
##                Model.Year.Cat
##                              0
```

Above, it can be seen that the Aftertreatment.Device.Cd and the Aftertreatment.Device.Desc have missing values. Because these columns may be significant in the model and because there are still many rows in the dataset, I will remove these rows from the original dataset and run the train/test split again.

```
nonelectric = nonelectric %>% drop_na()
```

Now, the split will be done again before proceeding to the linear regression model.

```
set.seed(101)

training.samples = nonelectric$CO2..g.mi. %>%
  createDataPartition(p = 0.8, list = FALSE)

training.data = nonelectric[training.samples,]
testing.data = nonelectric[-training.samples, ]
```

There are some initial unnecessary variables that can be removed before running the multiple linear regression model: - X (This is just an index) - Veh.Mfr.Code, Represented.Test.Veh.Make, Tested.Transmission.Type.Code, Drive.System.Code, and Aftertreatment.Device.Cd (We have the full name for all of these) - Police…Emergency.Vehicle (All “no’s” throughout and doesn’t apply to this dataset and what we are looking for at all) - Averaging.Method.Cd (This is just the way things are calculated rather than an actual metric - categorical)

Additionally, the two variables of Vehicle.Manufacturer.Name and Represented.Test.Veh.Model, which detail the make and model of each respective car, need to be left out. The reason for this is that the multiple linear regression model is unable to predict emissions for makes and models of cars that appear in the testing set but not the training set, preventing the model from working.

Other than those columns, all other terms will be used predict a full model and will tweaked based on results for additional models.

The following will be considered as categorical “dummy” variables in the model:

- Model.Year
- Vehicle.Type
- Tested.Transmission.Type
- Transmission.Lockup.
- Drive.System.Description
- Test.Fuel.Type.Description
- Averaging.Method.Desc

Emissions Model 1: Full model with all columns as predictors

```
emissions.model = lm(CO2..g.mi. ~ Model.Year.Cat + Test.Veh.Displacement..L. + Vehicle.Type + Rated.Horsepower + X..of.Cylinders.and.Rotors + Tested.Transmission.Type + X..of.Gears + Transmission.Lockup. + Drive.System.Description + Equivalent.Test.Weight..lbs.. + Axle.Ratio + N.V.Ratio + Test.Fuel.Type.Description + THC..g.mi. + CO..g.mi. + RND_ADJ_FE + DT.Inertia.Work.Ratio.Rating + DT.Absolute.Speed.Change.Ratg + Target.Coef.A..lbf. + Target.Coef.B..lbf.mph. + Target.Coef.C..lbf.mph..2. + Set.Coef.A..lbf. + Set.Coef.B..lbf.mph. + Set.Coef.C..lbf.mph..2. + Aftertreatment.Device.Desc, data = training.data)
```

```
options(max.print=999999)
```

```
summary(emissions.model)
```

```
##
## Call:
## lm(formula = CO2..g.mi. ~ Model.Year.Cat + Test.Veh.Displacement..L. +
##     Vehicle.Type + Rated.Horsepower + X..of.Cylinders.and.Rotors +
##     Tested.Transmission.Type + X..of.Gears + Transmission.Lockup. +
##     Drive.System.Description + Equivalent.Test.Weight..lbs.. +
##     Axle.Ratio + N.V.Ratio + Test.Fuel.Type.Description + THC..g.mi. +
##     CO..g.mi. + RND_ADJ_FE + DT.Inertia.Work.Ratio.Rating + DT.Absolute.Speed.Change.Ratg +
##     Target.Coef.A..lbf. + Target.Coef.B..lbf.mph. + Target.Coef.C..lbf.mph..2. +
##     Set.Coef.A..lbf. + Set.Coef.B..lbf.mph. + Set.Coef.C..lbf.mph..2. +
##     Aftertreatment.Device.Desc, data = training.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -400.31  -23.82   -4.84   16.11  650.26
##
## Coefficients:
##                                                                                 Estimate
## (Intercept)                                                                    5.092e+02
## Model.Year.Cat2019                                                             1.406e+00
## Model.Year.Cat2020                                                             3.491e+00
## Model.Year.Cat2021                                                             2.929e+00
## Model.Year.Cat2022                                                             4.811e+00
## Test.Veh.Displacement..L.                                                      6.268e+00
## Vehicle.TypeCar                                                                6.482e+00
## Vehicle.TypeTruck                                                             -6.079e+00
## Rated.Horsepower                                                               3.851e-02
## X..of.Cylinders.and.Rotors                                                     6.081e+00
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles) -7.612e+00
## Tested.Transmission.TypeAutomatic                                             -1.846e+00
## Tested.Transmission.TypeContinuously Variable                                  1.437e+01
## Tested.Transmission.TypeManual                                                -1.095e+01
## Tested.Transmission.TypeOther                                                 -1.380e+01
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)  8.761e+00
## Tested.Transmission.TypeSemi-Automatic                                        -8.860e-01
## X..of.Gears                                                                   -1.247e+00
## Transmission.Lockup.Y                                                         -4.032e+00
## Drive.System.Description2-Wheel Drive, Rear                                   -7.233e+00
## Drive.System.Description4-Wheel Drive                                         -1.259e+00
## Drive.System.DescriptionAll Wheel Drive                                       -3.997e+00
## Drive.System.DescriptionPart-time 4-Wheel Drive                               -2.087e+01
## Equivalent.Test.Weight..lbs..                                                 -7.149e-03
## Axle.Ratio                                                                    -5.767e+00
## N.V.Ratio                                                                      8.384e-01
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                              -1.113e+02
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                       4.280e+01
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                7.728e+01
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                            -1.117e+02
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                            -1.612e+02
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)         -1.700e+02
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                 -3.053e+01
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                -1.056e+02
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)        5.256e+01
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)       -1.725e+02
## THC..g.mi.                                                                     3.665e+02
## CO..g.mi.                                                                     -1.886e-03
## RND_ADJ_FE                                                                    -6.987e+00
## DT.Inertia.Work.Ratio.Rating                                                  -8.829e-01
## DT.Absolute.Speed.Change.Ratg                                                 -7.190e-01
## Target.Coef.A..lbf.                                                            5.970e-01
## Target.Coef.B..lbf.mph.                                                        3.866e+01
## Target.Coef.C..lbf.mph..2.                                                     1.545e+03
## Set.Coef.A..lbf.                                                              -2.414e-01
## Set.Coef.B..lbf.mph.                                                          -5.216e+00
## Set.Coef.C..lbf.mph..2.                                                        4.624e+01
## Aftertreatment.Device.DescNOx Adsorber                                         4.424e-01
## Aftertreatment.Device.DescOther                                               -1.989e+01
## Aftertreatment.Device.DescOxidation catalyst                                  -5.550e+00
```

```
## Aftertreatment.Device.DescSelective Catalytic Reduction                                -6.245e+00
## Aftertreatment.Device.DescThree-way catalyst                                            3.691e+01
##                                                                                       Std. Error
## (Intercept)                                                                            2.215e+01
## Model.Year.Cat2019                                                                     1.004e+00
## Model.Year.Cat2020                                                                     1.037e+00
## Model.Year.Cat2021                                                                     1.064e+00
## Model.Year.Cat2022                                                                     1.084e+00
## Test.Veh.Displacement..L.                                                              7.794e-01
## Vehicle.TypeCar                                                                        1.162e+00
## Vehicle.TypeTruck                                                                      1.419e+00
## Rated.Horsepower                                                                       5.050e-03
## X..of.Cylinders.and.Rotors                                                             5.887e-01
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles)  2.182e+00
## Tested.Transmission.TypeAutomatic                                                      2.224e+00
## Tested.Transmission.TypeContinuously Variable                                          2.761e+00
## Tested.Transmission.TypeManual                                                         2.266e+00
## Tested.Transmission.TypeOther                                                          2.526e+01
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)       2.943e+00
## Tested.Transmission.TypeSemi-Automatic                                                 2.161e+00
## X..of.Gears                                                                            2.839e-01
## Transmission.Lockup.Y                                                                  1.301e+00
## Drive.System.Description2-Wheel Drive, Rear                                            1.046e+00
## Drive.System.Description4-Wheel Drive                                                  2.145e+00
## Drive.System.DescriptionAll Wheel Drive                                                1.484e+00
## Drive.System.DescriptionPart-time 4-Wheel Drive                                        5.357e+00
## Equivalent.Test.Weight..lbs..                                                          8.541e-04
## Axle.Ratio                                                                             6.160e-01
## N.V.Ratio                                                                              6.735e-02
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                                       2.391e+01
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                               2.538e+01
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                        2.722e+01
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                                     1.946e+01
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                                     1.940e+01
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)                  1.940e+01
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                          2.147e+01
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                         1.924e+01
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)                2.174e+01
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)                3.134e+01
## THC..g.mi.                                                                            1.066e+01
## CO..g.mi.                                                                              9.244e-02
## RND_ADJ_FE                                                                             4.143e-02
## DT.Inertia.Work.Ratio.Rating                                                           4.075e-01
## DT.Absolute.Speed.Change.Ratg                                                          5.441e-01
## Target.Coef.A..lbf.                                                                    5.445e-02
## Target.Coef.B..lbf.mph.                                                                1.685e+00
## Target.Coef.C..lbf.mph..2.                                                             8.823e+01
## Set.Coef.A..lbf.                                                                       4.109e-02
## Set.Coef.B..lbf.mph.                                                                   1.115e+00
## Set.Coef.C..lbf.mph..2.                                                                3.147e+01
## Aftertreatment.Device.DescNOx Adsorber                                                 7.304e+00
## Aftertreatment.Device.DescOther                                                        6.098e+00
## Aftertreatment.Device.DescOxidation catalyst                                           4.094e+00
## Aftertreatment.Device.DescSelective Catalytic Reduction                                4.009e+00
## Aftertreatment.Device.DescThree-way catalyst                                           9.942e+00
##                                                                                         t value
## (Intercept)                                                                             22.990
## Model.Year.Cat2019                                                                       1.400
## Model.Year.Cat2020                                                                       3.367
## Model.Year.Cat2021                                                                       2.752
## Model.Year.Cat2022                                                                       4.437
## Test.Veh.Displacement..L.                                                                8.042
## Vehicle.TypeCar                                                                          5.578
## Vehicle.TypeTruck                                                                       -4.285
## Rated.Horsepower                                                                         7.625
## X..of.Cylinders.and.Rotors                                                              10.329
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles)  -3.489
## Tested.Transmission.TypeAutomatic                                                       -0.830
```

```
## Tested.Transmission.TypeContinuously Variable                                    5.206
## Tested.Transmission.TypeManual                                                  -4.834
## Tested.Transmission.TypeOther                                                   -0.546
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)  2.976
## Tested.Transmission.TypeSemi-Automatic                                          -0.410
## X..of.Gears                                                                     -4.391
## Transmission.Lockup.Y                                                           -3.100
## Drive.System.Description2-Wheel Drive, Rear                                     -6.914
## Drive.System.Description4-Wheel Drive                                           -0.587
## Drive.System.DescriptionAll Wheel Drive                                         -2.693
## Drive.System.DescriptionPart-time 4-Wheel Drive                                 -3.895
## Equivalent.Test.Weight..lbs..                                                   -8.371
## Axle.Ratio                                                                      -9.363
## N.V.Ratio                                                                       12.448
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                                -4.656
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                         1.686
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                  2.839
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                              -5.742
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                              -8.313
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)           -8.766
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                   -1.422
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                  -5.491
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)          2.418
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)         -5.505
## THC..g.mi.                                                                      34.381
## CO..g.mi.                                                                       -0.020
## RND_ADJ_FE                                                                    -168.664
## DT.Inertia.Work.Ratio.Rating                                                    -2.166
## DT.Absolute.Speed.Change.Ratg                                                   -1.321
## Target.Coef.A..lbf.                                                             10.965
## Target.Coef.B..lbf.mph.                                                         22.949
## Target.Coef.C..lbf.mph..2.                                                      17.512
## Set.Coef.A..lbf.                                                                -5.874
## Set.Coef.B..lbf.mph.                                                            -4.678
## Set.Coef.C..lbf.mph..2.                                                          1.469
## Aftertreatment.Device.DescNOx Adsorber                                           0.061
## Aftertreatment.Device.DescOther                                                 -3.262
## Aftertreatment.Device.DescOxidation catalyst                                    -1.356
## Aftertreatment.Device.DescSelective Catalytic Reduction                         -1.558
## Aftertreatment.Device.DescThree-way catalyst                                     3.713
##                                                                              Pr(>|t|)
## (Intercept)                                                                   < 2e-16
## Model.Year.Cat2019                                                           0.161665
## Model.Year.Cat2020                                                           0.000761
## Model.Year.Cat2021                                                           0.005927
## Model.Year.Cat2022                                                           9.18e-06
## Test.Veh.Displacement..L.                                                    9.40e-16
## Vehicle.TypeCar                                                              2.47e-08
## Vehicle.TypeTruck                                                            1.84e-05
## Rated.Horsepower                                                             2.56e-14
## X..of.Cylinders.and.Rotors                                                    < 2e-16
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles) 0.000485
## Tested.Transmission.TypeAutomatic                                            0.406625
## Tested.Transmission.TypeContinuously Variable                                1.96e-07
## Tested.Transmission.TypeManual                                               1.35e-06
## Tested.Transmission.TypeOther                                                0.584913
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles) 0.002920
## Tested.Transmission.TypeSemi-Automatic                                       0.681861
## X..of.Gears                                                                  1.13e-05
## Transmission.Lockup.Y                                                        0.001937
## Drive.System.Description2-Wheel Drive, Rear                                  4.86e-12
## Drive.System.Description4-Wheel Drive                                        0.557355
## Drive.System.DescriptionAll Wheel Drive                                      0.007082
## Drive.System.DescriptionPart-time 4-Wheel Drive                              9.84e-05
## Equivalent.Test.Weight..lbs..                                                 < 2e-16
## Axle.Ratio                                                                    < 2e-16
## N.V.Ratio                                                                     < 2e-16
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                             3.24e-06
```

```
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                                    0.091767
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                             0.004526
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                                          9.49e-09
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                                           < 2e-16
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)                        < 2e-16
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                               0.154911
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                              4.05e-08
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)                     0.015619
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)                     3.75e-08
## THC..g.mi.                                                                                  < 2e-16
## CO..g.mi.                                                                                   0.983722
## RND_ADJ_FE                                                                                  < 2e-16
## DT.Inertia.Work.Ratio.Rating                                                                0.030288
## DT.Absolute.Speed.Change.Ratg                                                               0.186389
## Target.Coef.A..lbf.                                                                         < 2e-16
## Target.Coef.B..lbf.mph.                                                                     < 2e-16
## Target.Coef.C..lbf.mph..2.                                                                  < 2e-16
## Set.Coef.A..lbf.                                                                            4.32e-09
## Set.Coef.B..lbf.mph.                                                                        2.92e-06
## Set.Coef.C..lbf.mph..2.                                                                     0.141827
## Aftertreatment.Device.DescNOx Adsorber                                                      0.951701
## Aftertreatment.Device.DescOther                                                             0.001110
## Aftertreatment.Device.DescOxidation catalyst                                                0.175202
## Aftertreatment.Device.DescSelective Catalytic Reduction                                     0.119300
## Aftertreatment.Device.DescThree-way catalyst                                                0.000206
##
## (Intercept)                                                                                 ***
## Model.Year.Cat2019
## Model.Year.Cat2020                                                                          ***
## Model.Year.Cat2021                                                                          **
## Model.Year.Cat2022                                                                          ***
## Test.Veh.Displacement..L.                                                                   ***
## Vehicle.TypeCar                                                                             ***
## Vehicle.TypeTruck                                                                           ***
## Rated.Horsepower                                                                            ***
## X..of.Cylinders.and.Rotors                                                                  ***
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles) ***
## Tested.Transmission.TypeAutomatic
## Tested.Transmission.TypeContinuously Variable                                               ***
## Tested.Transmission.TypeManual                                                              ***
## Tested.Transmission.TypeOther
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)           **
## Tested.Transmission.TypeSemi-Automatic
## X..of.Gears                                                                                 ***
## Transmission.Lockup.Y                                                                       **
## Drive.System.Description2-Wheel Drive, Rear                                                 ***
## Drive.System.Description4-Wheel Drive
## Drive.System.DescriptionAll Wheel Drive                                                     **
## Drive.System.DescriptionPart-time 4-Wheel Drive                                             ***
## Equivalent.Test.Weight..lbs..                                                               ***
## Axle.Ratio                                                                                  ***
## N.V.Ratio                                                                                   ***
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                                            ***
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                                    .
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                             **
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                                          ***
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                                          ***
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)                       ***
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                              ***
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)                     *
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)                     ***
## THC..g.mi.                                                                                  ***
## CO..g.mi.
## RND_ADJ_FE                                                                                  ***
## DT.Inertia.Work.Ratio.Rating                                                                *
## DT.Absolute.Speed.Change.Ratg
## Target.Coef.A..lbf.                                                                         ***
```

```
## Target.Coef.B..lbf.mph.                                                  ***
## Target.Coef.C..lbf.mph..2.                                               ***
## Set.Coef.A..lbf.                                                         ***
## Set.Coef.B..lbf.mph.                                                     ***
## Set.Coef.C..lbf.mph..2.
## Aftertreatment.Device.DescNOx Adsorber
## Aftertreatment.Device.DescOther                                          **
## Aftertreatment.Device.DescOxidation catalyst
## Aftertreatment.Device.DescSelective Catalytic Reduction
## Aftertreatment.Device.DescThree-way catalyst                             ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.4 on 17186 degrees of freedom
## Multiple R-squared:  0.856,  Adjusted R-squared:  0.8555
## F-statistic:  2003 on 51 and 17186 DF,  p-value: < 2.2e-16
```

Interestingly enough, aside from a few of the initial predictors, almost all of the predictors in the model appear to be significant in predicting emissions for a car.

However, it is very possible there may be multicollinearity in the current model, which occurs when at least two of the predictor variables in a model are highly correlated and result in redundancy, skewing the results and making the model unstable.

To detect the presence of multicollinearity, we can computer the variance inflation factor (VIF) score.

```
vif(emissions.model)
```

```
##                                     GVIF Df GVIF^(1/(2*Df))
## Model.Year.Cat                  1.218965  4        1.025059
## Test.Veh.Displacement..L.       9.530381  1        3.087131
## Vehicle.Type                    3.678897  2        1.384935
## Rated.Horsepower                5.510100  1        2.347360
## X..of.Cylinders.and.Rotors     10.908181  1        3.302754
## Tested.Transmission.Type       31.997164  7        1.280879
## X..of.Gears                     5.469504  1        2.338697
## Transmission.Lockup.            3.071643  1        1.752610
## Drive.System.Description        5.494194  4        1.237337
## Equivalent.Test.Weight..lbs..   4.902283  1        2.214110
## Axle.Ratio                      1.951033  1        1.396794
## N.V.Ratio                       1.836214  1        1.355070
## Test.Fuel.Type.Description    105.350942 10        1.262211
## THC..g.mi.                      2.141612  1        1.463425
## CO..g.mi.                       1.008894  1        1.004437
## RND_ADJ_FE                      2.308121  1        1.519250
## DT.Inertia.Work.Ratio.Rating   16.359583  1        4.044698
## DT.Absolute.Speed.Change.Ratg  16.323002  1        4.040174
## Target.Coef.A..lbf.             4.191159  1        2.047232
## Target.Coef.B..lbf.mph.         2.804017  1        1.674520
## Target.Coef.C..lbf.mph..2.      3.798719  1        1.949030
## Set.Coef.A..lbf.                2.118803  1        1.455611
## Set.Coef.B..lbf.mph.            1.226191  1        1.107335
## Set.Coef.C..lbf.mph..2.         1.396493  1        1.181733
## Aftertreatment.Device.Desc     40.527500  5        1.448021
```

Typically, predictors that exceed 5 can be considered to be highly correlated with other predictors. Since there are already many significant predictors, we will be extra conservative and remove the predictors of DT.Inertia.Work.Ratio.Rating and DT.Absolute.Speed.Change.Ratg from the model.

Combining this with the predictors that did not meet the 0.05 significance level, the predictors we will be removing to create a more “tuned” model to compare to the original are:

- DT.Inertia.Work.Ratio.Rating
- DT.Absolute.Speed.Change.Ratg
- Transmission.Lockup
- CO..g.mi.

(Note that if at least one dummy variable for a categorical variable is significant, all of them will be kept as a best practice at this stage of model tuning.)

Emissions Model 2: Removing multicollinearity from model and initial insignificant terms

```
emissions.model.2 = lm(CO2..g.mi. ~ Model.Year.Cat + Test.Veh.Displacement..L. + Vehicle.Type + Rated.Horsepower
+ X..of.Cylinders.and.Rotors + Tested.Transmission.Type + X..of.Gears + Drive.System.Description + Equivalent.Tes
t.Weight..lbs.. + Axle.Ratio + N.V.Ratio + Test.Fuel.Type.Description + THC..g.mi. + RND_ADJ_FE + Target.Coef.A..
lbf. + Target.Coef.B..lbf.mph. + Target.Coef.C..lbf.mph..2. + Set.Coef.A..lbf. + Set.Coef.B..lbf.mph. + Set.Coef.
C..lbf.mph..2. + Aftertreatment.Device.Desc, data = training.data)
```

```
summary(emissions.model.2)
```

```
##
## Call:
## lm(formula = CO2..g.mi. ~ Model.Year.Cat + Test.Veh.Displacement..L. +
##     Vehicle.Type + Rated.Horsepower + X..of.Cylinders.and.Rotors +
##     Tested.Transmission.Type + X..of.Gears + Drive.System.Description +
##     Equivalent.Test.Weight..lbs.. + Axle.Ratio + N.V.Ratio +
##     Test.Fuel.Type.Description + THC..g.mi. + RND_ADJ_FE + Target.Coef.A..lbf. +
##     Target.Coef.B..lbf.mph. + Target.Coef.C..lbf.mph..2. + Set.Coef.A..lbf. +
##     Set.Coef.B..lbf.mph. + Set.Coef.C..lbf.mph..2. + Aftertreatment.Device.Desc,
##     data = training.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -401.66  -23.91   -5.02   15.86  660.08
##
## Coefficients:
##                                                                                 Estimate
## (Intercept)                                                                    5.099e+02
## Model.Year.Cat2019                                                             1.613e+00
## Model.Year.Cat2020                                                             3.599e+00
## Model.Year.Cat2021                                                             3.247e+00
## Model.Year.Cat2022                                                             5.290e+00
## Test.Veh.Displacement..L.                                                      5.704e+00
## Vehicle.TypeCar                                                                6.454e+00
## Vehicle.TypeTruck                                                             -6.340e+00
## Rated.Horsepower                                                               4.075e-02
## X..of.Cylinders.and.Rotors                                                     5.967e+00
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles) -7.617e+00
## Tested.Transmission.TypeAutomatic                                             -5.333e+00
## Tested.Transmission.TypeContinuously Variable                                  1.315e+01
## Tested.Transmission.TypeManual                                                -1.093e+01
## Tested.Transmission.TypeOther                                                 -1.140e+01
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)  6.532e+00
## Tested.Transmission.TypeSemi-Automatic                                        -4.814e+00
## X..of.Gears                                                                   -1.413e+00
## Drive.System.Description2-Wheel Drive, Rear                                   -6.927e+00
## Drive.System.Description4-Wheel Drive                                         -3.626e-01
## Drive.System.DescriptionAll Wheel Drive                                       -3.793e+00
## Drive.System.DescriptionPart-time 4-Wheel Drive                               -2.007e+01
## Equivalent.Test.Weight..lbs..                                                 -6.195e-03
## Axle.Ratio                                                                    -6.546e+00
## N.V.Ratio                                                                      8.439e-01
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                              -1.024e+02
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                       4.620e+01
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                8.857e+01
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                            -1.057e+02
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                            -1.546e+02
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)         -1.632e+02
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                 -2.431e+01
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                -9.786e+01
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)        6.237e+01
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)       -1.684e+02
## THC..g.mi.                                                                     3.715e+02
## RND_ADJ_FE                                                                    -7.083e+00
## Target.Coef.A..lbf.                                                            5.449e-01
## Target.Coef.B..lbf.mph.                                                        3.675e+01
## Target.Coef.C..lbf.mph..2.                                                     1.507e+03
## Set.Coef.A..lbf.                                                              -2.512e-01
## Set.Coef.B..lbf.mph.                                                          -5.144e+00
## Set.Coef.C..lbf.mph..2.                                                        4.006e+01
## Aftertreatment.Device.DescNOx Adsorber                                         1.420e+00
## Aftertreatment.Device.DescOther                                               -2.061e+01
## Aftertreatment.Device.DescOxidation catalyst                                  -5.983e+00
## Aftertreatment.Device.DescSelective Catalytic Reduction                       -6.416e+00
## Aftertreatment.Device.DescThree-way catalyst                                   3.485e+01
##                                                                               Std. Error
## (Intercept)                                                                    2.227e+01
## Model.Year.Cat2019                                                             1.009e+00
```

```
## Model.Year.Cat2020                                                                1.041e+00
## Model.Year.Cat2021                                                                1.066e+00
## Model.Year.Cat2022                                                                1.085e+00
## Test.Veh.Displacement..L.                                                         7.804e-01
## Vehicle.TypeCar                                                                   1.168e+00
## Vehicle.TypeTruck                                                                 1.426e+00
## Rated.Horsepower                                                                  5.072e-03
## X..of.Cylinders.and.Rotors                                                        5.901e-01
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles)  2.186e+00
## Tested.Transmission.TypeAutomatic                                                 2.059e+00
## Tested.Transmission.TypeContinuously Variable                                     2.724e+00
## Tested.Transmission.TypeManual                                                    2.237e+00
## Tested.Transmission.TypeOther                                                     2.539e+01
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)  2.859e+00
## Tested.Transmission.TypeSemi-Automatic                                            2.000e+00
## X..of.Gears                                                                       2.805e-01
## Drive.System.Description2-Wheel Drive, Rear                                       1.051e+00
## Drive.System.Description4-Wheel Drive                                             2.156e+00
## Drive.System.DescriptionAll Wheel Drive                                           1.492e+00
## Drive.System.DescriptionPart-time 4-Wheel Drive                                   5.386e+00
## Equivalent.Test.Weight..lbs..                                                     8.526e-04
## Axle.Ratio                                                                        6.113e-01
## N.V.Ratio                                                                         6.770e-02
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                                  2.403e+01
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                          2.552e+01
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                   2.735e+01
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                                1.956e+01
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                                1.950e+01
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)             1.950e+01
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                     2.158e+01
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                    1.933e+01
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)           2.183e+01
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)           3.151e+01
## THC..g.mi.                                                                        1.067e+01
## RND_ADJ_FE                                                                        4.063e-02
## Target.Coef.A..lbf.                                                               5.457e-02
## Target.Coef.B..lbf.mph.                                                           1.678e+00
## Target.Coef.C..lbf.mph..2.                                                        8.853e+01
## Set.Coef.A..lbf.                                                                  4.115e-02
## Set.Coef.B..lbf.mph.                                                              1.120e+00
## Set.Coef.C..lbf.mph..2.                                                           3.165e+01
## Aftertreatment.Device.DescNOx Adsorber                                            7.344e+00
## Aftertreatment.Device.DescOther                                                   6.131e+00
## Aftertreatment.Device.DescOxidation catalyst                                      4.117e+00
## Aftertreatment.Device.DescSelective Catalytic Reduction                           4.031e+00
## Aftertreatment.Device.DescThree-way catalyst                                      9.988e+00
##                                                                                   t value
## (Intercept)                                                                        22.898
## Model.Year.Cat2019                                                                  1.598
## Model.Year.Cat2020                                                                  3.456
## Model.Year.Cat2021                                                                  3.045
## Model.Year.Cat2022                                                                  4.877
## Test.Veh.Displacement..L.                                                           7.309
## Vehicle.TypeCar                                                                     5.523
## Vehicle.TypeTruck                                                                  -4.446
## Rated.Horsepower                                                                    8.035
## X..of.Cylinders.and.Rotors                                                         10.111
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles)  -3.484
## Tested.Transmission.TypeAutomatic                                                  -2.590
## Tested.Transmission.TypeContinuously Variable                                       4.826
## Tested.Transmission.TypeManual                                                     -4.887
## Tested.Transmission.TypeOther                                                      -0.449
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)    2.285
## Tested.Transmission.TypeSemi-Automatic                                             -2.407
## X..of.Gears                                                                        -5.039
## Drive.System.Description2-Wheel Drive, Rear                                        -6.592
## Drive.System.Description4-Wheel Drive                                              -0.168
## Drive.System.DescriptionAll Wheel Drive                                            -2.542
```

```
## Drive.System.DescriptionPart-time 4-Wheel Drive                                      -3.726
## Equivalent.Test.Weight..lbs..                                                        -7.266
## Axle.Ratio                                                                          -10.708
## N.V.Ratio                                                                            12.464
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                                     -4.261
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                              1.810
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                       3.238
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                                   -5.407
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                                   -7.927
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)                -8.372
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                        -1.127
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                       -5.062
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)               2.857
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)              -5.344
## THC..g.mi.                                                                          34.810
## RND_ADJ_FE                                                                        -174.326
## Target.Coef.A..lbf.                                                                  9.985
## Target.Coef.B..lbf.mph.                                                             21.898
## Target.Coef.C..lbf.mph..2.                                                          17.026
## Set.Coef.A..lbf.                                                                    -6.103
## Set.Coef.B..lbf.mph.                                                                -4.592
## Set.Coef.C..lbf.mph..2.                                                              1.266
## Aftertreatment.Device.DescNOx Adsorber                                               0.193
## Aftertreatment.Device.DescOther                                                     -3.361
## Aftertreatment.Device.DescOxidation catalyst                                        -1.453
## Aftertreatment.Device.DescSelective Catalytic Reduction                             -1.591
## Aftertreatment.Device.DescThree-way catalyst                                         3.489
##                                                                                  Pr(>|t|)
## (Intercept)                                                                       < 2e-16
## Model.Year.Cat2019                                                               0.110063
## Model.Year.Cat2020                                                               0.000549
## Model.Year.Cat2021                                                               0.002329
## Model.Year.Cat2022                                                               1.08e-06
## Test.Veh.Displacement..L.                                                        2.80e-13
## Vehicle.TypeCar                                                                  3.37e-08
## Vehicle.TypeTruck                                                                8.79e-06
## Rated.Horsepower                                                                 9.98e-16
## X..of.Cylinders.and.Rotors                                                        < 2e-16
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles) 0.000495
## Tested.Transmission.TypeAutomatic                                                0.009617
## Tested.Transmission.TypeContinuously Variable                                    1.40e-06
## Tested.Transmission.TypeManual                                                   1.03e-06
## Tested.Transmission.TypeOther                                                    0.653423
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles) 0.022339
## Tested.Transmission.TypeSemi-Automatic                                           0.016096
## X..of.Gears                                                                      4.72e-07
## Drive.System.Description2-Wheel Drive, Rear                                      4.46e-11
## Drive.System.Description4-Wheel Drive                                            0.866444
## Drive.System.DescriptionAll Wheel Drive                                          0.011019
## Drive.System.DescriptionPart-time 4-Wheel Drive                                  0.000195
## Equivalent.Test.Weight..lbs..                                                    3.85e-13
## Axle.Ratio                                                                        < 2e-16
## N.V.Ratio                                                                         < 2e-16
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                                 2.05e-05
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                         0.070248
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                  0.001204
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                               6.48e-08
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                               2.37e-15
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)             < 2e-16
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                    0.259942
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                   4.20e-07
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)          0.004280
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)          9.21e-08
## THC..g.mi.                                                                        < 2e-16
## RND_ADJ_FE                                                                        < 2e-16
## Target.Coef.A..lbf.                                                               < 2e-16
## Target.Coef.B..lbf.mph.                                                           < 2e-16
## Target.Coef.C..lbf.mph..2.                                                        < 2e-16
```

```
## Set.Coef.A..lbf.                                                                1.06e-09
## Set.Coef.B..lbf.mph.                                                            4.41e-06
## Set.Coef.C..lbf.mph..2.                                                         0.205595
## Aftertreatment.Device.DescNOx Adsorber                                          0.846729
## Aftertreatment.Device.DescOther                                                 0.000780
## Aftertreatment.Device.DescOxidation catalyst                                    0.146131
## Aftertreatment.Device.DescSelective Catalytic Reduction                         0.111524
## Aftertreatment.Device.DescThree-way catalyst                                    0.000487
##
## (Intercept)                                                                     ***
## Model.Year.Cat2019
## Model.Year.Cat2020                                                              ***
## Model.Year.Cat2021                                                              **
## Model.Year.Cat2022                                                              ***
## Test.Veh.Displacement..L.                                                       ***
## Vehicle.TypeCar                                                                 ***
## Vehicle.TypeTruck                                                               ***
## Rated.Horsepower                                                                ***
## X..of.Cylinders.and.Rotors                                                      ***
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles) ***
## Tested.Transmission.TypeAutomatic                                               **
## Tested.Transmission.TypeContinuously Variable                                   ***
## Tested.Transmission.TypeManual                                                  ***
## Tested.Transmission.TypeOther
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles) *
## Tested.Transmission.TypeSemi-Automatic                                          *
## X..of.Gears                                                                     ***
## Drive.System.Description2-Wheel Drive, Rear                                     ***
## Drive.System.Description4-Wheel Drive
## Drive.System.DescriptionAll Wheel Drive                                         *
## Drive.System.DescriptionPart-time 4-Wheel Drive                                 ***
## Equivalent.Test.Weight..lbs..                                                   ***
## Axle.Ratio                                                                      ***
## N.V.Ratio                                                                       ***
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                                ***
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                        .
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                 **
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                              ***
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                              ***
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)           ***
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                  ***
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)         **
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)         ***
## THC..g.mi.                                                                      ***
## RND_ADJ_FE                                                                      ***
## Target.Coef.A..lbf.                                                             ***
## Target.Coef.B..lbf.mph.                                                         ***
## Target.Coef.C..lbf.mph..2.                                                      ***
## Set.Coef.A..lbf.                                                                ***
## Set.Coef.B..lbf.mph.                                                            ***
## Set.Coef.C..lbf.mph..2.
## Aftertreatment.Device.DescNOx Adsorber
## Aftertreatment.Device.DescOther                                                 ***
## Aftertreatment.Device.DescOxidation catalyst
## Aftertreatment.Device.DescSelective Catalytic Reduction
## Aftertreatment.Device.DescThree-way catalyst                                    ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.64 on 17190 degrees of freedom
## Multiple R-squared:  0.8543, Adjusted R-squared:  0.8539
## F-statistic:  2145 on 47 and 17190 DF,  p-value: < 2.2e-16
```

For the third and final model, we will remove the last of the insignificant variables below the 0.05 significance level, as well as those categorical variables where less than half of the dummy variables are significant. So, because of this, we will remove Set.Coef.C..lbf.mph..2.and also remove the categorical variable (and associated dummy variables) of Aftertreatment.Device.Desc.

Emissions Model 3: Removing all insignificant terms

```
emissions.model.3 = lm(CO2..g.mi. ~ Model.Year.Cat + Test.Veh.Displacement..L. + Vehicle.Type + Rated.Horsepower
+ X..of.Cylinders.and.Rotors + Tested.Transmission.Type + X..of.Gears + Drive.System.Description + Equivalent.Tes
t.Weight..lbs.. + Axle.Ratio + N.V.Ratio + Test.Fuel.Type.Description + THC..g.mi. + RND_ADJ_FE + Target.Coef.A..
lbf. + Target.Coef.B..lbf.mph. + Target.Coef.C..lbf.mph..2. + Set.Coef.A..lbf. + Set.Coef.B..lbf.mph., data = tra
ining.data)
```

```
summary(emissions.model.3)
```

```
##
## Call:
## lm(formula = CO2..g.mi. ~ Model.Year.Cat + Test.Veh.Displacement..L. +
##     Vehicle.Type + Rated.Horsepower + X..of.Cylinders.and.Rotors +
##     Tested.Transmission.Type + X..of.Gears + Drive.System.Description +
##     Equivalent.Test.Weight..lbs.. + Axle.Ratio + N.V.Ratio +
##     Test.Fuel.Type.Description + THC..g.mi. + RND_ADJ_FE + Target.Coef.A..lbf. +
##     Target.Coef.B..lbf.mph. + Target.Coef.C..lbf.mph..2. + Set.Coef.A..lbf. +
##     Set.Coef.B..lbf.mph., data = training.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -401.15  -23.83   -5.08   15.86  658.98
##
## Coefficients:
##                                                                                 Estimate
## (Intercept)                                                                    5.454e+02
## Model.Year.Cat2019                                                             1.585e+00
## Model.Year.Cat2020                                                             3.593e+00
## Model.Year.Cat2021                                                             3.038e+00
## Model.Year.Cat2022                                                             4.955e+00
## Test.Veh.Displacement..L.                                                      6.016e+00
## Vehicle.TypeCar                                                                6.217e+00
## Vehicle.TypeTruck                                                             -6.496e+00
## Rated.Horsepower                                                               3.818e-02
## X..of.Cylinders.and.Rotors                                                     5.865e+00
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles) -5.770e+00
## Tested.Transmission.TypeAutomatic                                             -3.771e+00
## Tested.Transmission.TypeContinuously Variable                                  1.415e+01
## Tested.Transmission.TypeManual                                                -9.368e+00
## Tested.Transmission.TypeOther                                                 -9.975e+00
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)  8.109e+00
## Tested.Transmission.TypeSemi-Automatic                                        -2.917e+00
## X..of.Gears                                                                   -1.510e+00
## Drive.System.Description2-Wheel Drive, Rear                                   -6.839e+00
## Drive.System.Description4-Wheel Drive                                          5.472e-01
## Drive.System.DescriptionAll Wheel Drive                                       -3.482e+00
## Drive.System.DescriptionPart-time 4-Wheel Drive                               -1.955e+01
## Equivalent.Test.Weight..lbs..                                                 -6.083e-03
## Axle.Ratio                                                                    -6.750e+00
## N.V.Ratio                                                                      8.521e-01
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                              -1.025e+02
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                       8.073e+00
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                8.875e+01
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                            -1.060e+02
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                            -1.545e+02
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)         -1.634e+02
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                 -6.461e+01
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                -9.775e+01
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)        6.343e+01
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)       -1.681e+02
## THC..g.mi.                                                                     3.750e+02
## RND_ADJ_FE                                                                    -7.089e+00
## Target.Coef.A..lbf.                                                            5.231e-01
## Target.Coef.B..lbf.mph.                                                        3.596e+01
## Target.Coef.C..lbf.mph..2.                                                     1.525e+03
## Set.Coef.A..lbf.                                                              -2.200e-01
## Set.Coef.B..lbf.mph.                                                          -4.480e+00
##                                                                              Std. Error
## (Intercept)                                                                    2.004e+01
## Model.Year.Cat2019                                                             1.010e+00
## Model.Year.Cat2020                                                             1.042e+00
## Model.Year.Cat2021                                                             1.067e+00
## Model.Year.Cat2022                                                             1.085e+00
## Test.Veh.Displacement..L.                                                      7.799e-01
## Vehicle.TypeCar                                                                1.168e+00
## Vehicle.TypeTruck                                                              1.427e+00
## Rated.Horsepower                                                               5.067e-03
```

```
## X..of.Cylinders.and.Rotors                                                          5.905e-01
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles)  2.169e+00
## Tested.Transmission.TypeAutomatic                                                   2.046e+00
## Tested.Transmission.TypeContinuously Variable                                       2.715e+00
## Tested.Transmission.TypeManual                                                      2.224e+00
## Tested.Transmission.TypeOther                                                       2.543e+01
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)    2.847e+00
## Tested.Transmission.TypeSemi-Automatic                                              1.983e+00
## X..of.Gears                                                                         2.804e-01
## Drive.System.Description2-Wheel Drive, Rear                                         1.052e+00
## Drive.System.Description4-Wheel Drive                                               2.154e+00
## Drive.System.DescriptionAll Wheel Drive                                             1.492e+00
## Drive.System.DescriptionPart-time 4-Wheel Drive                                     5.394e+00
## Equivalent.Test.Weight..lbs..                                                       8.527e-04
## Axle.Ratio                                                                          6.115e-01
## N.V.Ratio                                                                           6.779e-02
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                                    2.407e+01
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                            2.368e+01
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                     2.739e+01
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                                  1.959e+01
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                                  1.953e+01
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)               1.953e+01
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                       1.942e+01
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                      1.936e+01
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)             2.186e+01
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)             3.156e+01
## THC..g.mi.                                                                         1.068e+01
## RND_ADJ_FE                                                                         4.068e-02
## Target.Coef.A..lbf.                                                                5.455e-02
## Target.Coef.B..lbf.mph.                                                            1.667e+00
## Target.Coef.C..lbf.mph..2.                                                         8.436e+01
## Set.Coef.A..lbf.                                                                   4.092e-02
## Set.Coef.B..lbf.mph.                                                               1.116e+00
##                                                                                    t value
## (Intercept)                                                                         27.213
## Model.Year.Cat2019                                                                   1.570
## Model.Year.Cat2020                                                                   3.450
## Model.Year.Cat2021                                                                   2.848
## Model.Year.Cat2022                                                                   4.566
## Test.Veh.Displacement..L.                                                            7.714
## Vehicle.TypeCar                                                                      5.321
## Vehicle.TypeTruck                                                                   -4.552
## Rated.Horsepower                                                                     7.535
## X..of.Cylinders.and.Rotors                                                           9.933
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles)  -2.660
## Tested.Transmission.TypeAutomatic                                                   -1.844
## Tested.Transmission.TypeContinuously Variable                                        5.212
## Tested.Transmission.TypeManual                                                      -4.212
## Tested.Transmission.TypeOther                                                       -0.392
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)     2.848
## Tested.Transmission.TypeSemi-Automatic                                              -1.471
## X..of.Gears                                                                         -5.386
## Drive.System.Description2-Wheel Drive, Rear                                         -6.500
## Drive.System.Description4-Wheel Drive                                                0.254
## Drive.System.DescriptionAll Wheel Drive                                             -2.334
## Drive.System.DescriptionPart-time 4-Wheel Drive                                     -3.623
## Equivalent.Test.Weight..lbs..                                                       -7.134
## Axle.Ratio                                                                         -11.039
## N.V.Ratio                                                                           12.569
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                                    -4.257
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                             0.341
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                      3.240
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                                  -5.413
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                                  -7.912
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)               -8.366
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                       -3.327
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                      -5.048
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)              2.901
```

```
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)                    -5.327
## THC..g.mi.                                                                                 35.123
## RND_ADJ_FE                                                                               -174.250
## Target.Coef.A..lbf.                                                                         9.589
## Target.Coef.B..lbf.mph.                                                                    21.565
## Target.Coef.C..lbf.mph..2.                                                                 18.071
## Set.Coef.A..lbf.                                                                           -5.377
## Set.Coef.B..lbf.mph.                                                                       -4.014
##                                                                                          Pr(>|t|)
## (Intercept)                                                                               < 2e-16
## Model.Year.Cat2019                                                                       0.116502
## Model.Year.Cat2020                                                                       0.000562
## Model.Year.Cat2021                                                                       0.004406
## Model.Year.Cat2022                                                                       5.00e-06
## Test.Veh.Displacement..L.                                                                1.28e-14
## Vehicle.TypeCar                                                                          1.05e-07
## Vehicle.TypeTruck                                                                        5.35e-06
## Rated.Horsepower                                                                         5.13e-14
## X..of.Cylinders.and.Rotors                                                                < 2e-16
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles) 0.007813
## Tested.Transmission.TypeAutomatic                                                        0.065256
## Tested.Transmission.TypeContinuously Variable                                            1.89e-07
## Tested.Transmission.TypeManual                                                           2.54e-05
## Tested.Transmission.TypeOther                                                            0.694906
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)         0.004406
## Tested.Transmission.TypeSemi-Automatic                                                   0.141392
## X..of.Gears                                                                              7.29e-08
## Drive.System.Description2-Wheel Drive, Rear                                              8.25e-11
## Drive.System.Description4-Wheel Drive                                                    0.799431
## Drive.System.DescriptionAll Wheel Drive                                                  0.019608
## Drive.System.DescriptionPart-time 4-Wheel Drive                                          0.000292
## Equivalent.Test.Weight..lbs..                                                            1.02e-12
## Axle.Ratio                                                                                < 2e-16
## N.V.Ratio                                                                                 < 2e-16
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                                         2.09e-05
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                                 0.733168
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                          0.001198
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                                       6.28e-08
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                                       2.69e-15
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)                     < 2e-16
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                            0.000881
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                           4.52e-07
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)                  0.003720
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)                  1.01e-07
## THC..g.mi.                                                                                < 2e-16
## RND_ADJ_FE                                                                                < 2e-16
## Target.Coef.A..lbf.                                                                       < 2e-16
## Target.Coef.B..lbf.mph.                                                                   < 2e-16
## Target.Coef.C..lbf.mph..2.                                                                < 2e-16
## Set.Coef.A..lbf.                                                                         7.66e-08
## Set.Coef.B..lbf.mph.                                                                     6.00e-05
##
## (Intercept)                                                                              ***
## Model.Year.Cat2019
## Model.Year.Cat2020                                                                       ***
## Model.Year.Cat2021                                                                       **
## Model.Year.Cat2022                                                                       ***
## Test.Veh.Displacement..L.                                                                ***
## Vehicle.TypeCar                                                                          ***
## Vehicle.TypeTruck                                                                        ***
## Rated.Horsepower                                                                         ***
## X..of.Cylinders.and.Rotors                                                               ***
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles) **
## Tested.Transmission.TypeAutomatic                                                        .
## Tested.Transmission.TypeContinuously Variable                                            ***
## Tested.Transmission.TypeManual                                                           ***
## Tested.Transmission.TypeOther
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)         **
```

```
## Tested.Transmission.TypeSemi-Automatic
## X..of.Gears                                                                  ***
## Drive.System.Description2-Wheel Drive, Rear                                  ***
## Drive.System.Description4-Wheel Drive
## Drive.System.DescriptionAll Wheel Drive                                      *
## Drive.System.DescriptionPart-time 4-Wheel Drive                              ***
## Equivalent.Test.Weight..lbs..                                                ***
## Axle.Ratio                                                                   ***
## N.V.Ratio                                                                    ***
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                             ***
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)              **
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                           ***
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                           ***
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)        ***
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                ***
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                               ***
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)      **
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)      ***
## THC..g.mi.                                                                   ***
## RND_ADJ_FE                                                                   ***
## Target.Coef.A..lbf.                                                          ***
## Target.Coef.B..lbf.mph.                                                      ***
## Target.Coef.C..lbf.mph..2.                                                   ***
## Set.Coef.A..lbf.                                                             ***
## Set.Coef.B..lbf.mph.                                                         ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.71 on 17196 degrees of freedom
## Multiple R-squared:  0.8538, Adjusted R-squared:  0.8534
## F-statistic:  2449 on 41 and 17196 DF,  p-value: < 2.2e-16
```

Now, the metrics of the three models can be compared to determine which one can be used to best predict a car's CO_2 emissions.

```
pred1 = emissions.model %>% predict(testing.data)
p1 = data.frame(
  RMSE = RMSE(pred1, testing.data$CO2..g.mi.),
  R2 = R2(pred1, testing.data$CO2..g.mi.)
)
pred2 = emissions.model.2 %>% predict(testing.data)
p2 = data.frame(
  RMSE = RMSE(pred2, testing.data$CO2..g.mi.),
  R2 = R2(pred2, testing.data$CO2..g.mi.)
)
pred3 = emissions.model.3 %>% predict(testing.data)
p3 = data.frame(
  RMSE = RMSE(pred3, testing.data$CO2..g.mi.),
  R2 = R2(pred3, testing.data$CO2..g.mi.)
)
```

```
combined = rbind(p1, p2, p3)

combined = cbind(combined, c(summary(emissions.model)$fstatistic[1], summary(emissions.model.2)$fstatistic[1], su
mmary(emissions.model.3)$fstatistic[1]))

combined=cbind(combined, c(summary(emissions.model)$adj.r.squared, summary(emissions.model.2)$adj.r.squared, summ
ary(emissions.model.3)$adj.r.squared))

combined=cbind(combined,c(summary(emissions.model)$sigma,summary(emissions.model.2)$sigma, summary(emissions.mode
l.3)$sigma))

combined=cbind(combined, c("Model 1", "Model 2", "Model 3"))
colnames(combined)[c(3,4,5,6)] = c("F-Statistic", "Adj R2", "RSE", "Model Name")
```

```
library(kableExtra)
```

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

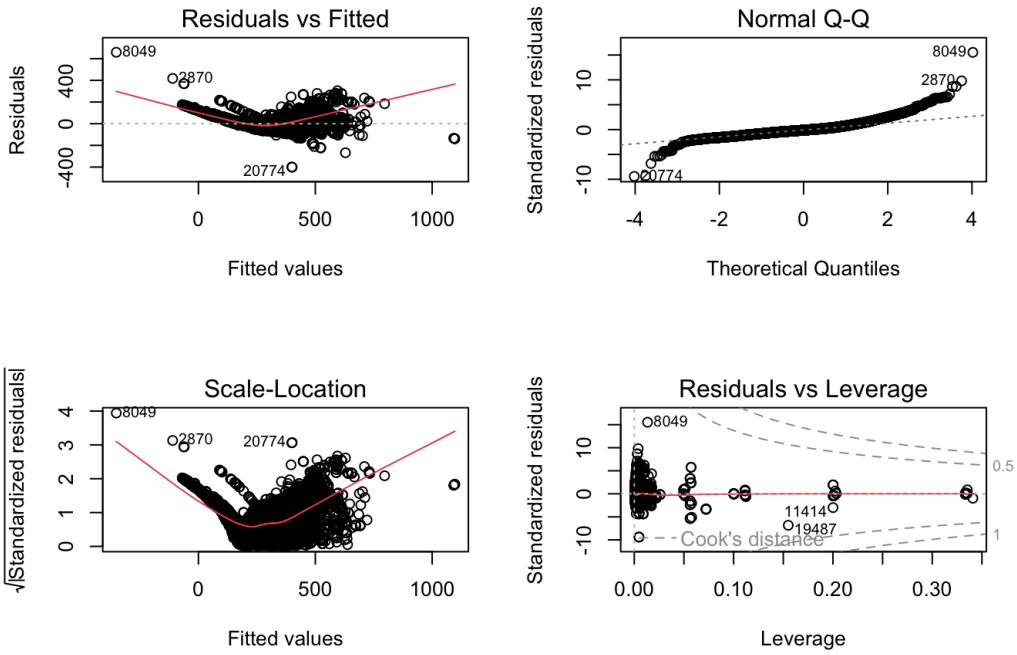
```
combined %>%
  kbl() %>%
  kable_classic(full_width = F, html_font = "Cambria")
```

RMSE	R2	F-Statistic	Adj R2	RSE	Model Name
67.19786	0.6954633	2002.677	0.8555428	42.40103	Model 1
67.93032	0.6903372	2144.816	0.8539192	42.63865	Model 2
67.94107	0.6903654	2449.245	0.8534456	42.70771	Model 3

Overall, the models seem to perform relatively equally, with Adjusted R^2 scores around 85% for all models. However, because Model 3 has the least number of predictor terms in the model, and knowing that the addition of predictor terms inflates the R^2, we can say that Model 3 is the best predictor of a car's emissions.

Lastly, we want to check for any outliers or high leverage points in the chosen model.

```
par(mfrow=c(2,2))
plot(emissions.model.3)
```



Looking at the plots above, particularly the Residuals vs Fitted and the Scale-Location plots, we can see that linearity appears to be violated. Due to the parabola shape of the data, it is possible that quadratic regression could be a better fit for this data.

To see if a quadratic term could improve this model, we will add single squared regression term to the predictor variables.

Because the predictor with the highest influence on the model (largest F-statistic) is RND_ADJ_FE, or Miles per Gallon, with an F-statistic of -174.250, we will add a quadratic term for this predictor to see if it improves the model.

```
nonelectric$RND_ADJ_FE_2 = nonelectric$RND_ADJ_FE^2
```

```
set.seed(101)

training.samples = nonelectric$CO2..g.mi. %>%
  createDataPartition(p = 0.8, list = FALSE)

training.data = nonelectric[training.samples,]
testing.data = nonelectric[-training.samples, ]
```

Emissions Model 4: Quadratic Regression Model

```
emissions.model.4 = lm(CO2..g.mi. ~ Model.Year.Cat + Test.Veh.Displacement..L. + Vehicle.Type + Rated.Horsepower
+ X..of.Cylinders.and.Rotors + Tested.Transmission.Type + X..of.Gears + Drive.System.Description + Equivalent.Tes
t.Weight..lbs.. + Axle.Ratio + N.V.Ratio + Test.Fuel.Type.Description + THC..g.mi. + RND_ADJ_FE + Target.Coef.A..
lbf. + Target.Coef.B..lbf.mph. + Target.Coef.C..lbf.mph..2. + Set.Coef.A..lbf. + Set.Coef.B..lbf.mph. + RND_ADJ_F
E_2, data = training.data)
```

```
summary(emissions.model.4)
```

```
##
## Call:
## lm(formula = CO2..g.mi. ~ Model.Year.Cat + Test.Veh.Displacement..L. +
##     Vehicle.Type + Rated.Horsepower + X..of.Cylinders.and.Rotors +
##     Tested.Transmission.Type + X..of.Gears + Drive.System.Description +
##     Equivalent.Test.Weight..lbs.. + Axle.Ratio + N.V.Ratio +
##     Test.Fuel.Type.Description + THC..g.mi. + RND_ADJ_FE + Target.Coef.A..lbf. +
##     Target.Coef.B..lbf.mph. + Target.Coef.C..lbf.mph..2. + Set.Coef.A..lbf. +
##     Set.Coef.B..lbf.mph. + RND_ADJ_FE_2, data = training.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -641.54  -11.26   -1.61   10.92  244.82
##
## Coefficients:
##                                                                                 Estimate
## (Intercept)                                                                    8.568e+02
## Model.Year.Cat2019                                                             5.260e-01
## Model.Year.Cat2020                                                             1.715e+00
## Model.Year.Cat2021                                                             7.048e-01
## Model.Year.Cat2022                                                             1.855e+00
## Test.Veh.Displacement..L.                                                      3.074e+00
## Vehicle.TypeCar                                                               -6.975e-01
## Vehicle.TypeTruck                                                             -5.158e+00
## Rated.Horsepower                                                               3.217e-02
## X..of.Cylinders.and.Rotors                                                     1.751e+00
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles)  4.313e+00
## Tested.Transmission.TypeAutomatic                                              7.911e+00
## Tested.Transmission.TypeContinuously Variable                                  1.028e+01
## Tested.Transmission.TypeManual                                                 4.536e+00
## Tested.Transmission.TypeOther                                                  1.105e+01
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)  1.243e+01
## Tested.Transmission.TypeSemi-Automatic                                         9.087e+00
## X..of.Gears                                                                   -7.366e-01
## Drive.System.Description2-Wheel Drive, Rear                                   -8.034e+00
## Drive.System.Description4-Wheel Drive                                         -7.545e+00
## Drive.System.DescriptionAll Wheel Drive                                       -7.647e+00
## Drive.System.DescriptionPart-time 4-Wheel Drive                               -1.764e+01
## Equivalent.Test.Weight..lbs..                                                 -4.306e-03
## Axle.Ratio                                                                    -1.095e+00
## N.V.Ratio                                                                      4.163e-01
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                              -1.075e+02
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                      -3.731e+01
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                1.618e+02
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                            -1.263e+02
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                            -1.380e+02
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)         -2.192e+02
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                 -7.364e+01
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                -1.126e+02
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)        8.883e+01
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)       -1.751e+02
## THC..g.mi.                                                                     2.095e+02
## RND_ADJ_FE                                                                    -2.140e+01
## Target.Coef.A..lbf.                                                            3.796e-01
## Target.Coef.B..lbf.mph.                                                        1.446e+01
## Target.Coef.C..lbf.mph..2.                                                     6.247e+02
## Set.Coef.A..lbf.                                                              -1.826e-01
## Set.Coef.B..lbf.mph.                                                          -6.985e-01
## RND_ADJ_FE_2                                                                   1.777e-01
##                                                                               Std. Error
## (Intercept)                                                                    1.209e+01
## Model.Year.Cat2019                                                             6.028e-01
## Model.Year.Cat2020                                                             6.216e-01
## Model.Year.Cat2021                                                             6.368e-01
## Model.Year.Cat2022                                                             6.477e-01
## Test.Veh.Displacement..L.                                                      4.657e-01
## Vehicle.TypeCar                                                                6.984e-01
## Vehicle.TypeTruck                                                              8.517e-01
```

```
## Rated.Horsepower                                                                3.024e-03
## X..of.Cylinders.and.Rotors                                                      3.531e-01
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles)  1.296e+00
## Tested.Transmission.TypeAutomatic                                               1.223e+00
## Tested.Transmission.TypeContinuously Variable                                   1.621e+00
## Tested.Transmission.TypeManual                                                  1.330e+00
## Tested.Transmission.TypeOther                                                   1.518e+01
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)  1.699e+00
## Tested.Transmission.TypeSemi-Automatic                                          1.186e+00
## X..of.Gears                                                                     1.674e-01
## Drive.System.Description2-Wheel Drive, Rear                                     6.279e-01
## Drive.System.Description4-Wheel Drive                                           1.286e+00
## Drive.System.DescriptionAll Wheel Drive                                         8.905e-01
## Drive.System.DescriptionPart-time 4-Wheel Drive                                 3.219e+00
## Equivalent.Test.Weight..lbs..                                                   5.089e-04
## Axle.Ratio                                                                      3.663e-01
## N.V.Ratio                                                                       4.053e-02
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                                1.437e+01
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                        1.413e+01
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                 1.635e+01
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                              1.169e+01
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                              1.165e+01
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)           1.166e+01
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                   1.159e+01
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                  1.156e+01
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)         1.305e+01
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)         1.883e+01
## THC..g.mi.                                                                     6.440e+00
## RND_ADJ_FE                                                                     8.473e-02
## Target.Coef.A..lbf.                                                            3.256e-02
## Target.Coef.B..lbf.mph.                                                        1.002e+00
## Target.Coef.C..lbf.mph..2.                                                     5.060e+01
## Set.Coef.A..lbf.                                                               2.442e-02
## Set.Coef.B..lbf.mph.                                                           6.664e-01
## RND_ADJ_FE_2                                                                   1.007e-03
##                                                                                t value
## (Intercept)                                                                     70.860
## Model.Year.Cat2019                                                               0.873
## Model.Year.Cat2020                                                               2.758
## Model.Year.Cat2021                                                               1.107
## Model.Year.Cat2022                                                               2.864
## Test.Veh.Displacement..L.                                                        6.601
## Vehicle.TypeCar                                                                 -0.999
## Vehicle.TypeTruck                                                               -6.056
## Rated.Horsepower                                                                10.640
## X..of.Cylinders.and.Rotors                                                       4.959
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles)   3.329
## Tested.Transmission.TypeAutomatic                                                6.470
## Tested.Transmission.TypeContinuously Variable                                    6.346
## Tested.Transmission.TypeManual                                                   3.412
## Tested.Transmission.TypeOther                                                    0.728
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)   7.317
## Tested.Transmission.TypeSemi-Automatic                                           7.665
## X..of.Gears                                                                     -4.401
## Drive.System.Description2-Wheel Drive, Rear                                    -12.795
## Drive.System.Description4-Wheel Drive                                           -5.867
## Drive.System.DescriptionAll Wheel Drive                                         -8.588
## Drive.System.DescriptionPart-time 4-Wheel Drive                                 -5.479
## Equivalent.Test.Weight..lbs..                                                   -8.461
## Axle.Ratio                                                                      -2.988
## N.V.Ratio                                                                       10.271
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                                -7.484
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                        -2.640
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                  9.891
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                             -10.805
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                             -11.843
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)          -18.798
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                   -6.353
```

```
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                                  -9.740
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)                          6.808
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)                         -9.296
## THC..g.mi.                                                                                      32.537
## RND_ADJ_FE                                                                                    -252.592
## Target.Coef.A..lbf.                                                                             11.657
## Target.Coef.B..lbf.mph.                                                                         14.421
## Target.Coef.C..lbf.mph..2.                                                                      12.344
## Set.Coef.A..lbf.                                                                                -7.478
## Set.Coef.B..lbf.mph.                                                                            -1.048
## RND_ADJ_FE_2                                                                                   176.325
##                                                                                               Pr(>|t|)
## (Intercept)                                                                                    < 2e-16
## Model.Year.Cat2019                                                                            0.382908
## Model.Year.Cat2020                                                                            0.005820
## Model.Year.Cat2021                                                                            0.268368
## Model.Year.Cat2022                                                                            0.004192
## Test.Veh.Displacement..L.                                                                     4.20e-11
## Vehicle.TypeCar                                                                               0.317984
## Vehicle.TypeTruck                                                                             1.42e-09
## Rated.Horsepower                                                                               < 2e-16
## X..of.Cylinders.and.Rotors                                                                    7.16e-07
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles)     0.000874
## Tested.Transmission.TypeAutomatic                                                             1.00e-10
## Tested.Transmission.TypeContinuously Variable                                                 2.27e-10
## Tested.Transmission.TypeManual                                                                0.000648
## Tested.Transmission.TypeOther                                                                 0.466720
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)              2.65e-13
## Tested.Transmission.TypeSemi-Automatic                                                        1.89e-14
## X..of.Gears                                                                                   1.08e-05
## Drive.System.Description2-Wheel Drive, Rear                                                    < 2e-16
## Drive.System.Description4-Wheel Drive                                                         4.52e-09
## Drive.System.DescriptionAll Wheel Drive                                                        < 2e-16
## Drive.System.DescriptionPart-time 4-Wheel Drive                                               4.34e-08
## Equivalent.Test.Weight..lbs..                                                                  < 2e-16
## Axle.Ratio                                                                                    0.002809
## N.V.Ratio                                                                                      < 2e-16
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                                              7.54e-14
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                                      0.008301
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                                < 2e-16
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                                             < 2e-16
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                                             < 2e-16
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)                          < 2e-16
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                                 2.16e-10
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                                 < 2e-16
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)                       1.02e-11
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)                        < 2e-16
## THC..g.mi.                                                                                     < 2e-16
## RND_ADJ_FE                                                                                     < 2e-16
## Target.Coef.A..lbf.                                                                            < 2e-16
## Target.Coef.B..lbf.mph.                                                                        < 2e-16
## Target.Coef.C..lbf.mph..2.                                                                     < 2e-16
## Set.Coef.A..lbf.                                                                              7.88e-14
## Set.Coef.B..lbf.mph.                                                                          0.294588
## RND_ADJ_FE_2                                                                                   < 2e-16
##
## (Intercept)                                                                                   ***
## Model.Year.Cat2019
## Model.Year.Cat2020                                                                            **
## Model.Year.Cat2021
## Model.Year.Cat2022                                                                            **
## Test.Veh.Displacement..L.                                                                     ***
## Vehicle.TypeCar
## Vehicle.TypeTruck                                                                             ***
## Rated.Horsepower                                                                              ***
## X..of.Cylinders.and.Rotors                                                                    ***
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles)     ***
## Tested.Transmission.TypeAutomatic                                                             ***
```

```
## Tested.Transmission.TypeContinuously Variable                                    ***
## Tested.Transmission.TypeManual                                                   ***
## Tested.Transmission.TypeOther
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles) ***
## Tested.Transmission.TypeSemi-Automatic                                           ***
## X..of.Gears                                                                      ***
## Drive.System.Description2-Wheel Drive, Rear                                      ***
## Drive.System.Description4-Wheel Drive                                            ***
## Drive.System.DescriptionAll Wheel Drive                                          ***
## Drive.System.DescriptionPart-time 4-Wheel Drive                                  ***
## Equivalent.Test.Weight..lbs..                                                    ***
## Axle.Ratio                                                                       **
## N.V.Ratio                                                                        ***
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                                 ***
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                         **
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                  ***
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                               ***
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                               ***
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)            ***
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                    ***
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                   ***
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)          ***
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)          ***
## THC..g.mi.                                                                       ***
## RND_ADJ_FE                                                                       ***
## Target.Coef.A..lbf.                                                              ***
## Target.Coef.B..lbf.mph.                                                          ***
## Target.Coef.C..lbf.mph..2.                                                       ***
## Set.Coef.A..lbf.                                                                 ***
## Set.Coef.B..lbf.mph.
## RND_ADJ_FE_2                                                                     ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.49 on 17195 degrees of freedom
## Multiple R-squared:  0.9479, Adjusted R-squared:  0.9478
## F-statistic:  7454 on 42 and 17195 DF,  p-value: < 2.2e-16
```

```
pred4 = emissions.model.4 %>% predict(testing.data)
p4 = data.frame(
  RMSE = RMSE(pred4, testing.data$CO2..g.mi.),
  R2 = R2(pred4, testing.data$CO2..g.mi.)
)
```

```
combined = rbind(p1, p2, p3, p4)

combined = cbind(combined, c(summary(emissions.model)$fstatistic[1], summary(emissions.model.2)$fstatistic[1], su
mmary(emissions.model.3)$fstatistic[1], summary(emissions.model.4)$fstatistic[1]))

combined=cbind(combined, c(summary(emissions.model)$adj.r.squared, summary(emissions.model.2)$adj.r.squared, summ
ary(emissions.model.3)$adj.r.squared, summary(emissions.model.4)$adj.r.squared))

combined=cbind(combined,c(summary(emissions.model)$sigma,summary(emissions.model.2)$sigma, summary(emissions.mode
l.3)$sigma, summary(emissions.model.4)$sigma))

combined=cbind(combined, c("Model 1", "Model 2", "Model 3", "Model 4"))
colnames(combined)[c(3,4,5,6)] = c("F-Statistic", "Adj R2", "RSE", "Model Name")
```

```
library(kableExtra)
combined %>%
  kbl() %>%
  kable_classic(full_width = F, html_font = "Cambria")
```

RMSE	R2	F-Statistic	Adj R2	RSE	Model Name
67.19786	0.6954633	2002.677	0.8555428	42.40103	Model 1
67.93032	0.6903372	2144.816	0.8539192	42.63865	Model 2

RMSE	R2	F-Statistic	Adj R2	RSE	Model Name
67.94107	0.6903654	2449.245	0.8534456	42.70771	Model 3
390.62920	0.0712754	7453.846	0.9478072	25.48660	Model 4

Results

As we can see in the model results, Model 4 appears to be a huge improvement in every way on the previous Model 3. However, when looking at the RMSE for Model 4, we can see that it is extremely large, suggesting that this quadratic regression is badly overfitting, and therefore is not a good predictor.

Therefore, the best model is still the multiple linear regression model of Model 3.

```
summary(emissions.model.3)
```

```
##
## Call:
## lm(formula = CO2..g.mi. ~ Model.Year.Cat + Test.Veh.Displacement..L. +
##     Vehicle.Type + Rated.Horsepower + X..of.Cylinders.and.Rotors +
##     Tested.Transmission.Type + X..of.Gears + Drive.System.Description +
##     Equivalent.Test.Weight..lbs.. + Axle.Ratio + N.V.Ratio +
##     Test.Fuel.Type.Description + THC..g.mi. + RND_ADJ_FE + Target.Coef.A..lbf. +
##     Target.Coef.B..lbf.mph. + Target.Coef.C..lbf.mph..2. + Set.Coef.A..lbf. +
##     Set.Coef.B..lbf.mph., data = training.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -401.15  -23.83   -5.08   15.86  658.98
##
## Coefficients:
##                                                                                 Estimate
## (Intercept)                                                                    5.454e+02
## Model.Year.Cat2019                                                             1.585e+00
## Model.Year.Cat2020                                                             3.593e+00
## Model.Year.Cat2021                                                             3.038e+00
## Model.Year.Cat2022                                                             4.955e+00
## Test.Veh.Displacement..L.                                                      6.016e+00
## Vehicle.TypeCar                                                                6.217e+00
## Vehicle.TypeTruck                                                             -6.496e+00
## Rated.Horsepower                                                               3.818e-02
## X..of.Cylinders.and.Rotors                                                     5.865e+00
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles) -5.770e+00
## Tested.Transmission.TypeAutomatic                                             -3.771e+00
## Tested.Transmission.TypeContinuously Variable                                  1.415e+01
## Tested.Transmission.TypeManual                                                -9.368e+00
## Tested.Transmission.TypeOther                                                 -9.975e+00
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)  8.109e+00
## Tested.Transmission.TypeSemi-Automatic                                        -2.917e+00
## X..of.Gears                                                                   -1.510e+00
## Drive.System.Description2-Wheel Drive, Rear                                   -6.839e+00
## Drive.System.Description4-Wheel Drive                                          5.472e-01
## Drive.System.DescriptionAll Wheel Drive                                       -3.482e+00
## Drive.System.DescriptionPart-time 4-Wheel Drive                               -1.955e+01
## Equivalent.Test.Weight..lbs..                                                 -6.083e-03
## Axle.Ratio                                                                    -6.750e+00
## N.V.Ratio                                                                      8.521e-01
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                              -1.025e+02
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                       8.073e+00
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                8.875e+01
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                            -1.060e+02
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                            -1.545e+02
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)         -1.634e+02
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                 -6.461e+01
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                -9.775e+01
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)        6.343e+01
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)       -1.681e+02
## THC..g.mi.                                                                     3.750e+02
## RND_ADJ_FE                                                                    -7.089e+00
## Target.Coef.A..lbf.                                                            5.231e-01
## Target.Coef.B..lbf.mph.                                                        3.596e+01
## Target.Coef.C..lbf.mph..2.                                                     1.525e+03
## Set.Coef.A..lbf.                                                              -2.200e-01
## Set.Coef.B..lbf.mph.                                                          -4.480e+00
##                                                                              Std. Error
## (Intercept)                                                                    2.004e+01
## Model.Year.Cat2019                                                             1.010e+00
## Model.Year.Cat2020                                                             1.042e+00
## Model.Year.Cat2021                                                             1.067e+00
## Model.Year.Cat2022                                                             1.085e+00
## Test.Veh.Displacement..L.                                                      7.799e-01
## Vehicle.TypeCar                                                                1.168e+00
## Vehicle.TypeTruck                                                              1.427e+00
## Rated.Horsepower                                                               5.067e-03
```

```
## X..of.Cylinders.and.Rotors                                                          5.905e-01
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles)  2.169e+00
## Tested.Transmission.TypeAutomatic                                                   2.046e+00
## Tested.Transmission.TypeContinuously Variable                                       2.715e+00
## Tested.Transmission.TypeManual                                                      2.224e+00
## Tested.Transmission.TypeOther                                                       2.543e+01
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)    2.847e+00
## Tested.Transmission.TypeSemi-Automatic                                              1.983e+00
## X..of.Gears                                                                         2.804e-01
## Drive.System.Description2-Wheel Drive, Rear                                         1.052e+00
## Drive.System.Description4-Wheel Drive                                               2.154e+00
## Drive.System.DescriptionAll Wheel Drive                                             1.492e+00
## Drive.System.DescriptionPart-time 4-Wheel Drive                                     5.394e+00
## Equivalent.Test.Weight..lbs..                                                       8.527e-04
## Axle.Ratio                                                                          6.115e-01
## N.V.Ratio                                                                           6.779e-02
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                                    2.407e+01
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                            2.368e+01
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                     2.739e+01
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                                  1.959e+01
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                                  1.953e+01
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)               1.953e+01
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                       1.942e+01
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                      1.936e+01
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)             2.186e+01
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)             3.156e+01
## THC..g.mi.                                                                         1.068e+01
## RND_ADJ_FE                                                                         4.068e-02
## Target.Coef.A..lbf.                                                                5.455e-02
## Target.Coef.B..lbf.mph.                                                            1.667e+00
## Target.Coef.C..lbf.mph..2.                                                         8.436e+01
## Set.Coef.A..lbf.                                                                   4.092e-02
## Set.Coef.B..lbf.mph.                                                               1.116e+00
##                                                                                    t value
## (Intercept)                                                                         27.213
## Model.Year.Cat2019                                                                   1.570
## Model.Year.Cat2020                                                                   3.450
## Model.Year.Cat2021                                                                   2.848
## Model.Year.Cat2022                                                                   4.566
## Test.Veh.Displacement..L.                                                            7.714
## Vehicle.TypeCar                                                                      5.321
## Vehicle.TypeTruck                                                                   -4.552
## Rated.Horsepower                                                                     7.535
## X..of.Cylinders.and.Rotors                                                           9.933
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles)  -2.660
## Tested.Transmission.TypeAutomatic                                                   -1.844
## Tested.Transmission.TypeContinuously Variable                                        5.212
## Tested.Transmission.TypeManual                                                      -4.212
## Tested.Transmission.TypeOther                                                       -0.392
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)     2.848
## Tested.Transmission.TypeSemi-Automatic                                              -1.471
## X..of.Gears                                                                         -5.386
## Drive.System.Description2-Wheel Drive, Rear                                         -6.500
## Drive.System.Description4-Wheel Drive                                                0.254
## Drive.System.DescriptionAll Wheel Drive                                             -2.334
## Drive.System.DescriptionPart-time 4-Wheel Drive                                     -3.623
## Equivalent.Test.Weight..lbs..                                                       -7.134
## Axle.Ratio                                                                         -11.039
## N.V.Ratio                                                                           12.569
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                                    -4.257
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                             0.341
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                      3.240
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                                  -5.413
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                                  -7.912
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)               -8.366
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                       -3.327
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                      -5.048
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)              2.901
```

```
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)                    -5.327
## THC..g.mi.                                                                                 35.123
## RND_ADJ_FE                                                                               -174.250
## Target.Coef.A..lbf.                                                                         9.589
## Target.Coef.B..lbf.mph.                                                                    21.565
## Target.Coef.C..lbf.mph..2.                                                                 18.071
## Set.Coef.A..lbf.                                                                           -5.377
## Set.Coef.B..lbf.mph.                                                                       -4.014
##                                                                                          Pr(>|t|)
## (Intercept)                                                                               < 2e-16
## Model.Year.Cat2019                                                                       0.116502
## Model.Year.Cat2020                                                                       0.000562
## Model.Year.Cat2021                                                                       0.004406
## Model.Year.Cat2022                                                                       5.00e-06
## Test.Veh.Displacement..L.                                                                1.28e-14
## Vehicle.TypeCar                                                                          1.05e-07
## Vehicle.TypeTruck                                                                        5.35e-06
## Rated.Horsepower                                                                         5.13e-14
## X..of.Cylinders.and.Rotors                                                                < 2e-16
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles) 0.007813
## Tested.Transmission.TypeAutomatic                                                        0.065256
## Tested.Transmission.TypeContinuously Variable                                            1.89e-07
## Tested.Transmission.TypeManual                                                           2.54e-05
## Tested.Transmission.TypeOther                                                            0.694906
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)         0.004406
## Tested.Transmission.TypeSemi-Automatic                                                   0.141392
## X..of.Gears                                                                              7.29e-08
## Drive.System.Description2-Wheel Drive, Rear                                              8.25e-11
## Drive.System.Description4-Wheel Drive                                                    0.799431
## Drive.System.DescriptionAll Wheel Drive                                                  0.019608
## Drive.System.DescriptionPart-time 4-Wheel Drive                                          0.000292
## Equivalent.Test.Weight..lbs..                                                            1.02e-12
## Axle.Ratio                                                                                < 2e-16
## N.V.Ratio                                                                                 < 2e-16
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                                         2.09e-05
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur                                 0.733168
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)                          0.001198
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                                       6.28e-08
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                                       2.69e-15
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)                     < 2e-16
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                            0.000881
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                                           4.52e-07
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)                  0.003720
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)                  1.01e-07
## THC..g.mi.                                                                                < 2e-16
## RND_ADJ_FE                                                                                < 2e-16
## Target.Coef.A..lbf.                                                                       < 2e-16
## Target.Coef.B..lbf.mph.                                                                   < 2e-16
## Target.Coef.C..lbf.mph..2.                                                                < 2e-16
## Set.Coef.A..lbf.                                                                         7.66e-08
## Set.Coef.B..lbf.mph.                                                                     6.00e-05
##
## (Intercept)                                                                              ***
## Model.Year.Cat2019
## Model.Year.Cat2020                                                                       ***
## Model.Year.Cat2021                                                                       **
## Model.Year.Cat2022                                                                       ***
## Test.Veh.Displacement..L.                                                                ***
## Vehicle.TypeCar                                                                          ***
## Vehicle.TypeTruck                                                                        ***
## Rated.Horsepower                                                                         ***
## X..of.Cylinders.and.Rotors                                                               ***
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated Manual with paddles) **
## Tested.Transmission.TypeAutomatic                                                        .
## Tested.Transmission.TypeContinuously Variable                                            ***
## Tested.Transmission.TypeManual                                                           ***
## Tested.Transmission.TypeOther
## Tested.Transmission.TypeSelectable Continuously Variable (e.g. CVT with paddles)         **
```

```
## Tested.Transmission.TypeSemi-Automatic
## X..of.Gears                                                                  ***
## Drive.System.Description2-Wheel Drive, Rear                                  ***
## Drive.System.Description4-Wheel Drive
## Drive.System.DescriptionAll Wheel Drive                                      *
## Drive.System.DescriptionPart-time 4-Wheel Drive                              ***
## Equivalent.Test.Weight..lbs..                                                ***
## Axle.Ratio                                                                   ***
## N.V.Ratio                                                                    ***
## Test.Fuel.Type.DescriptionCARB Phase II Gasoline                             ***
## Test.Fuel.Type.DescriptionCold CO Diesel 7-15 ppm Sulfur
## Test.Fuel.Type.DescriptionCold CO E10 Premium Gasoline (Tier 3)              **
## Test.Fuel.Type.DescriptionCold CO Premium (Tier 2)                           ***
## Test.Fuel.Type.DescriptionCold CO Regular (Tier 2)                           ***
## Test.Fuel.Type.DescriptionE85 (85% Ethanol 15% EPA Unleaded Gasoline)        ***
## Test.Fuel.Type.DescriptionFederal Cert Diesel 7-15 PPM Sulfur                ***
## Test.Fuel.Type.DescriptionTier 2 Cert Gasoline                               ***
## Test.Fuel.Type.DescriptionTier 3 E10 Premium Gasoline (9 RVP @Low Alt.)      **
## Test.Fuel.Type.DescriptionTier 3 E10 Regular Gasoline (9 RVP @Low Alt.)      ***
## THC..g.mi.                                                                   ***
## RND_ADJ_FE                                                                   ***
## Target.Coef.A..lbf.                                                          ***
## Target.Coef.B..lbf.mph.                                                      ***
## Target.Coef.C..lbf.mph..2.                                                   ***
## Set.Coef.A..lbf.                                                             ***
## Set.Coef.B..lbf.mph.                                                         ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.71 on 17196 degrees of freedom
## Multiple R-squared:  0.8538, Adjusted R-squared:  0.8534
## F-statistic:  2449 on 41 and 17196 DF,  p-value: < 2.2e-16
```

Looking again at the summary of Emissions Model 3, we can see that the top 3 predictors that hold the most weight in predicting a car's CO_2 based on respective t-values are:

1. Miles per Gallon (RND_ADJ_FE): -174.25

As the number of miles per gallon a car is able to achieve increases, its CO_2 emissions go down.

2. Total hydrocarbon emissions (THC..g.mi.): 35.12
3. Electric Dynamometer Coefficient/mph (Target.Coef.B..lbf.mph.): 21.90

(This is the measure of force, speed, and power required to operate the car being measured)

As the total hydrocarbon emissions and the electric dynamometer coefficient increase, the CO_2 emissions produced by a car will also increase.