ANLY-511 Final Project Report

ANLY-511-04 Group 3

12/12/2022

DATA CLEANING CODE START

```
## Warning: package 'ISLR2' was built under R version 4.0.5
```

DATA CLEANING CODE END

Analyzing Relationships Between Car Design and Fuel Emissions

Authors

- Mia Mayerhofer
- Matt Moriarty
- Natalie Smith
- Madelyne Ventura
- Linlin Wang

I. 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.

ADD FIGURE 1 TO MARKDOWN

II. Data

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 XY below.

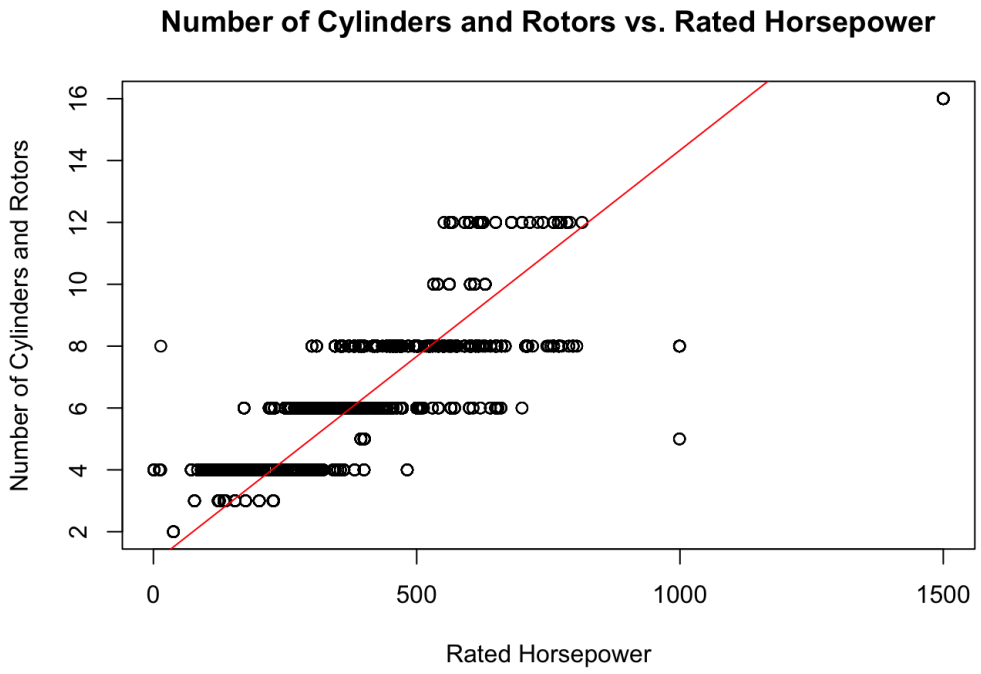


Figure XY

T-TESTS CODE START

T-TESTS CODE END

III. Exploratory Data Analysis (EDA)

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.

ADD FIGURES 2 and 3 TO MARKDOWN

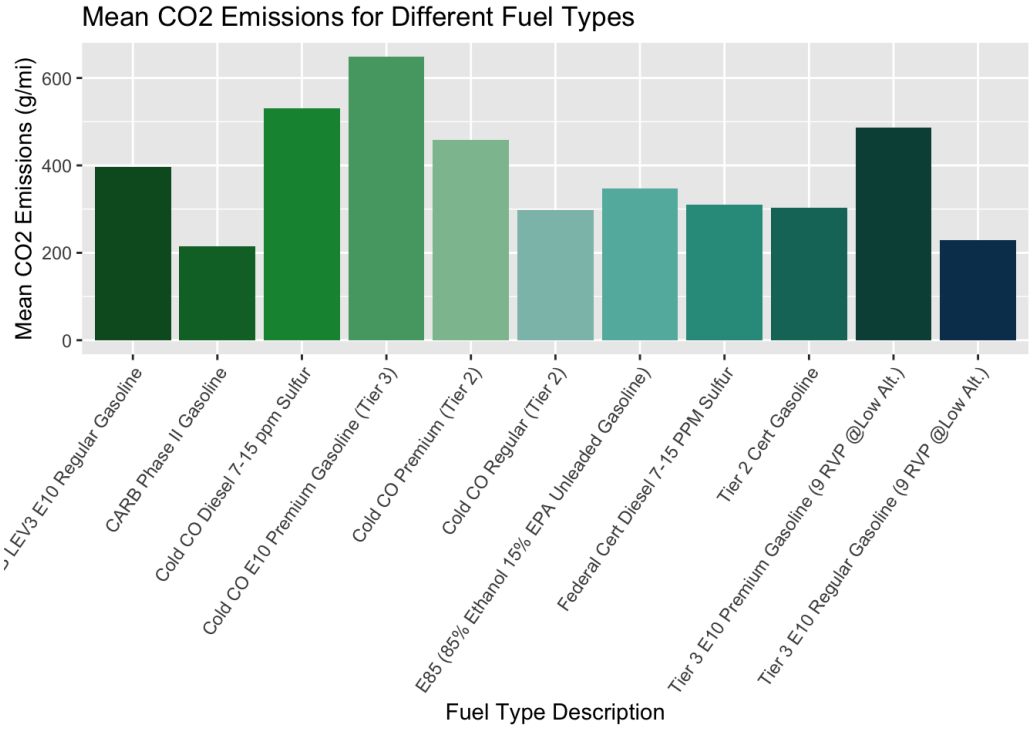


Figure 2

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

In Figure 2 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 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 4).

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 4

ADD FIGURE 4 TO MARKDOWN

In the boxplots shown in Figure 5, 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.

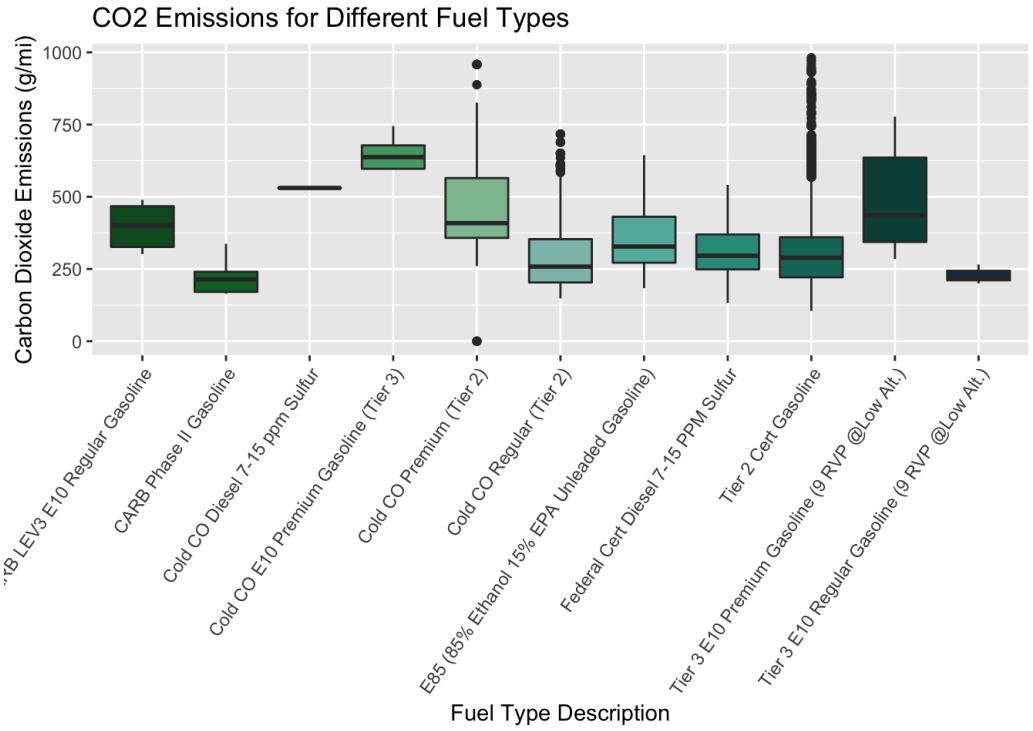


Figure 5

ADD FIGURE 5 TO MARKDOWN

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 6 look closer at the distributions for these two most common fuel types: federal certified diesel 7-15 PPM sulfur and tier 2 certified gasoline.

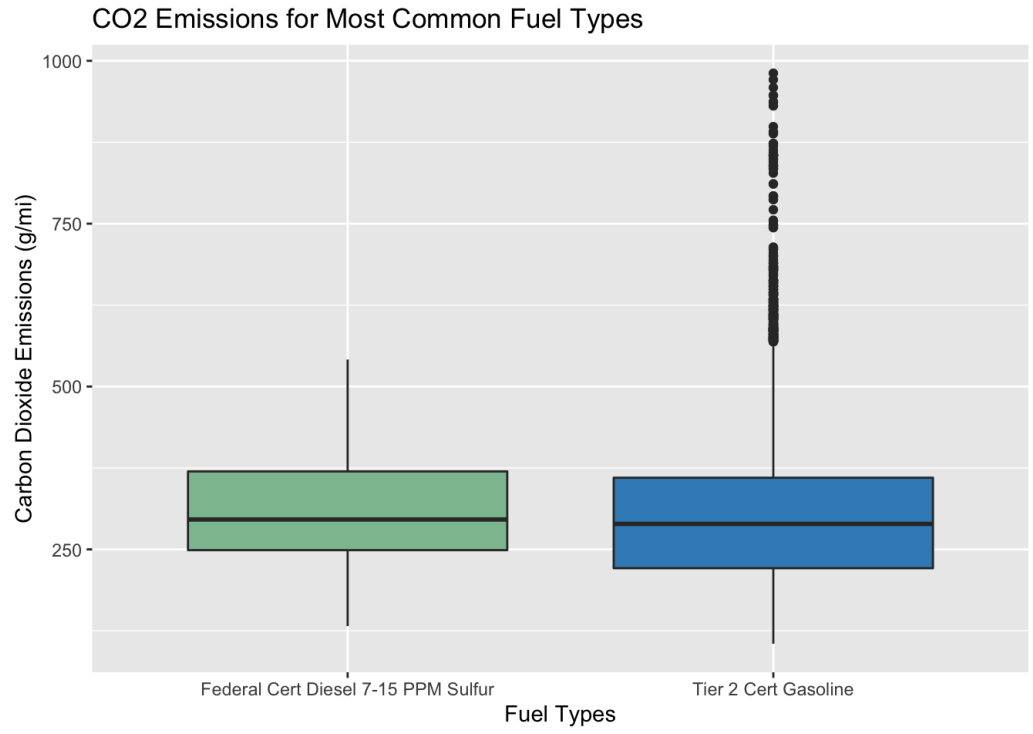


Figure 6

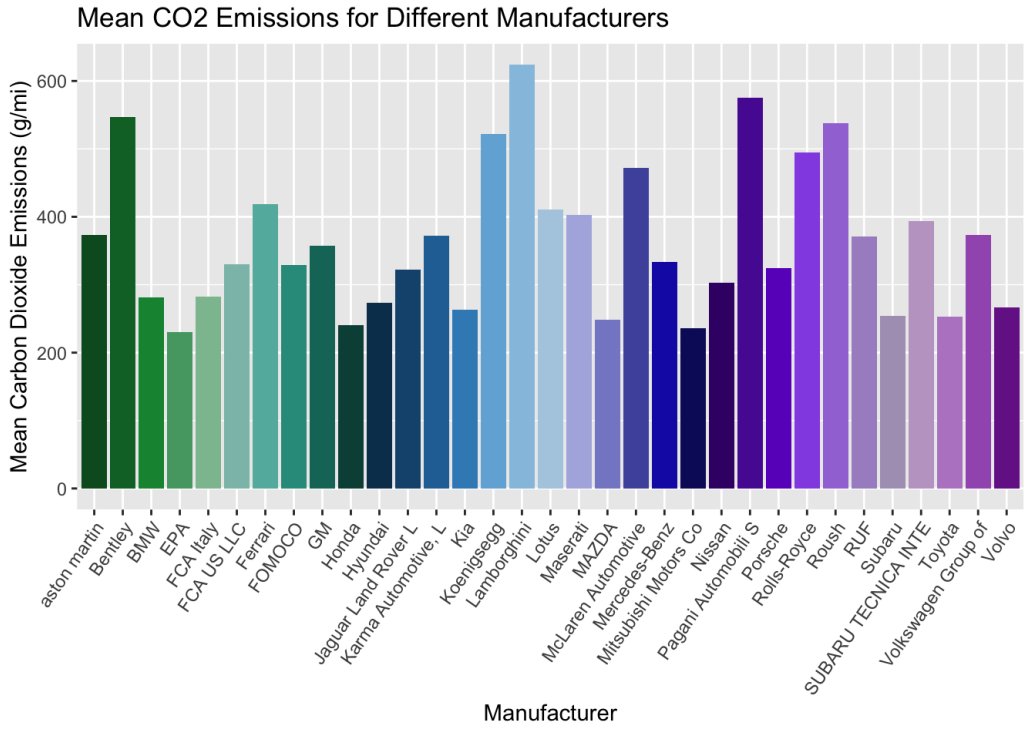


Figure 7

ADD FIGURES 6 and 7 TO MARKDOWN

The bar plot shown in Figure 7 highlights which vehicle manufacturers have the highest and lowest mean carbon dioxide emissions. Figures 8 and 9 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 8 *(left) and* ***Figure 9*** *(right)*

ADD FIGURES 8 and 9 TO MARKDOWN

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 10 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.

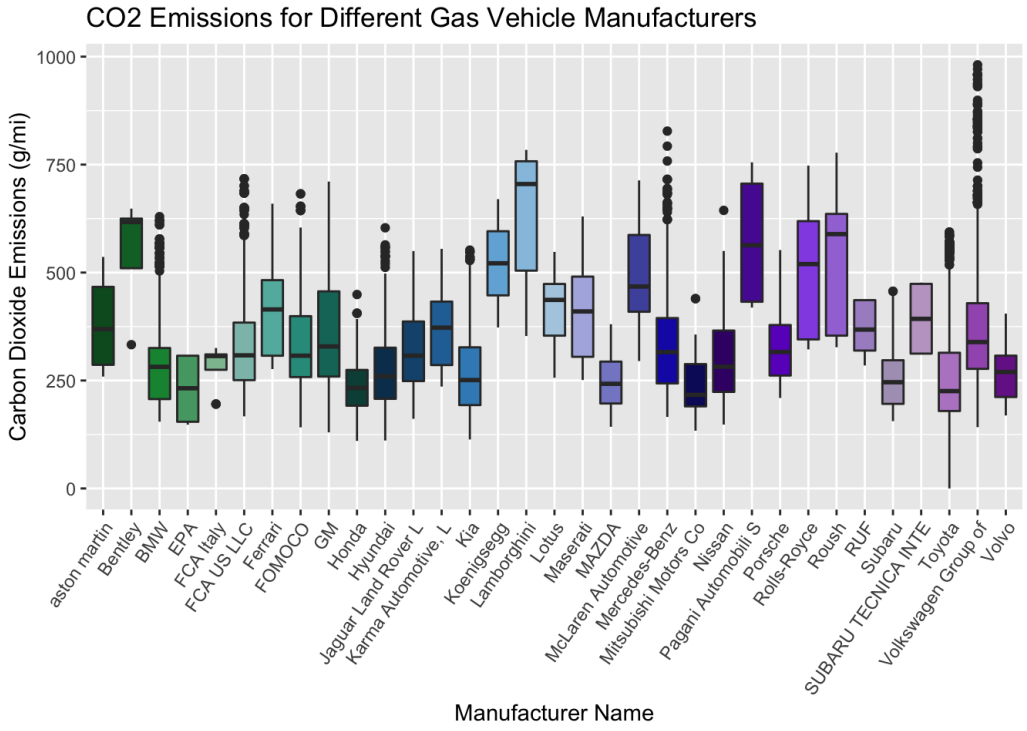


Figure 10

ADD FIGURE 10 TO MARKDOWN

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 11 below look closer at the distributions for these two most common manufacturers: General Motors and Toyota.

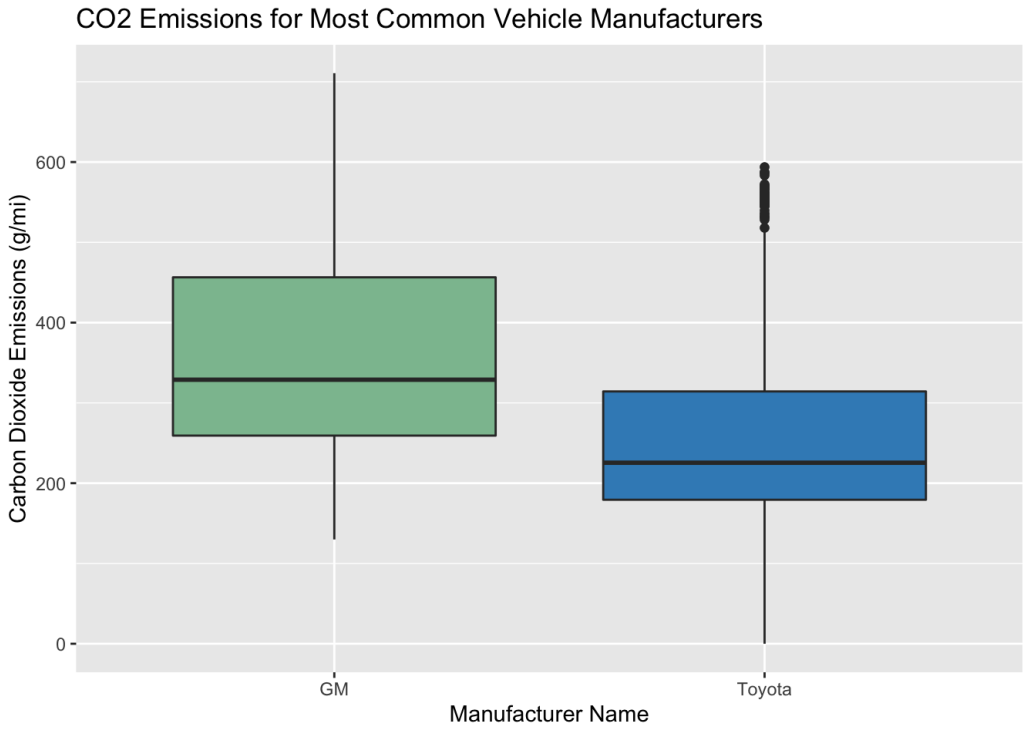


Figure 11

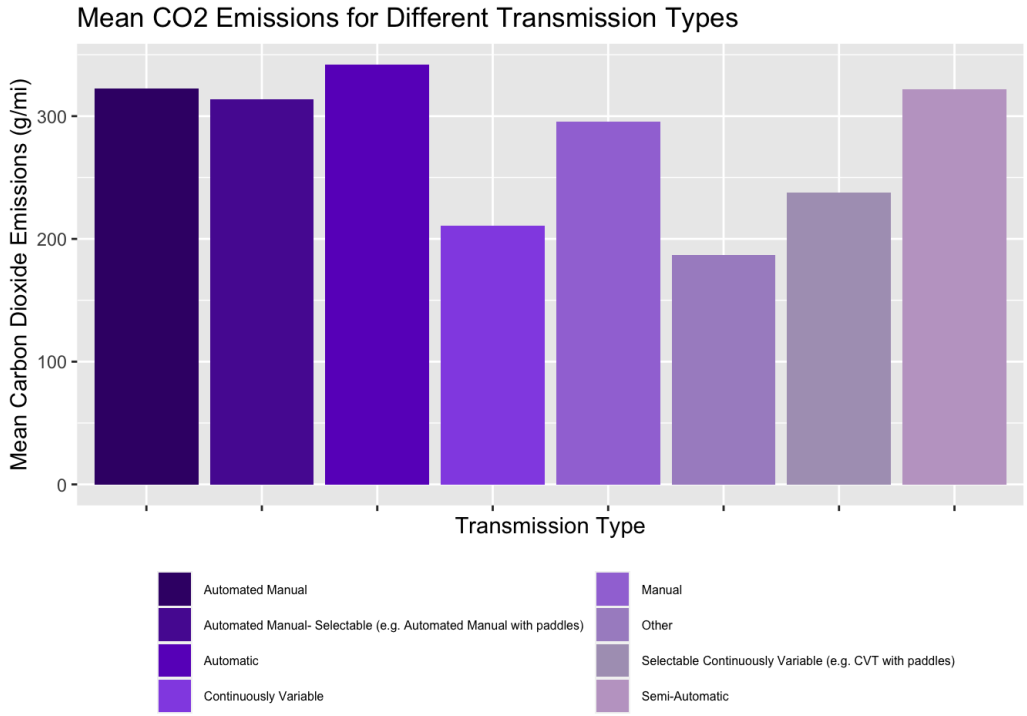


Figure 12

ADD FIGURES 11 and 12 TO MARKDOWN

Figure 12 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 13 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
Selectable Continuously Variable (e.g. CVT with paddles)	237.6764
Continuously Variable	210.7348
Other	187.1900

Figure 13

ADD FIGURE 13 TO MARKDOWN

In the boxplots shown in Figure 14, 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 15 below look closer at the distributions for these two transmission types.

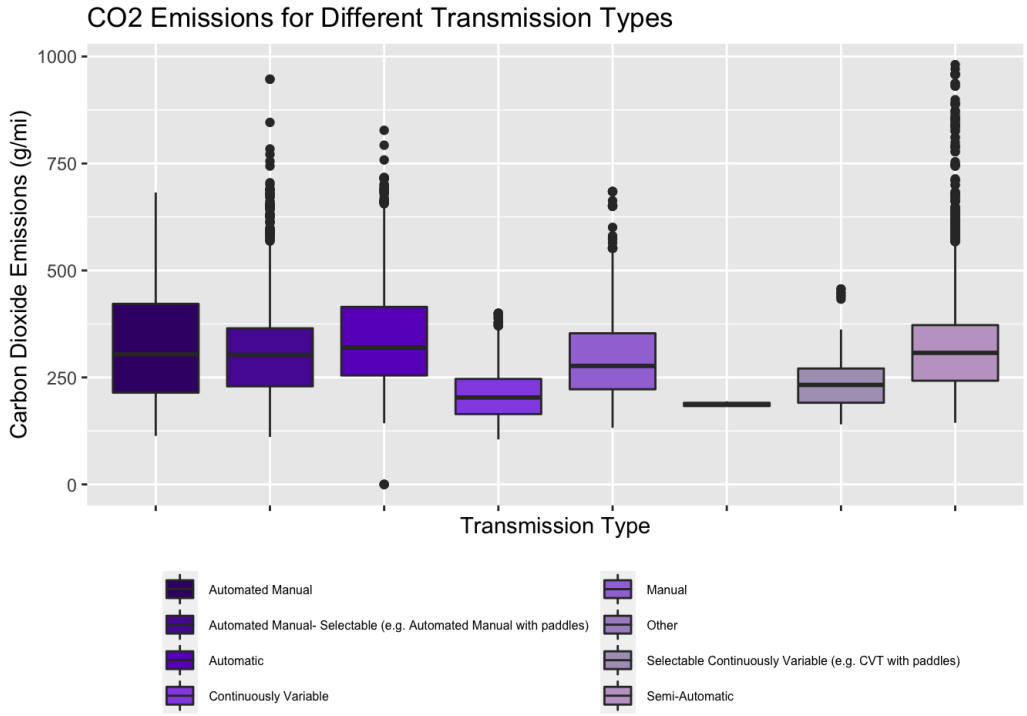
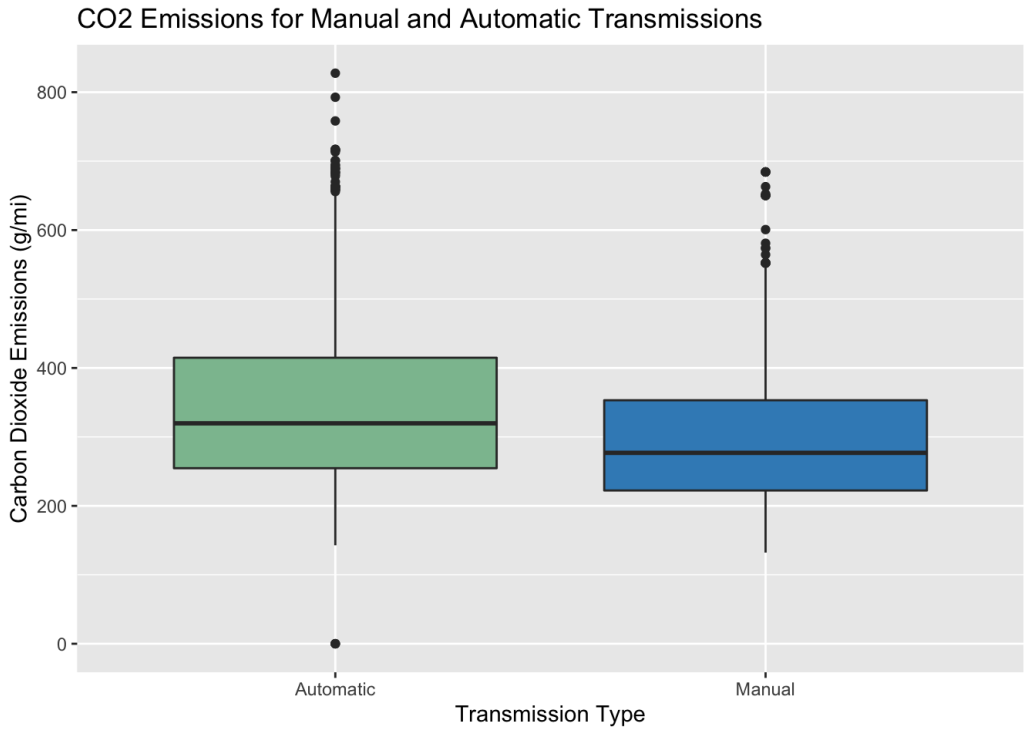


Figure 14



ADD FIGURES 14 and 15 TO MARKDOWN

IV. 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

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. From the below result table, 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 1

ADD TABLE 1 TO MARKDOWN

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 manufactures?*

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 manufactures.*
- *H_A: There is a significant difference in CO_2, CO, and THC between the different vehicle manufactures.*

Performing Pillai's Trace, Hotelling-Lawley, Wilks, and Roy separately give the following results. 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 2

ADD TABLE 2 TO MARKDOWN

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.

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 3

ADD TABLE 3 TO MARKDOWN

Chi-Squared Test of Independence

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. Figure 16 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 / 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

ADD FIGURE 16 TO MARKDOWN

- *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 Figure 17. 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 / Fuel Emission Level	Low	Medium	High
Two-Wheel Drive, Front	5332	3820	7

Drive System / Fuel Emission Level	Low	Medium	High
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 5

ADD FIGURE 17 TO MARKDOWN

- *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 Figure 18. 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.

Drive System / Fuel Emission Level	Low	Medium	High
Automated Manual	255	430	79
Automated Manual-Selectable	553	1115	151
Automatic	1411	4081	603
Continuously Variable	2168	671	0
Manual	634	932	72
Other	4	0	0
Selectable Continuously Variable	454	308	0
Semi-Automatic	2179	5122	516

Table 6

ADD FIGURE 18 TO MARKDOWN

T-Tests

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.

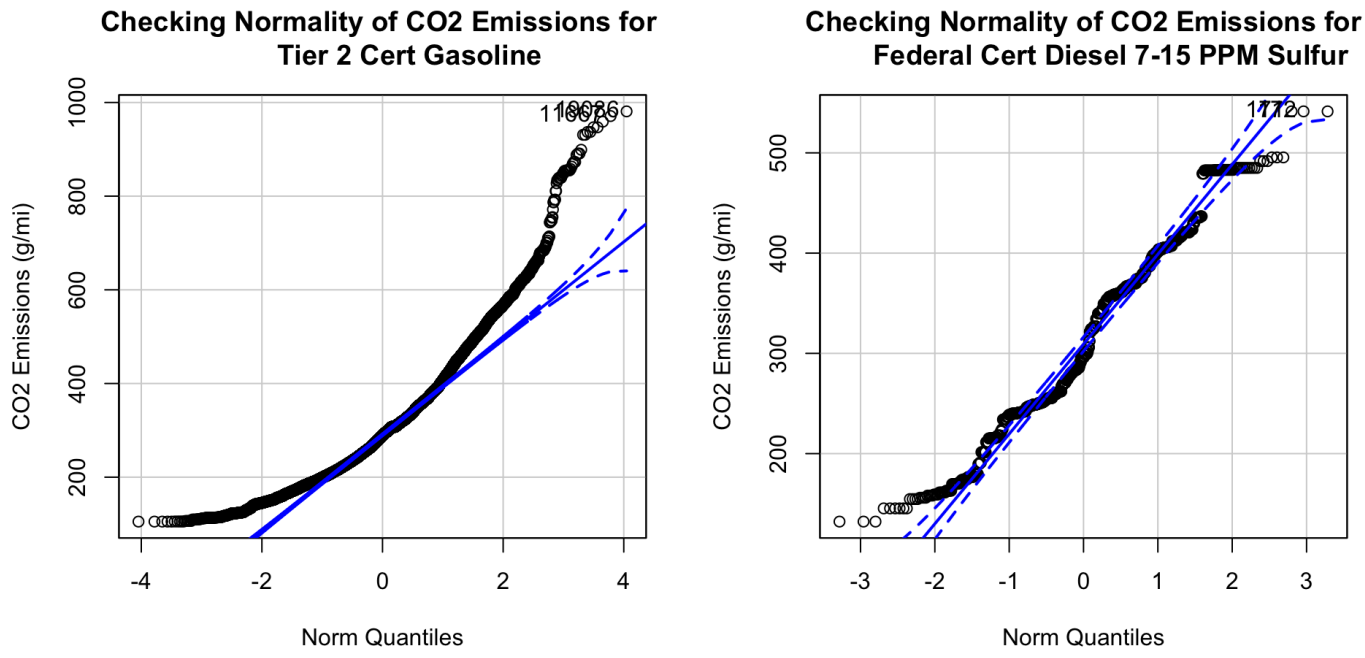


Figure 19

ADD FIGURE 19 TO MARKDOWN

It is clear from the two QQ-plots 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 11, 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.

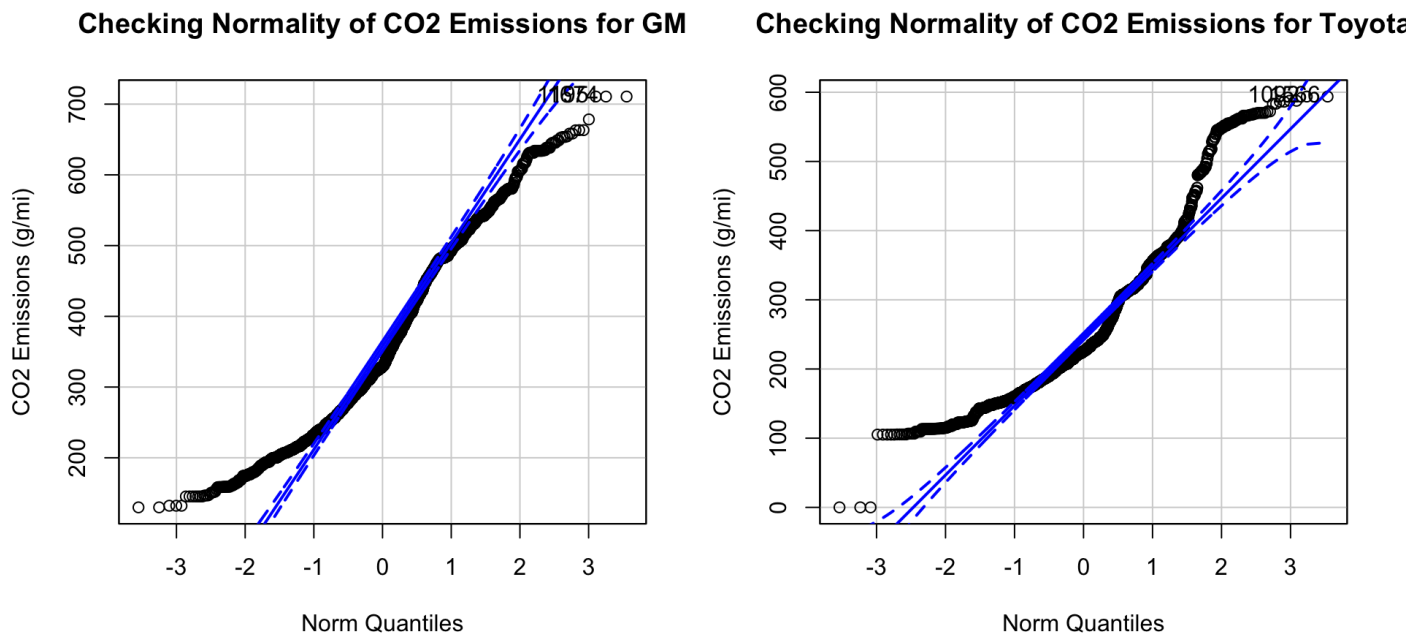


Figure 20

ADD FIGURE 20 TO MARKDOWN

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.

- *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 15, 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 the QQ-plots; 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.

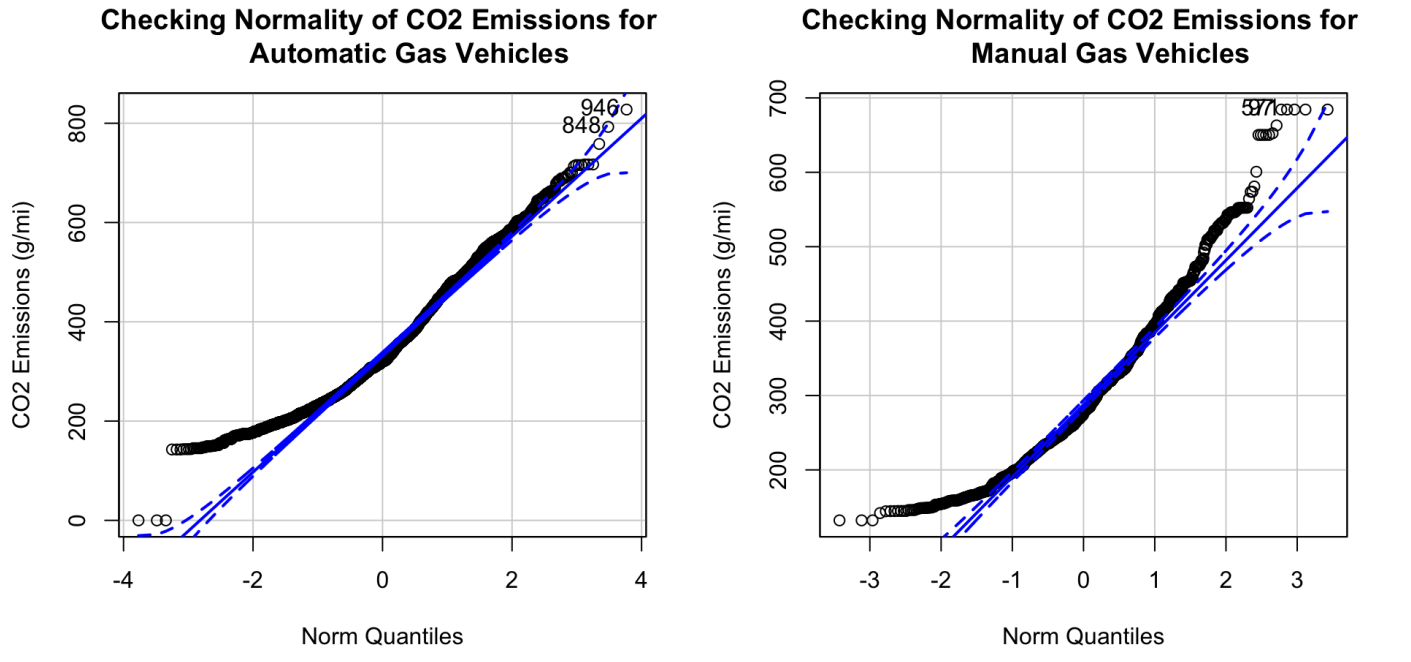


Figure 21

ADD FIGURE 21 TO MARKDOWN

Linear Regression

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 YY:

```
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 22

ADD FIGURE YY TO MARKDOWN

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 YY:

```
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 23

ADD FIGURE YY TO MARKDOWN

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 YY:

```
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 24

ADD FIGURE YY TO MARKDOWN

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.

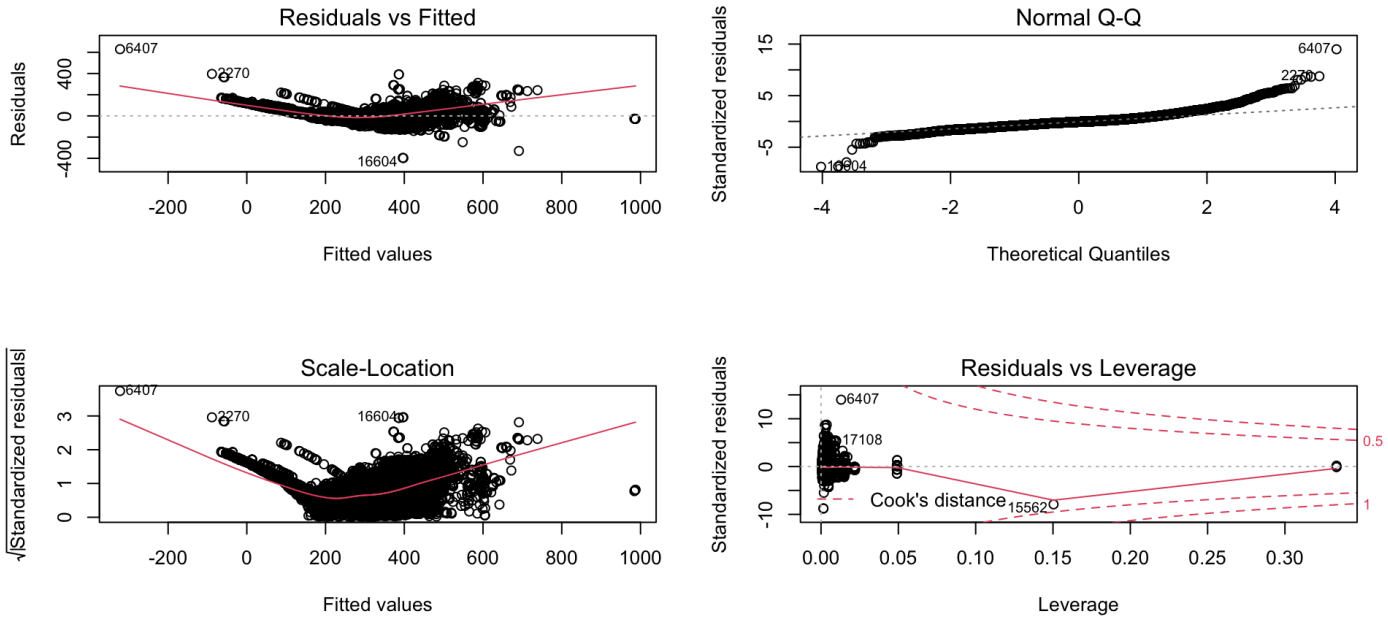


Figure 25

ADD FIGURE YY TO MARKDOWN

Looking at Figure YY, 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 YY:

```
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 25

ADD FIGURE YY TO MARKDOWN

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.

Results

ADD RESULTS TO MARKDOWN

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