The Relationship Between Carbon Dioxide Emissions and Petroleum Consumption

Graham Ward

Author Note

This study was requested and paid for by the automaker, Nissan America. A subsidiary of parent company Nissan, headquartered in Yokohama Japan. Analysis and report were carried out during the first half of Q1 2024.

Abstract

This study was conducted to examine the question of whether there is a statistically significant relationship between oil consumption and the amount of carbon dioxide emissions for vehicles in the United States. The null hypothesis is there is no statistically significant relationship between the two variables. The study controls for multiple vehicle characteristics and examines the role each play in contributing to carbon dioxide emissions. In this study, the null hypothesis was rejected in favor of the null hypothesis at a 95% confidence level, indicating a statistically significant relationship between the variables. This study leveraged a secondary data set which was collected from *FuelEconomy.gov* and consists of 43,177 observations of vehicles produced between 1984 and 2020 within the United States from 138 different manufacturers. A multiple variable regression model was developed to depict the relationship between various vehicle characteristics to gain insight into the creation of fuel-efficient vehicles and identify business opportunities allowing opportunities for Nissan to gain a competitive advantage in different markets by identifying areas where we can develop fuel efficient technology. While we were able to reject the null hypothesis in favor of the alternate, the presence of multicollinearity was still present at the end of the study, necessitating further analysis to apply further statistical inference in order to develop predictions. Additionally, this study's final model did not account for any nominal variables and as a result is limited in terms of its robust interpretability.

The Relationship Between Carbon Dioxide Emissions and Petroleum Consumption

Background

The Environmental Protection Agency (EPA) estimated during the year 2021 the Transportation sector accounted for 28% of all Greenhouse gas (GHG) emissions. (EPA, 2023) This makes the Transportation sector the largest contributor of GHG emissions. As a part of the transportation sector, the automotive industry is vital to reducing the amount of carbon dioxide (CO2) emissions. Fuel efficient vehicles that minimize CO2 emissions are necessary to help meet the global climate goals as set by the 2015 Paris Accords. At Nissan, we live by the motto “Innovation that excites.” Therefore, understanding the relationship between CO2 emissions and petroleum consumption, as well as other vehicle characteristics, will give us a competitive advantage within the automotive industry by designing industry leading fuel-efficient vehicles across all our consumer bases. To accomplish this, it is necessary to develop a working model capturing the various factors which contribute to the CO2 emissions of a vehicle.

Purpose

The objective of this analysis is to statistically investigate the association of primary fuel tailpipe CO2 emissions in grams per mile to annual primary-fuel petroleum consumption in barrels after controlling for combined miles-per-gallon for the primary fuel type, vehicle manufacturer, make, engine displacement, engine cylinders, combined luggage and passenger volume in cubic feet, vehicle type, transmission type, and primary fuel type. In this study, the null hypothesis that annual oil consumption in barrels has no statistically significant effect on CO2 emissions. Nissan is currently exploring different ways to capitalize market share on the creation of fuel-efficient vehicle market. By identifying trends and building a thorough understanding of the current and past fuel-efficient vehicle landscape, Nissan can recognize untapped markets and invest in building out areas where fuel efficient vehicles lack, allowing space for innovation.

Generalized Object Formula

Equation 1 represents the generalized linear object formula specifying the dependent variable of CO2 emissions in grams per mile as a product of a constant term, then the independent variables' petroleum consumption, the number of miles per gallon, the vehicle's manufacturer, and an error term that represents other independent variables which are not present in equation 1. The equation provides a basic framework for us to examine the relationship between CO2 and petroleum consumption. From this point one can add additional variables to determine the most innovative and fuel-efficient vehicles to design.

$$CO_2 = \beta_0 + \beta_1 PetroleumConsumption + \beta_2 MPG + \beta_3 Manufacturer + \varepsilon \quad (1)$$

Method

Data Collection and Data Pre-Processing

The data underlying this study is from the United States Department of Energy and contains fuel economy information on vehicles produced from 1980 through 2020. (U.S. Department of Energy, 2020). The data was collected on a Lenovo ThinkStation P330 tower with an Intel Xeon E-2186G 6-Core Processor with the following specifications: 64 GB (16GBx4) DDR 2666MHz ECC UDIMM, RAID 1, 1.0 Terrabyte M.2 PCIe Opal SSD (x2), and running Windows 10 Pro for Workstations 64. The software used to manipulate the data was MatLab R2020b, Microsoft Excel for Microsoft Office 365 (64-bit), SAS Enterprise Guide 7.1 (64-bit), and Tableau 2020.3. After downloading the data from FuelEconomy.gov, a CSV was opened in Microsoft Excel and converted into a table. That final Excel table was then imported into MatLab for distribution curve fitting and imported into SAS for further statistical analysis.

Population Characteristics

The full dataset contains 43,177 observations of vehicles manufactured during the period 1984-2020 (U.S. Department of Energy, 2020). This sample contains vehicles from 138 different vehicle manufacturers broken down by the vehicle's various defining characteristics. For the purposes of this

study, we control for the following characteristics and their associated variables: combined miles per gallon, vehicle manufacturer, engine displacement, engine cylinders, combined luggage and passenger volume, vehicle type, transmission type, and primary fuel type. There are five categorical variables of interest: vehicle type (vehtype), emissions category (emissionscat), vehicle manufacturer (make_id), transmission type (transtype_id), and primary fuel type (prifueltype). The discrete variables include the number of vehicle cylinders (cylinders), miles per gallon (Comb08) and volume (volume). Continuous variables include CO2 emissions in grams per mile (CO2TailpipeGPM), annual petroleum consumption (barrels08) and engine displacement (displ).

Descriptive Statistics

Examining Table 1, The Table of Descriptive Statistics we note there are 43,177 observations across most of the variables. Data is missing for 258 cylinders observations and 256 displ observations. Examining the breakdown of CO2 emissions there is a mean of 462.7668 grams per mile and a standard deviation of 124.7720 grams per mile. Annual petroleum consumption measured in barrels per year has a mean value of 17.1532 barrels per year and a standard deviation of 4.6629. The variable comb08, which measures the vehicle's average miles per gallon had a mean of 20.8459 miles per gallon and a standard deviation of 8.2242 miles per gallon. As mentioned previously, cylinders, a discrete variable, and since one cannot have a partial cylinder, the mode of this variable is 4.0000 indicating the most common engine type in this dataset is a 4-cylinder engine. Engine displacement, a continuous variable has a mean of 3.2867 liters and a standard deviation of 1.3570 liters. The final continuous variable is volume with a mean of 66.9290 cubic feet and a standard deviation of 69.0417 cubic feet. The other variables are categorical and while they have values for the descriptive statistics table, there is no significant statistical interpretation for these values. They were encoded numerically, meaning their current form has no interpretable value because we introduced order in variables that don't have order by nature.

Relationship of CO2 Emissions to Annual Oil Consumption

Figure 1 is a scatter plot visualizing the relationship between the two quantitative continuous variables co2TailpipeGPM and Barrels08. This graph then is a visualization of the primary relationship we are interested in. The scatter plot highlights the positive linear relationship between the amount of CO2 emissions and the barrels of fuel the vehicle uses annually. This visualization supports research done by Mickūnaitis et al. on the idea that the more petroleum consumption a vehicle has annually the higher the CO2 emissions in grams per mile (2007).

Bivariate Frequency

Table 2 is a Bivariate Frequency Table exploring the proportions of primary fuel types broken down by the different vehicle types and transmission types. Both Vehicle Type and Transmission Type were found to be statistically significant across the primary fuel type categories. Relative to the population, there were proportionately more Unknown vehicles that consumed Regular Gasoline (53.4%), Diesel (57.3%), and Natural Gas (56.7%). Additionally, when compared against the population, Electric vehicles were proportionately higher in the hatchback vehicle type (40.9%), indicating a majority of the electric vehicles were hatchback. Premium Gasoline held proportionately more Passenger 2-Door (24.7%) and Passenger 4-Door vehicles (37.8%) in its sample relative to the population. Of note regarding transmission type, when it comes to manual transmission vehicles there are zero records for vehicles that consume Midgrade gasoline, Natural Gas, or Electric vehicles. Proportionate to the population, automatic transmission vehicles that consume Premium Gasoline, Midgrade Gasoline, Natural Gasoline and Electric vehicle types have a higher proportion of their sample being automatic transmission vehicles.

Table of Associations

Table 3 is a proportional analysis of the variables Primary Fuel Type, Vehicle Type and Transmission Type broken down across six different emissions categories. Of these variables all were

found to be statistically significant (p<.0001) in determining the emissions category of vehicles. The emissions categories are Ultra-Low Emission, Very-Low Emission, Low Emission, Standard, Polluter and Gross Polluter. The proportions of each category are compared to the population breakdown for an understanding of how the categorical breakdowns compare to the overall population. Regular Gasoline, the most common type, according to the population with 28,733 (66.5%) was proportionately higher in the Very-Low Emission (81.0%), Low Emission (73.2%), Polluter (73.9%) and Gross Polluter (67.0%). Unknown vehicle types were found to be proportionately more prevalent Polluters and Gross Polluters. This aligns with how the data is structured as Unknown vehicle types contain our all our trucks and performance automotives. If we are ready to produce more fuel-efficient vehicles, we may wish to consider looking at producing Hatchback and Passenger 4-Door vehicle types. Across the categories of Ultra-Low, Very-Low and Low Emissions these vehicle types are proportionately much more emission friendly than our Unknown vehicle types.

Results

Probability Density Function

Figure 2 depicts the probability density function for CO2 emissions modeled in equation 2, overlaid on a histogram which was fitted to a normal distribution by creating 52 equal sized bins, which allowed us to fit the histogram to the normalized probability density function in equation 2. The bins have a width of 24.4148 grams per mile of CO2 emissions. The shape of the histogram closely resembles the normalized distribution of equation 2 which is the red line in Figure 2. The peak is located right at the population mean (μ) of 465.538. By fitting our data to a normal distribution this will allow the performing of statistical tests and the development of confidence intervals, allowing us to accurately draw conclusions about the population of vehicles.

$$PDF_{CO2} = f(x;\ \mu = 465.538, \sigma = 119.88) = \begin{cases} \frac{1}{[(\sqrt{2\pi})(119.88)]} e^{-\left[\frac{(x-465.538)^2}{[(2)(119.88)]}\right]^2}, x \geq 0 \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad , x < 0 \end{cases} \quad (2)$$

Correlation

The most highly correlated variables as reported in Table 4 are the amount of CO2 emissions and the amount of petroleum consumed annually. This relationship has a Pearson Correlation Coefficient equal to .9885 which indicates a strong linear relationship and that cars consuming more Petroleum produce more CO2 emissions in the atmosphere. Emissions were strongly positively correlated with engine displacement (.7954), engine cylinders (.7438) and the emissions category (0.8894). This indicates larger sized vehicles produce more CO2 emissions. Additionally, engine size and cylinders were also found to be highly correlated with a correlation coefficient equal to .9046. Conversely, the most negatively correlated variables are CO2 emissions and the vehicle's miles per gallon having a correlation value of -.9184. This matches the narrative in which more fuel-efficient vehicles produce less CO2 emissions. This is further supported through miles per gallon also being negatively correlated with petroleum consumption (-.9050) as well as emissions category variable (-.8415). The more miles per gallon a vehicle gets, the more efficient it is at burning fuel and as a result more fuel-efficient vehicles produce less CO2 emissions.

The manufacturer and primary fuel type were found to not be correlated with any variables. This indicates the manufacturer and primary fuel type does not have a strong relationship with the other variables. Another pair of variables that did not share strong correlation were the number of engine cylinders and the vehicle type suggesting different vehicles ran various engine cylinders in their vehicles and that one specific number of engine cylinders was not guaranteed present in a specific vehicle type. It should be noted that determining correlation with the nominal variables is inherently difficult in their current state.

Chi-Squared Discussion

Table 5 is a Two-Way Contingency table showing the relationship between the two categorical variables vehicle type and emissions category. The table contains the observed values, the expected values, and the Chi-Squared contributions from each possible combination of vehicle and emission type. When conducting our test of homogeneity and examining Table 5, the results of the Chi-Square Test of Homogeneity revealed the categorical variables had a Chi-Square value of 9,406.17. This value was too far removed from the 15 degrees of freedom to determine if emissions and vehicle type are homogenous variables, and additional research is needed to be able to draw any insights from this relationship.

Multicollinearity

Multicollinearity can be present when two or more independent variables explain similar changes in the dependent variable. We can identify multicollinearity through multiple ways. The first is to examine the Pearson Correlation Coefficient values for any value greater than .9 (Devore, 2016). Independent variables suspected of having multicollinearity will reach this threshold. As shown in Table 4, the variables cylinders and displ have a correlation coefficient equal to .9046, indicating the potential presence of multicollinearity. However, correlation does not always mean multicollinearity and we must therefore turn our attention to different indicators such as the Variance Inflation Factor (VIF) when we run the regression through statistical software. For this test, a value between 5 and 10 indicates the variable may have multicollinearity with one of the other independent variables (Bhandari, 2020).

Initial Linear Regression Models

From the generalized linear object Formula represented in equation 1, an initial multiple regression model was constructed as depicted in equation 3 and includes the variables controlled for in this study. Figure 4 is the output of the model after running it through statistical software. Key characteristics of this model is the high adjusted R-Squared of .9828 which tells us the model is

explaining 98.28% of the change in the dependent variable CO2 emissions, after accounting for the number of parameters. Additionally, the high F-value of 245,913 is indicative of the model's predictive power and helps in rejecting the null hypothesis in combination with our p-values (Glen, 2022). All variables were found to be statistically significant (p<.0001) except for make_id (.1093). In the model the β_0 had a value of 79.5532 which gives us the baseline CO2 emissions in grams per mile. The primary variable of interest, barrels08 had a coefficient equal to 20.7438, indicating a one barrel in annual petroleum consumption is associated with an increase of 20.7438 CO2 grams per mile while holding all other variables constant.

$$\begin{aligned} CO2tailpipeGPM &= \beta_0 + \beta_1(barrels08) + \beta_2(comb08) + \beta_3(make\_id) + \beta_4(displ) \\ &+ \beta_5(cylinders) + \beta_6(volume) + \beta_7(vehtype) + \beta_8(emissionscat) \\ &+ \beta_9(transtype\_id) + \beta_{10}(prifueltype) + \varepsilon \end{aligned} \quad (3)$$

The VIF of the variables in Figure 3, show there are four variables with VIF's greater than 5. These variables are barrels08 (9.3468), comb08 (6.0801), displ (7.2995), and cylinders (6.4795). This indicates these independent variables may suffer from multicollinearity, especially since the pairs barrels08 and comb08 were highly correlated as well as displ and cylinders.

Final Linear Regression Models

The final regression model can be found in equation 4. In the final model, the categorical variables were removed because of the need for creation of indicator variables to interpret the results of the regression model accurately. Additionally, the final regression model has an adjusted R-Squared value of .9812, indicating this model predicts 98.12% of the change in the dependent variable. This is slightly less than our original model with the categorical variables. This model still suffers from multicollinearity problems due to four of the five variables having VIF's greater than 5. However, all variables were found to be statistically significant at the 95% confidence level (p<.0001). Based on the

results of the final model, a one barrel (42 gallon) increase in annual petroleum consumption is associated with an increase of 22.3663 grams of CO2 emissions per mile.

$$CO2tailpipeGPM = \beta_0 + \beta_1(barrels08) + \beta_2(comb08) + \beta_4(displ) + \beta_5(cylinders) + \beta_6(volume) + \varepsilon \quad (4)$$

Discussion

Hypothesis Substantiated?

In the initial model and the final model, the null hypothesis that the variable annual oil consumption in barrels had no significant effect on CO2 emissions was able to be rejected in favor of the alternative. In both models the variable barrels08 was found to be statistically significant ($p<.0001$) at a 95% confidence level. Additionally given the strong relationship visualized in Figure 1 it is apparent the two variables share a close relationship.

In addition to the rejection of the null hypothesis, the nominal variables, despite being removed in the final model, provided valuable information. The breakdown of the characteristics within each category yield insights such as the ability to design electric Passenger 2-Door and 4-Door as that demographic has room to grow. Currently both 2-Door and 4-Door types are underrepresented in electric vehicles compared to the population, indicating a gap Nissan could capitalize on. Budget could be allocated to the R&D team to innovate electric vehicles in these categories, allowing Nissan to gain competitive advantage, leading to more market share, higher stock prices and brand recognition as a leader in creating electric vehicles.

Unfortunately, any further interpretation of the results of the study requires further analysis. Due to concerns arising from the presence of multicollinearity, the actual coefficients are not correct and thus caution should be taken when interpreting the results of the final regression model. Additionally, the nominal variables that were removed due to improper encoding, should be

reintroduced correctly encoded as indicator variables to gain insight into the important relationships that were dropped from the final model.

Study Strengths

This dataset benefits from being a robust representation of the United States vehicle population. The 43,177 observations of vehicle fuel efficiency give us deep insight into the relationships between CO2 emissions and different vehicle characteristics. Benefits of a large dataset like this include allowing us to test many different relationships and develop models with increased accuracy across a range of variables. This dataset also represents a large part of the population of vehicles driven within the United States allowing for us to draw accurate conclusions about CO2 emissions within the United States.

Study Weaknesses

Some weaknesses of this dataset include the fact it is a secondary dataset which can introduce bias from the original creator in its collection and cleaning. Additionally, the interpretability of these results outside of the United States may not be applicable. This is because the dataset represents vehicles in the United States. Other countries may have different regulation requirements, vehicle models, and fuel efficiencies. Depending upon these different factors the results may not tell the same story. By expanding this dataset to include current vehicles as well as vehicles from different countries we can start to gain a global view on CO2 emissions and vehicle characteristics to gain a better understanding of fuel efficiencies by country. This will allow us to tailor our marketing strategy globally, not just within the United States.

References

Bhandari, A. (2020, March 19). What is Multicollinearity? Here's Everything You Need to Know. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/

Devore, J. L. (2016). *Probability and Statistics for Engineering and the Sciences (9th ed.)*. Cengage Learning US. https://bookshelf.vitalsource.com/books/9781305465329

EPA. (2023, October 5). *Sources of Greenhouse Gas Emissions*. United States Environmental Protection Agency. https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions

Glen, S. (2022). *F Statistic / F Value: Definition and How to Run an F-Test*. Statistics How To. https://www.statisticshowto.com/probability-and-statistics/f-statistic-value-test/

Mickūnaitis, V., Pikūnas, A., & Igor Mackoit, I. (2007). REDUCING FUEL CONSUMPTION AND CO2 EMISSION IN MOTOR CARS. ResearchGate. https://www.researchgate.net/publication/26541377_Reducing_fuel_consumption_and_CO2_emission_in_motor_cars

U.S. Department of Energy. (2020, December). https://www.fueleconomy.gov/feg/ws/index.shtml

Appendix

Table 1

Table of Descriptive Statistics

	co2TailpipeGpm	*barrels08*	*comb08*	*cylinders*	*displ*	*volume*
Mean	462.7668	17.1532	20.8459	5.7129	3.2867	66.9290
Standard Error	0.6005	0.0224	0.0396	0.0085	0.0066	0.3323
Median	444.3500	16.4805	20.0000	6.0000	3.0000	86.0000
Mode	493.7222	18.3117	18.0000	4.0000	2.0000	0.0000
Standard Deviation	124.7720	4.6629	8.2242	1.7642	1.3570	69.0417
Sample Variance	15568.0563	21.7425	67.6366	3.1123	1.8415	4766.7538
Kurtosis	2.0730	2.1176	64.7490	1.0592	-0.4919	-0.0313
Skewness	0.4175	0.3689	6.3508	0.8982	0.6581	0.6476
Range	1269.5714	47.0271	134.0000	14.0000	8.4000	538.0000
Minimum	0.0000	0.0600	7.0000	2.0000	0.0000	0.0000
Maximum	1269.5714	47.0871	141.0000	16.0000	8.4000	538.0000
Sum	19980883.9925	740622.2582	900064.0000	245181.0000	141063.5000	2889792.0000
Count	43177	43177	43177	42917	42919	43177
		Make Id	*Veh Type*	*EmissionsCat*	*transtypeid*	*primaryfueltype*
Mean		62.9295	1.2462	4.0361	1.2998	2.4524
Standard Error		0.1837	0.0062	0.0034	0.0022	0.0048
Median		52.0000	1.0000	4.0000	1.0000	3.0000
Mode		23.0000	0.0000	4.0000	1.0000	3.0000
Standard Deviation		38.1697	1.2863	0.7093	0.4587	0.9879
Sample Variance		1456.9291	1.6546	0.5030	0.2104	0.9759
Kurtosis		-0.9874	-1.6207	3.1600	-1.2272	-0.3332
Skewness		0.4143	0.3169	-0.1273	0.8660	-0.4231
Range		137.0000	3.0000	5.0000	2.0000	5.0000
Minimum		1.0000	0.0000	1.0000	0.0000	1.0000
Maximum		138.0000	3.0000	6.0000	2.0000	6.0000
Sum		2717108.0000	53807.0000	174268.0000	56122.0000	105886.0000
Count		43177	43177	43177	43177	43177

TABLE 2

Characteristics of 43,177 Sample Vehicle Models by Primary Fuel Type

Variable	Population *N* (%) (N=43,177)	Premium Gasoline *n* (%) (n=12,801)	Midgrade Gasoline *n* (%) (n=130)	Regular Gasoline *n* (%) (n=28,733)	Diesel *n* (%) (n=1,196)	Natural Gas *n* (%) (n=60)	Electricity *n* (%) (n=257)	*p* value*
Vehicle Type								<.0001
Unknown (0)	19,730 (45.7%)	3,491 (27.3%)	90 (69.2%)	15,346 (53.4%)	685 (57.3%)	34 (56.7%)	84 (32.7%)	
Hatchback (1)	5,070 (11.7%)	1,313 (10.3%)	0 (0.0%)	3,535 (12.3%)	115 (9.6%)	2 (3.3%)	105 (40.9%)	
Passenger 2-Door (2)	6,394 (14.8%)	3,157 (24.7%)	12 (9.2%)	3,120 (10.9%)	103 (8.6%)	1 (1.7%)	1 (0.4%)	
Passenger 4-Door (3)	11,983 (27.8%)	4,840 (37.8%)	28 (21.5%)	6,732 (23.4%)	293 (24.5%)	23 (38.3%)	67 (26.1%)	
Transmission Type								<.0001
Automatic (1)	30,210 (70.0%)	9,411 (73.5%)	130 (100.0%)	19,588 (68.2%)	773 (64.6%)	60 (100.0%)	248 (100.0%)	
Manual (2)	12,956 (30.0%)	3,390 (26.5%)	0 (0.0%)	9,143 (31.8%)	423 (35.4%)	0 (0.0%)	0 (0.0%)	

TABLE 3

Association of Emissions Category by Fuel Type and Other Characteristics

Variable	Population *N* (%) (N=43,177)	Ultra-Low Emission *n* (%) (n=321)	Very-Low Emission *n* (%) (n=384)	Low Emission *n* (%) (n=5,556)	Standard *n* (%) (n=29,543)	Polluter *n* (%) (n=5,899)	Gross Polluter *n* (%) (n=1,474)	*p* value*
Primary Fuel Type								<.0001
Premium Gasoline (1)	12,801 (29.6%)	24 (7.5%)	70 (18.2%)	1,169 (21.0%)	9,798 (33.2%)	1,262 (21.4%)	478 (32.4%)	
Midgrade Gasoline (2)	130 (0.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	124 (0.4%)	6 (0.1%)	0 (0.0%)	
Regular Gasoline (3)	28,733 (66.5%)	40 (12.5%)	311 (81.0%)	4,066 (73.2%)	18,971 (64.2%)	4,358 (73.9%)	987 (67.0%)	
Diesel (4)	1,196 (2.8%)	0 (0.0%)	0 (0.0%)	303 (5.5%)	629 (2.1%)	259 (4.4%)	5 (0.3%)	
Natural Gas (5)	60 (0.1%)	0 (0.0%)	3 (0.8%)	18 (0.3%)	21 (0.1%)	14 (0.2%)	4 (0.3%)	
Electricity (6)	257 (0.6%)	257 (80.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Vehicle Type								<.0001
Unknown (0)	19,730 (45.7%)	91 (28.3%)	50 (13.0%)	739 (13.3%)	12,579 (42.6%)	5,119 (86.8%)	1,152 (78.2%)	
Hatchback (1)	5,070 (11.7%)	122 (38.0%)	128 (33.3%)	1,820 (32.8%)	2,952 (10.0%)	47 (0.8%)	1 (0.1%)	
Passenger 2-Door (2)	6,394 (14.8%)	7 (2.2%)	11 (2.9%)	703 (12.7%)	5,193 (17.6%)	339 (5.7%)	141 (9.6%)	
Passenger 4-Door (3)	11,983 (27.8%)	101 (31.5%)	195 (50.8%)	2,294 (41.3%)	8,819 (29.9%)	394 (6.7%)	180 (12.2%)	
Transmission Type								<.0001
Automatic (1)	30,210 (70.0%)	312 (100.0%)	301 (78.4%)	3,202 (57.6%)	20,730 (70.2%)	4,557 (77.3%)	1,108 (75.2%)	
Manual (2)	12,956 (30.0%)	0 (0.0%)	83 (21.6%)	2,354 (42.4%)	8,813 (29.8%)	1,341 (22.7%)	365 (24.8%)	

TABLE 4

Pearson Correlation Coefficients (N=42,917)

	co2TailpipeGpm	barrels08	comb08	make_id	displ	cylinders	volume	vehtype	emissionscat	prifueltype
co2TailpipeGpm	1.0000	.9885	-.9184	-.2157	.7954	.7438	-.4323	-.3626	.8894	-.1128
barrels08	.9885	1.0000	-.9050	-.2117	.7843	.7337	-.4266	-.3580	.8791	-.1084
comb08	-.9184	-.9050	1.0000	.2072	-.7327	-.6863	.4161	.3313	-.8415	.1234
make_id	-.2157	-.2117	.2072	1.0000	-.2823	-.2670	.1165	.0940	-.1755	.0710
displ	.7954	.7843	-.7327	-.2823	1.0000	.9046	-.3628	-.2631	.6703	-.2149
cylinders	.7438	.7337	-.6863	-.2670	.9046	1.0000	-.2648	-.1524	.6185	-.2181
volume	-.4323	-.4266	.4161	.1165	-.3628	-.2648	1.0000	.7418	-.3627	.0498
vehtype	-.3626	-.3580	.3313	.0940	-.2631	-.1524	.7418	1.0000	-.3054	-.0340
emissionscat	.8894	.8791	-.8415	-.1755	.6703	.6185	-.3627	-.3054	1.0000	-.0874
prifueltype	-.1128	-.1084	.1234	.0710	-.2149	-.2181	.0498	-.0340	-.0874	1.0000

Note: All correlation values resulted in a p -value < .0001.

Table 5

Chi Square Two Way Contingency Table

	Ultra-Low Emission	Very-Low Emission	Low Emission	Standard	Polluter	Gross Polluter	Total
Hatchback	188.57	152.45	2,089.61	77.06	601.87	171.09	3,280.65
Passenger 2-Door	34.57	36.99	17.44	152.96	327.12	27.36	596.44
Passenger 4-Door	1.59	73.37	366.77	46.86	943.98	128.28	1,560.87
Unknown	21.14	89.72	1,275.95	62.81	2,178.73	339.86	3,968.21
							9,406.17

Chi-SQ	9,406.17
All Estimated Counts >5	TRUE
Degrees of Freedom	15
Devore Table A.11	<.0001

Additional research needed to determine homogeneity between vehicle type and emission category

Figure 1

Scatter Plot of Emissions to Barrels

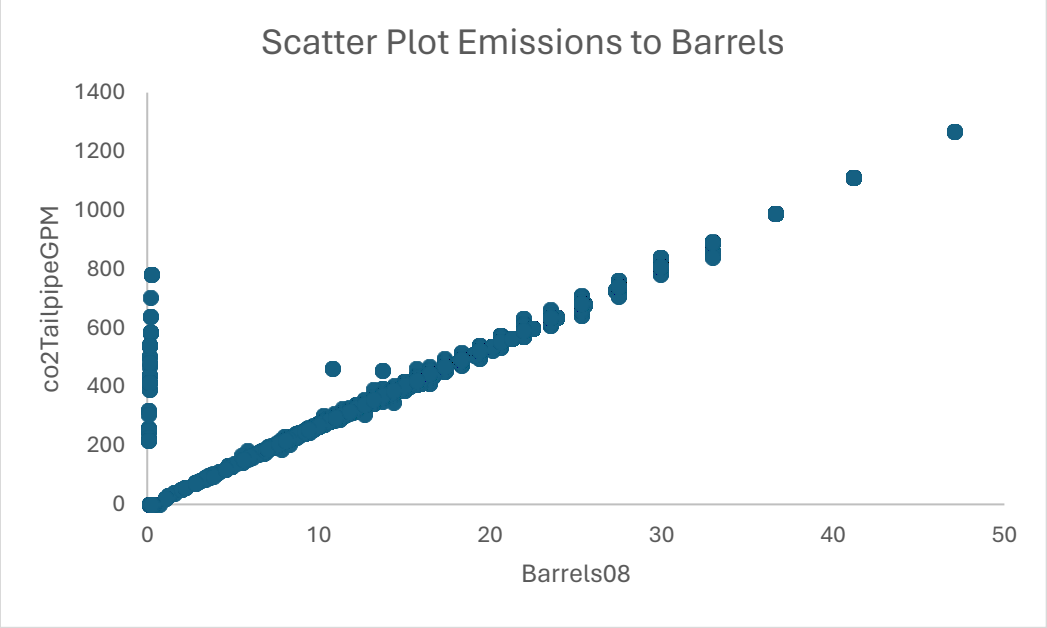

Figure 2

Normalized CO2 Emissions Distribution

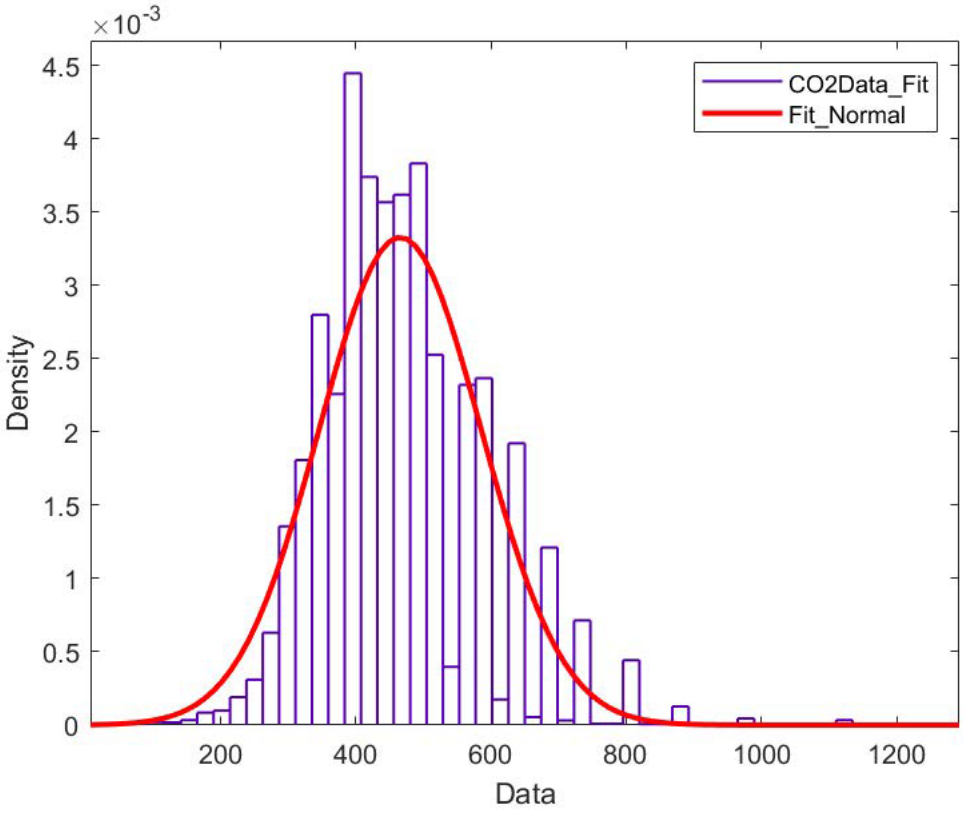

Figure 3

Initial Linear Regression Model

Linear Regression - Originally Planned Model

The REG Procedure
Model: MODEL1
Dependent Variable: co2TailpipeGpm

Number of Observations Read	43177
Number of Observations Used	42917
Number of Observations with Missing Values	260

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	606206477	60620648	245913	<.0001
Error	42906	10576874	246.51270		
Corrected Total	42916	616783351			

Root MSE	15.70072	R-Square	0.9829
Dependent Mean	465.54124	Adj R-Sq	0.9828
Coeff Var	3.37257		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	79.55324	1.56729	50.76	<.0001	0
barrels08	1	20.74377	0.05162	401.83	<.0001	9.34681
comb08	1	-2.53562	0.03490	-72.65	<.0001	6.08008
make_id	1	0.00333	0.00208	1.60	0.1093	1.08998
displ	1	2.17223	0.15090	14.39	<.0001	7.29952
cylinders	1	2.26299	0.10936	20.69	<.0001	6.47947
volume	1	-0.00779	0.00172	-4.54	<.0001	2.44999
vehtype	1	-0.41356	0.09181	-4.50	<.0001	2.43104
emissionscat	1	12.87660	0.24511	52.53	<.0001	4.71274
transtype_id	1	0.77998	0.17104	4.56	<.0001	1.07364
prifueltype	1	2.99025	0.08772	34.09	<.0001	1.21372

Figure 4

Final Linear Regression Model

Trimming Regression Model -- Removed Categorical Non-Indicator Variables

The REG Procedure
Model: MODEL1
Dependent Variable: co2TailpipeGpm

Number of Observations Read	43177
Number of Observations Used	42917
Number of Observations with Missing Values	260

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	605162908	121032582	446939	<.0001
Error	42911	11620443	270.80335		
Corrected Total	42916	616783351			

Root MSE	16.45610	R-Square	0.9812
Dependent Mean	465.54124	Adj R-Sq	0.9812
Coeff Var	3.53483		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	123.29415	1.41325	87.24	<.0001	0
barrels08	1	22.36625	0.04614	484.77	<.0001	6.79630
comb08	1	-2.78911	0.03512	-79.42	<.0001	5.60279
displ	1	3.20908	0.15361	20.89	<.0001	6.88532
cylinders	1	0.61407	0.10808	5.68	<.0001	5.76127
volume	1	-0.01339	0.00130	-10.31	<.0001	1.27869