Secondary Analysis of carbon dioxide emissions and petroleum consumption

Mayel Espino

University of San Diego

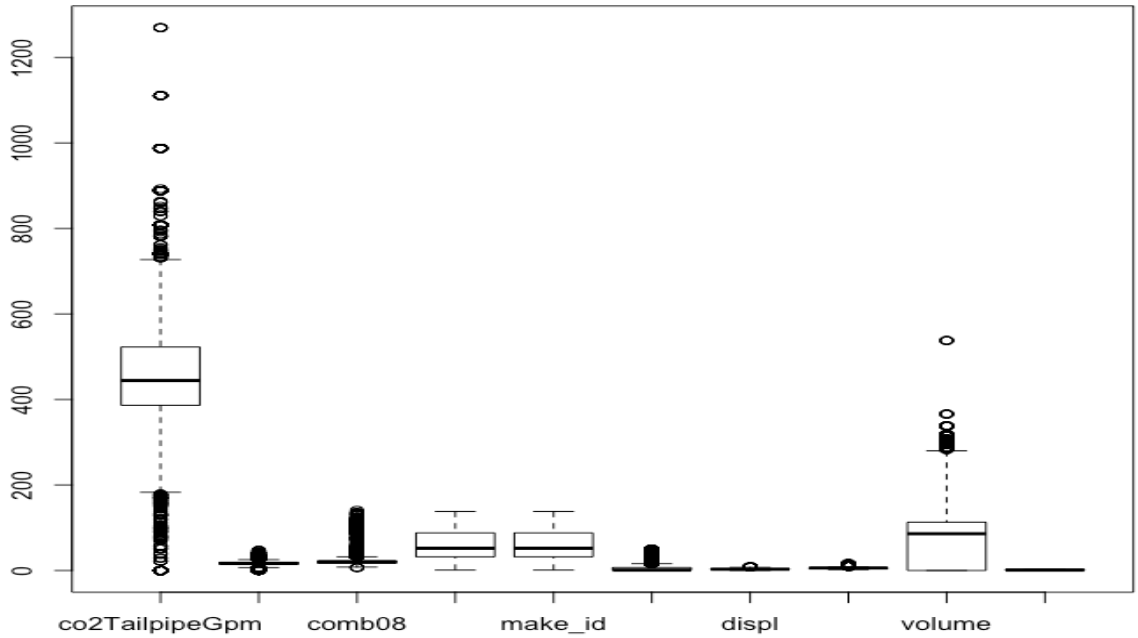
Abstract

This is a secondary analysis performed on the data set provided by the EPA, with 43,177 records, and nighty one variables, from which nine were initially considered to formulate the model. The hypothesis postulated in this study is that the carbon dioxide (CO2) emissions can be predicted by a statistical model based on a few key statistics. SAS was used to perform statistical analysis and tests such as bivariate, multivariant, chi, multicollinearity and regression tests. Based on the results of these tests the original model was revised and a new model with a slightly lower level of predictability was adopted, trading off for higher confidence and less variability. The conclusions and implications from this study are that the key variables selected for the final model have the most influence on the CO2 emission and that other variables which could have been considered, actually obscures and diminishes the strength of the model. The application of statistical tools and methods ensure the integrity of the model. The possible uses for the model selected as well as considerations for future studies are proposed in the discussion section of this study, also discussed are suggestions for further study. There are very few efforts more vital to our wellbeing, and that of future generations, than battling man made pollution and the impact it has on our environment.

Keywords: CO2, emissions, EPA.

Table of Contents

Boxplot of key variables



Introduction

Global warming is ubiquitous in the news coverage, and in the political discourse. With the effects that Greenhouse gas (GHG) emissions on global warming becoming undeniable and considering that twenty eight percent of the total greenhouse emissions in the United States originate from transportation, according to the Environmental Protection Agency (n.d.) https://www.epa.gov/transportation-air-pollution-and-climate-change/carbon-pollution-transportation, the need for a statistical model to study this in greater detail is clear. The objective of this study is to apply the statistical tools and methods we learned in class to formulate a statistical model. The statistical model will be tested and conclusions drawn about the hypotheses postulated on the model. The hypotheses will be used to determine what are the principal characteristics of the population that have the greatest effect on the emission of greenhouse gases. Further hypotheses should be formulated and conclusions drawn from this model given its relevance to the immediate future of humanity.

Best efforts will be made to ensure that the model created for this study and the hypotheses postulated in this study are unbiased and mathematically sound, so that that they can be trusted and proven useful for future use.

The research objective is to statistically investigate the primary-fuel type carbon dioxide emissions, in grams per mile, to primary-fuel petroleum consumption in barrels after controlling for combined miles-per-gallon for the primary fuel-type, vehicle manufacturer, make, model, engine displacement, engine cylinders, and combined luggage and passenger volume, in cubic feet. This is a secondary study, the original study was done by the EPA and the sample data was acquired from said study. The statistical module formulated in this research can be used for a variety of goals, however the primary goal of this research is to use this model to educate the consumers of the impact that the vehicle configuration they choose has on the environment.

The following is the initial generalized object formula or general lineal model ("general lineal model", 2021) co2TailpipeGpm = $\beta_0 + \beta_1$ cylinders + β_2 barrels08 + β_3 comb08 + β_4 displ + β_5 volume + β_6 vehtype + β_7 make + β_8 mfrCode + β_9 make_id + ε (1)

This formula helps us describe or model the relationships between all the variables involved. The letters β_n, β_0 to β_9 in this case, are the coefficients which indicate in what degree each variable in the term in the equation affects the dependent variable, while holding all the other variables the same. In this case the dependent variable is co2TailpipeGpm.

The GLM also accounts for an error ε alongside all the factors of our equation. The error is not a mistake but rather a deviation from our model. The error ε , also called υ is the unknown that the rest of the variables β_0, β_1 cylinders … β_9 make_id, do not account for.

β_0 is the baseline value for our model, when plotting the regression line on a graph, this is the location on the Y axis that our regression line intercepts.

The variables being multiplied by the βn coefficient, for example: cylinders are called predictor variables and the variable on the left hand side of the equation is called the dependent variable.

The hypothesis of this study postulates is that the behavior of the co2TailpipeGpm can be described in terms of variables on the righthand side of the GLM formula shown above, therefore allowing us to make good predictions based on the model, (Turner, 2008) Introduction to generalized linear models.

Method

The data is comprised of a variety of metrics on attributes that define an automobile. We were provided a data sample obtained from FuelEconomy.gov web services. The size is considerable, with over forty three thousand records. The sample has a mix of pure electric vehicles 0.5952% , Premium and Electricity 0.2131%, Premium Gas or Electricity 0.1228%, Regular Gas and

Electricity 0.1320%, Regular Gas or Electricity 0.0093%, and internal combustion vehicles 98.93%. For all vehicles the same variables are measured and recoded, there are two categories of variables: key variables and control variables. The key variables are

primary fuel tailpipe carbon dioxide emissions in grams per mile, and annual primary-fuel petroleum consumption in barrels. The control variables are: Combined miles-per-gallon for the primary fuel type, make, engine displacement in litters, engine cylinders, combined luggage and passenger volume in cubic feet, vehicle type, transmission type, and primary fuel type.

The following table provides descriptive statistics for the key variables.

From the counts in Table 3 we can observe that the majority of the sample population are categorized as either “Other” or “blank”. This categorial data is of limited use for our model or our hypothesis. The sample proportions across transmission type and fuel type present a clear pattern that may be considered as a predictor variable in our model.

The boxplot graph in Figure 1, clearly shows the distribution of the key variables side by side and from the comparison we can derive that the sample population distributions are very dissimilar. Most are discrete variables, a few are categorical, such as make. What each variable measures is quite different and loosely coupled. This means we will need to process and model each of the variables independently.

Table 4 and Table 5 are the bivariate table which illustrates the sample proportions of two of our categorical variables in relationship to each other. The variables being considered are vehicle type and transmission type.

From the tables we should note some key relationships between the proportions of the sample population and specific categories. One key relationship is the proportional amount of unknown vehicle type to the all the fuel types. From this I conclude that comparing fuel type to vehicle type is of very limited value. On the other hand, the comparison of transmission type to fuel type is very useful.

Table 4 shows a similarly the associations between four variables, in this case the fuel type, vehicle type and transmission type against emissions category. In this table, we can see patterns in the relationships between the variables that are intuitive and logical, for example: The fact that the lowest emissions category has the largest proportion in the electrical vehicle variable, or that the gross polluters have the highest proportion in the most popular fuel time, regular gasoline. Additionally the way that the proportions are related to the transmission type is also expected, for example no electric vehicles have manual transmission.

Because the patterns seen in this associations table are consistent to what we expect to see in the general population, the sample data seems to be representative of the hypothetical population. Another useful conclusion is that the relationship between fuel type and emissions category is useful for our model. We took these relationships in to account to select the predictor variables for the statistical model we propose in the results section.

Results

Now we consider the probability density function for the CO2 emissions:

$$PDF_{co2} = f(x;\mu = 465.538, \sigma = 119.88) = \begin{cases} \frac{1}{[(\sqrt{2\pi})(119.88)]} e^{-[\frac{(x-465.538)^2}{[(2)(119.88)]^2}]} & x \geq 0 \\ 0 & x < 0 \end{cases} \qquad (2)$$

The probability density function (PDF) describes mathematically the probability of all possible values of our variable of interest, in this case CO2 emissions, these values plotted on a graph give us the distribution of all the probabilities. Considering the PDF we select the most appropriate statistical formulas and processes to apply.

From the CO2 emissions probability plot on Figure 2 we see that the sample data is very nearly normally distributed. We can see that the values at either end, that is the tails, are the farthest from the normal distribution. This is consistent from the boxplot for CO2 in Figure 1. The values in the tails, the outliers, correspond to the measurements from pure electric and hybrid automobiles and internal combustion automobiles that are extreme polluters.

Table 5 and Table 6 contain the correlation coefficients of CO2 emissions, fuel consumption, fuel type, automobile make, engine displacement, emissions category and transmission type. From the tables we arrive at a few key observations, there is a strong correlation, grater than .9, between The CO2 emissions and : The fuel consumption, the engine displacement, the number of cylinders, and the emissions category. There is also strong but inverse correlation, greater than .9, between the CO2 emissions and the combined fuel type.

On the other hand there is a marked lack of correlation between CO2 emissions and: automobile make, automobile type and primary fuel type. This is unsurprising and a good indication that these variables may not be very useful for our model or hypothesis.

There are other strong correlations, positive and negative between annual fuel consumption and: combined fuel consumption, engine displacement, cylinders and emissions category. These are unsurprising and not very useful for our model or hypothesis. What we are testing here is if the polluter category is independent from the vehicle classification: two door, four door, hatchback and unknown.

Revising the model

The original model was described in formula (1), now we can examine and remove some of the predictor variables to arrive at a final model.

Predictor variables not selected

The first set of variables are the easiest to address, the initial model (1) does not include them precisely because they do not contribute to the prediction power or ability of the model. In particular the categorical variable “unknown” . There is the temptation of reducing the error term ε by taking in to account these variables. Even though efforts could be made to statistically predict and fill in the values, in order to recategorize the data in this field, the return on investment does not merit these efforts. The data would add very little additional prediction ability or accuracy to the model. Many other metrics could be mentioned here, which are in the data set but not taken in to account for the model, for example: drive or sub-model. There are a couple of metrics which I would initially consider for the model, and that perhaps may be used for another research project, such as: Highway and City millage rating, because this is often used as a criteria for selecting an automobile.

After reviewing the bivariate, associations and the Pearson’s correlation tables, it becomes clear and demonstrable by the statistical calculations that the following variables are not contributing significantly to the predictability power, or statistical significance of the model:

- make_id. – manufacturer
- vehtype. - categorized vehicle type
- prifueltype. - primary fuel type
- trans_type. – Transmission type

These metrics under-correlated to our dependent variable, co2TailpipeGpm, contribute very little to the model.

Intercorrelated and Inflationary variables

Some other variables, on initial consideration, seem to contribute quite a bit to the prediction power of the model, but upon further application of statistical methods we come to realize that they are not required and in some cases they actually amplify the effect of another variable already in the model, doing more harm than good by increasing the variability in the model's predictions. These variables are:

- barrels08. – Annual petroleum consumption
- comb08. – Combined miles per gallon
- emissionscat. – Emissions category

The metrics barrels08, comb08 and emissionscat are highly correlated to the dependent variable and to each other. They do not contribute to the model any significant value that the displacement and cylinders predictors do not also contribute. On initial consideration, cylinders and displacement would also seem to be equally redundant predictors. After researching the relationship between cylinders and displacement to engine performance, I came to realize that the different combinations have different performance. Say we have two engines with the same displacement but different number of cylinders, the revolutions per minute (RPM) increase and torque decreses. So the performance of the engine will be different dependent on this combination of predictor variables, AnonymousIsAnnon (2016, May).

The omission of the variables mentioned above is further reinforced by the linear regression results:

Parameter Estimates

variable	DF	Parameter Estimate	Standard error	t value	Pr > |t|	Variance Inflation
intercept	1	249.699	1.321	188.9	<.0001	0
displ	1	48.764	0.608	80.11	<.0001	6.049
cylinders	1	13.379	0.452	29.57	<.0002	5.649
volume	1	-0.312	0.005	-59.09	<.0003	1.182

Final model

Hence, we arrive at the final model:

$$co2TailpipeGpm = \beta_0 + \beta_1 \text{ cylinders} + \beta_2 \text{ displ} + \beta_3 \text{ volume} + \varepsilon \quad (3)$$

$$co2TailpipeGpm = 249.699 + 13.379 * \text{cylinders} + 48.764 * \text{displ} + -0.312 \text{ volume} + \varepsilon \quad (4)$$

Discussion

Strengths of the model

The strengths of this study are in the statistical tools and methods that were applied, and in particular in the model, the hypothesis formulation, the tests and determination of the hypothesis. The statistical methodology we followed insures that we are unbiased in our conclusions, free from any preconceived notions we may or may not be aware of. To that end we follow reliable techniques and statistical tests based on mathematical principles and a history of experimental and/or empirical knowledge.

Weakness of the model

The final model has only three variables from the nine initially considered, furthermore in the data set collected there many more variables. These three variables may not be enough to explain fully the CO2 emissions. These were selected from the variables available in the dataset, other variables that could have been considered are for example: Demographic information, driving habits, traffic patterns etcetera.

It is tempting to entertain various combinations of these variables to be added to the model in order to make it more robust ant the error term, ε, smaller. Here we need to strike the balance between having too few or too many variables. Too few variables and our model becomes marginally useful, too much variability with an ε term that is substantial. Too many variables and our model becomes too sensitive to the variability in our data, even though our ε term becomes small. Another danger of too many variables would be the influence, multicollinearity , and other effects of certain variables on the other variables and the model overall.

The process of collecting the data was not explored in our secondary research, in the implications for the future section, some more variables are proposed. From the conditions and content of the data, it can be inferred that the collection tools and methods can be improved and therefore are a weakness in this study.

Implications for future research

In the future research gathering other variables, dealing with: driving habits, demographics, consumer preferences, geography, traffic and climate patterns. All of which are worth considering to revise our model after careful application of all the same methods and calculations in this research paper. Consumer preferences and demographics have a large impact, as stated in (Shigeta & Hosseini, 2021) “This shows that American customers demand vehicles with strong engines. On the other hand, Japanese customers demand a higher fuel economy; therefore, hybrid, electric, and other compact cars are popular.”. Perhaps this models and others like this can be used in educating the American customers on the impact of their choices.

Applications of model

The model selected in this research can be applied to calculate the CO2 emissions, based on the variables provided, on any data for a variety of purposes. There are a variety of applications of this model : The purpose could be to educate the consumers, to find the optimal pricing of vehicles while remaining within the parameters of present regulations or to present arguments to change present regulations. There is good reason to believe that educating the educating the consumers with facts, data such as this model provides is effective: Environmental Protection Agency(2021, February, 20) History of Reducing Air Pollution from Transportation in the

United States, judging by the effects of lower CO_2 emissions on the presence of led in the blood for example. There are other success stories in this article and that provides an optimistic outlook for the future.

Appendix

List of figures

Figure 1

Boxplot of key variables

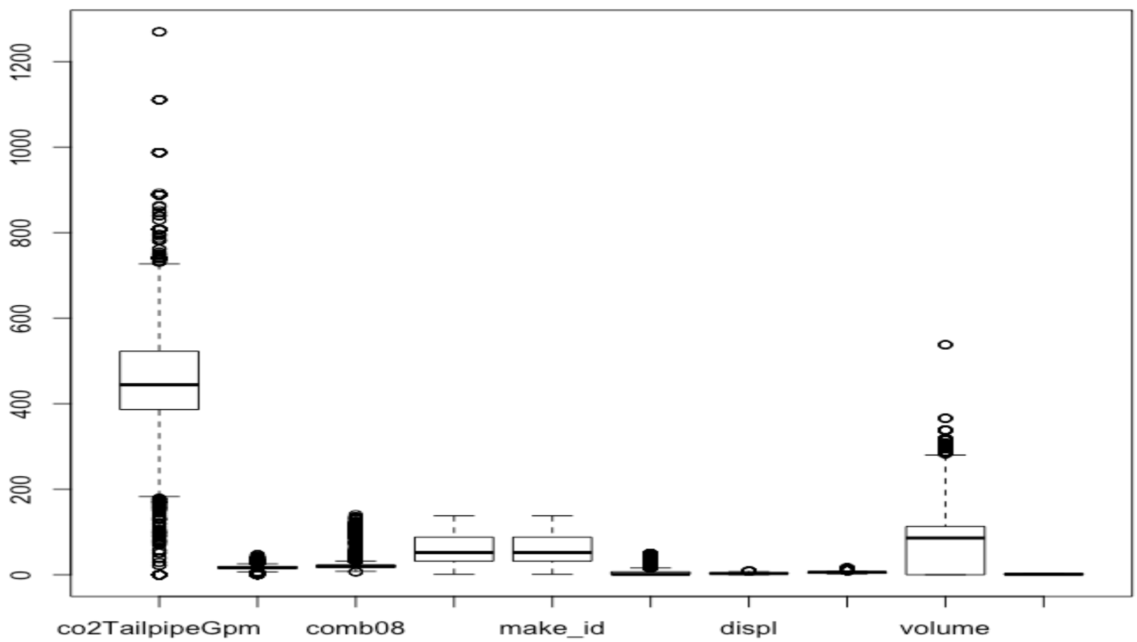
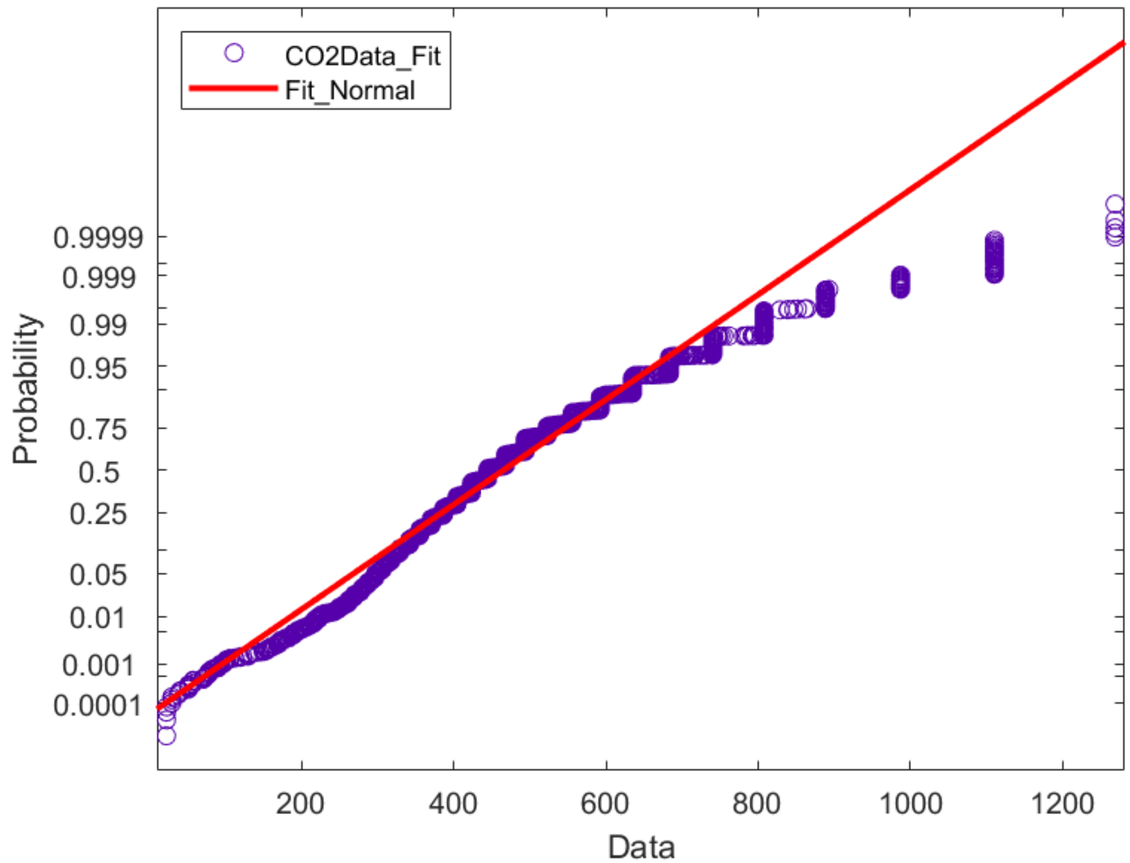


Figure 2

Normal probability plot



List of Tables

Table 1

Descriptive statistics, discrete data distribution and location

	co2TailpipeGpm	*barrels08*	*comb08*	*displ*	*volume*
Minimum	0.0	0.06	7.00	0.000	0.00
1st Quartile	386.4	14.33	17.00	2.200	0.00
Median	444.4	16.48	20.00	3.000	86.00

Mean	462.8	17.15	20.85	3.287	66.93
3rd Quartile	522.8	19.39	23.00	4.300	113.00
Maximum	1269.6	47.09	141.00	8.400	538.00

Table 2

Descriptive statistics, categorical data counts

make	*count*	*mfrCode*	*count*
Chevrolet	0.0	blank	30808
Ford	386.4	GMX	1786
Dodge	444.4	BMX	1378
GMC	462.8	FMX	1023
Toyota	522.8	CRX	964
BMW	1269.6	TYX	941
		Other	6277

Table 3

Bivariate table

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%)	

* *p* values based on Pearson chi-square test of association.

Table 4

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%)	

* *p* values based on Pearson chi-square test of association.

Table 5

Pearson Correlation Coefficients (N=42,917)

	co2TailpipeGpm	barrels08	comb08	make_id	displ	cylinders	volume	vehtype	emissionscat	transtype_id
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 6

Contingency and chi-square table

		Hatchback	Passenger 2-Door	Passenger 4-Door	Unknown	Total
GROSS POLLUTER	Frequency	1.00	141.00	180.00	1,152.00	1,474.00
	Expected	173.08	218.28	409.08	673.55	-
	Percent	-	0.33	0.42	2.67	3.41
LOW EMISSION	Frequency	1,820.00	703.00	2,294.00	739.00	5,556.00
	Expected	652.41	822.78	1,542.00	2,538.80	-
	Percent	4.22	1.63	5.31	1.71	12.87
POLLUTER	Frequency	47.00	339.00	394.00	5,119.00	5,899.00
	Expected	692.68	873.57	1,637.20	2,695.60	-
	Percent	0.11	0.79	0.91	11.86	13.66
STANDARD	Frequency	2,952.00	5,193.00	8,819.00	12,579.00	29,543.00
	Expected	3,469.00	4,375.00	8,199.10	13,500.00	-
	Percent	6.84	12.03	20.43	29.13	68.42
ULTRA-LOW EMISSION	Frequency	122.00	7.00	101.00	91.00	321.00
	Expected	37.69	47.54	89.09	146.68	-
	Percent	0.28	0.02	0.23	0.21	0.74
VERY-LOW EMISSION	Frequency	128.00	11.00	195.00	50.00	384.00
	Expected	45.09	56.87	106.57	175.47	-
	Percent	0.30	0.03	0.45	0.12	0.89
TOTAL	Frequency	5,070.00	6,394.00	11,983.00	19,730.00	43,177.00
	Percent	11.74	14.81	27.75	45.70	100.00

Degrees of freedom = 15, α = 0.5, we get a p-value < 2.2 e^{-16} (nearly zero).

References

Environmental protection Agency (2021, Febriary, 20) Transportation and climate change https://www.epa.gov/transportation-air-pollution-and-climate-change/carbon-pollution-transportation

Environmental Protection Agency (2021, February, 20) Fast facts on transportation greenhouse gas emissions https://www.epa.gov/greenvehicles/fast-facts-transportation-greenhouse-gas-emissions

Turner, Heather (2008) Introduction to generalized linear models http://statmath.wu.ac.at/courses/heather_turner/glmCourse_001.pdf

AnonymousIsAnnon (2016, May) In Reddit forum (https://www.reddit.com/r/Autos/comments/4htqh8/engine_displacement_vs_amount_of_cylinders/

Environmental Protection Agency(2021, February, 20) History of Reducing Air Pollution from Transportation in the United States https://www.epa.gov/transportation-air-pollution-and-climate-change/accomplishments-and-success-air-pollution-transportation

Shigeta, N.; Hosseini, S.E. Sustainable Development of the Automobile Industry in the United States, Europe, and Japan with Special Focus on the Vehicles' Power Sources. *Energies* **2021**,*14*,78. https://dx.doi. org/10.3390/en14010078