Relation between Carbon Dioxide emission with different vehicle characteristics

S M Sultan Mahmud Rahat

Shiley-Marcos School of Engineering, University of San Diego

Masters of Applied Data Science

Abstract

Emission of CO2 is directly related with the global warming. The main objective of the analysis to find out the association of different variable such as primary fuel type, annual petroleum consumption, vehicle manufacturer, engine displacement, transmission type, engine cylinders, and combined luggage and passenger volume in cubic feet etc. with CO2 emission. Initial hypothesis was that these variables have no effect on CO2 emission. After going through several statistical analysis by the data collected from FuelEconomy.gov website, for weak correlation coefficient and multicollinearity of the variables, many variables resonate with the hypothesis. But engine displacement, engine cylinders, and combined luggage and passenger volume in cubic feet are strongly related with each other and these variables are counterintuitive to the hypothesis. From the statistical analysis, emission of CO2 shows a normal distribution with a mean and standard deviation of 465.538 and 119.88. By bivariate analysis, unknown vehicle types use highest portion for midgrade gasoline, regular gasoline and natural gas as a primary fuel. For the weak correlation coefficient and multicollinearity, only three variables; engine displacement, engine cylinders, and combined volume are only used to make a regression model with a coefficient of 6.04887, 5.64859 and 1.18174 respectively. The p-value of the regression model is less than 0.001 that makes the null hypothesis rejected for any confidence level of 95%. It maintains R^2 value of 0.6632 which hinders the model of being overfitted. So, this final regression model may use to predict and estimate CO_2 emission in different dependent variable.

Keywords: CO_2 emission, vehicle, regression model, p-value, statistical analysis, correlation coefficient

Contents

List of Tables

List of Figures

List of Equation

Introduction

Climate change is the most alarming and disastrous event happening in 21 centuries. It shows a drastic change in climate all over the world. Recent events like Pakistan and Bangladesh's recent flood in 2022, the Australian wildfire in 2020, Megadrought and wildfire in California state, etc. Carbon dioxide, the main primary gas among all greenhouse gases, contributes to entrapping heat inside the earth and causing climate change. 25% of Carbon dioxide emissions are produced by the transport section. (Zhang et al., 2019) And passenger car contributes 41% of this total emission. (Global Transport CO2 Emissions Breakdown 2020 | Statista, n.d.) So passenger car contributes a significant impact on global catastrophes. It is important to get measure carbon dioxide emission to see the impact of different variables like the primary fuel type, annual petroleum consumption, vehicle manufacturer, engine displacement, transmission type, engine cylinders, and combined luggage and passenger volume in cubic feet. The relation between these variables with carbon dioxide emission can depict a clear picture that may help the government and the customer market to push the car market toward a fuel-efficient car industry. Studying a wide range of variables will help the manufacturer to make new policies and production lines to reduce CO_2 emissions. Government regulation may be easier to impose with less stringent rules as all variables are not weighted the same impact on tailpipe CO_2 emission.

Generalized Object Formula

The generalized object equation used to estimate CO_2 is described in equation 1.

$$CO_2 = \beta_\circ + \beta_1\, AnnualPetroleumConsumption + \beta_2\, MPG + \beta_3\, Manufacturer + \beta_4\, EngineDisplacement + \beta_5\, Volume + \beta_6\, Cylinder +$$

$$\beta_7\, TransmissionType \begin{pmatrix} 1 = Automatic \\ 2 = Manual \end{pmatrix} + \beta_8\, VehicleType \begin{pmatrix} 0{=}Unknown \\ 1{=}Hatchback \\ 2{=}Passenger\ 2{-}Door \\ 3{=}Passenger\ 4{-}Door \end{pmatrix} +$$

$$\beta_9\, PrimaryFuelType \begin{pmatrix} 1{=}Premium\ gasoline \\ 2{=}Midgrade\ Gasoline \\ 3{=}Regular\ Gasoline \\ 4{=}Diesel \\ 5{=}Natural\ Gas \\ 6{=}Electricity \end{pmatrix} + \epsilon \quad (1)$$

where,

β_0 = the y-intercept of the equation

β_n = Coefficient of the corresponding variables of vehicles

ϵ = an error term for the observations which do not follow the regression line.

CO_2 emission has a different relationship with the different characteristics of the vehicles. To quantify the CO_2 emission, equation 1 is used. It depicts all the relations of CO_2 emission with various characteristics of the vehicle by implying multiple linear regression with proper constant and error terms. This prototype model can help to guide car manufacturers to pre-mass production and design cars, also government can make proper instructions to force the automobile industry more precisely. The hypothesis from the equation is important to find the relation between different variables with CO_2 emission. β_0 is the intercept of the equation and β_n is the corresponding variable to make the multivariable regression but some will not follow the regression line, for these, ϵ is applied as an error term for the observation.

Method

Data collection and preprocessing:

The data which is used in the project are collected from FuelEconomy.gov and imported as a CSV file and opened in Microsoft Excel. Other than converting data into table formats, there are no

manipulation occurred in the dataset. This data will help to make different regression model to statistical analysis and prediction.

Population characteristics

A total population of 43,177 samples of the different vehicles is gathered as a population and 49 variables are used but among these, 11 variables: - fuel tailpipe carbon dioxide emissions in grams per mile (co2TailpipeGpm), annual petroleum consumption in barrels (barrels08), combined miles-per-gallon for the primary fuel type (comb08), engine cylinders (cylinders), vehicle manufacturer Id (make_id), engine displacement in liters (displ), combined luggage, and passenger volume in cubic feet (volume), categorized vehicle type (vehtype), emissions category (emissionscat), transmission type (transtype_id), and primary fuel type (prifueltype), are used to make different correlations to get a conclusion. Among these variables, tailpipe carbon dioxide emissions in grams per mile, and annual petroleum consumption in barrels show continuous data where they both are positively skewed but Combined MPG and engine cylinder show discrete data, and manufacturer (Division), model name (carline) and fuel type show nominal data. Though Fuel type could be ordinal data as it shows an order of hierarchy, for electricity and diesel, it has turned into a nominal variable.

In the fuel type, the top 3 modes of using fuel are Regular gasoline, Premium gasoline, and diesel with 27294, 12529, and 1196 respectively. Among engine cylinders, the top 3 modes of using cylinders are cylinder- 4, cylinder- 6, and cylinder- 8 with counts of 16820, 14862, and 9218 respectively. On the other hand, cylinder-16 shows the least number of samples with a count of 14. In these samples, some vehicles are from 1985-2007 and some vehicles are from 2011-2016. EPA test is used to measure tailpipe CO_2 measurement for 2013 models and beyond.

Methodology

Table 1 shows descriptive information about the continuous variables as the discrete variable shows a different standard deviation and different sample variance than another continuous variable.

The average quantity of consuming barrels of petroleum is 16.48 barrels with a standard deviation of 4.66 and a variance of 21.74. Total consumption of petroleum by the sample vehicle is 740622 barrels and they show a moderate positive skewness of 0.37. Figure 2, shows that 91.57% population of the total vehicles consume 10 to 24 barrels of petroleum per year. The average quality of tailpipe CO_2 emission is 444.36 GPM with a standard deviation of 124.77 GPM and a sample deviation of 15568.06 GPM. They also show a positive skewness. From figure 1, the histogram depicts that a total of 82.73% of the population emits 300 to 599 MPG CO_2. Engine displacement means of the population is 3.24 with a standard deviation of 1.42 and sample variance of 2.02 and the total sum of the liter used for the displacement of the engine is 20396.1 liter. Table 1 shows all mean, median, and standard deviation of the continuous variable but not the discrete variables because the discrete variable can't be computed standard deviation and sample variance with the normal process. Discrete variables like engine cylinders have a different relation with variables like CO_2 emission and annual consumption of petroleum. In Figure 3: the boxplot shows that the category of discrete variables like a cylinder is quite visible and some of the data from point 9 are vague and unstructured. It shows that cylinder-6 is used by the major portion of the population, and it has some countable outliers which may change the descriptive statistics of cylinder-6 against CO_2 emissions. Cylinder-2 has the least variation among other engine cylinders. By using cylinder 3 and cylinder 4, the least emission of CO_2 is possible. In figure 4, the Cylinder engine is plotted against the annual consumption of petroleum. It shows that cylinder 16 has the least amount of outlier but it is responsible for most CO_2 emissions and petroleum consumption. Cylinder 3 and cylinder 4 have minimum consumption of petroleum but cylinder 4 and cylinder 6 have a wide range of outliers. Like figure 3, cylinder 2 has the least variety of samples in terms of consumption.

Visualization and explanation of Tailpipe CO2 emission

In Figure 5, The boxplot of the CO_2 emission depicts a clear picture of the frequency of cars on different levels of emission with different groups of fuel types, general gasoline, premium gasoline,

diesel, natural gas, and midgrade gasoline. This boxplot will help the US EPA (Environmental Protective Agency) assumption of the carbon emission on large scale to push the car manufacturer

Natural gas has the highest Q3 value (583.83 GMP) but they have the least Q1 value (253.99 GMP). So, they maintain the highest IQR value of (331.84 GMP). That means they are responsible for the highest CO_2 rate and lowest CO_2 rate. These vague conclusions and outcomes have happened because of their small count sample (60 counts). According to these samples, it is concluded that government should take more samples of natural gas fuel-type cars to get a concrete conclusion. The manufacturer also should focus on their CO_2 emission technology as their emission of CO_2 is not stable.

Midgrade Gasoline is a good choice for using CO_2 as its upper extreme is minimum (614 GMP) and have the minimum IQR value (67 GMP), but they have the highest mean, median, Q1, and lower extreme value which is 503.92,513,466 and 377 GMP respectively. It seems that the manufacturer has a good grasp on using midgrade gasoline. But the sample is inadequate (count 130) to get a strong conclusion. But if they need to be tested under a wide variety of samples, otherwise their actual picture can't be depicted.

In terms of diesel, the counting sample is moderate (count 1196). They show the highest value (122.94 GMP) in the range of Q3-Q2. And their median is lowest among other groups, though the median of all groups is closely similar. But their CO_2 emission is quite predictable and within range.

Premium gasoline has an inexhaustible number of outliers, but their range in (Q2-Q1) and Q3 is lowest than other groups of fuel, 45.05 and 516 GMP respectively. Customers should get aware of using premium gasoline as they have a lot of outliers, so the inappropriate choice can lead to a lot of CO_2 emissions. Their mean is only 467.64 GMP though they have a lot of outliers, that's why if the manufacturer can concern with those outliers, premium gasoline can be a good choice.

Samples taken from regular gasoline are high (count 28733). So, the government should reduce its sample size or increase the size of other groups so that a similar weighted comparison can be made.

As they have the highest number of samples that's why their outliers are also the highest, and their lower extreme is the lowest of 174 GMP but their IQR is only 331.84 GMP. So, their CO_2 emissions are in control and easy to predict.

Bivariate Frequency Analysis

In Table 3, the bivariate frequency table of 43,177 sample vehicles is given by primary fuel type. Among 43177 observations, the total number of hatchbacks, passenger - door, and passenger 4- doors are 5070, 6394, and 11983 which are respectively 11.7%, 14.8%, and 27.8% of the total population. But the maximum vehicle type is "Unknown" which covers 45.7% of the total sample size with a count of 19,730. Figure 7, illustrates that passenger 4-door used premium gasoline and natural gas in larger proportions with 37.81% and 38.33%. But they use midgrade gasoline with a portion of 21.54% only. But Passenger 2-Door shows a different picture. They are used 0.39% of electricity fuel type; it may happen because maximum Passenger 2- Door cars are a luxury product which is known for high power density and speed but with electricity is very hard to produce this power and speed that's why premium gasoline is used as a primary fuel type for these types of vehicles and it is 24.66%. Hatchback cars are used electricity vastly with a percentage of 40.86%, but only 3.33% of these cars use natural gas and they do not use any midgrade gasoline as their primary fuel. All these "unknowns" covers a maximum portion of the samples. So, these car uses midgrade gasoline most of the time which is 69.23% and they use premium gasoline the least with a 27.27% proportion.

From Table 3, in transmission type, there are 2 types of vehicles that can be seen. 1) Automatic and 2) Manual. Maximum vehicles are used Automatic (70.0%). But this type is significantly dominant in to use of natural gas and electricity with 100%. But a majority of manual vehicles use regular gasoline and premium gasoline which is 31.8% and 26.5%, respectively, and midgrade gasoline, electricity, and natural gas- these fuels are not used by any manual vehicles.

Results

Probability Density Function

In Figure 6, 52 bin is arranged to a normal probability distribution. In this distribution, the average mean of the Tailpipe CO_2 emission is 456.538 so half of the area under the curve will be close to .5 but it is not a perfectly normal distribution. If the probability plot is applied, the diverted data are easily seen. From Figure 6, it has been seen that small values are less diverted than the higher value. There is some data missing also in the histogram. After 692.11 GPM only 3% of the vehicles are identified to cross which is 1,497 counts among 42,920 counts of vehicles. These slight numbers divert the normal probability distribution from being perfect.

$$PDF_{CO_2} = f\ (x; \mu = 465.538, \sigma = 119.88 = \begin{cases} \frac{1}{(\sqrt{2}\pi)(119.88)} e^{-\left[{(x-465.538)^2}/{2[\ (119.881^2)]}\right]},\ x \geq 0 \\ 0,\ x < 0 \end{cases} \quad \text{.....(2)}$$

Where,

PDF_{CO_2}= Probability function of tailpipe CO_2 emission.

μ = Population mean

$\sigma =$ Population standard deviation

Association of Emissions Category by fuel type and other characteristics

From Table 4, in primary fuel type, there are a total of 43,177 samples of observation which are organized according to their level of emission by each fuel type. In Gasoline, the maximum vehicle maintains a standard emission rate of 33.2% and low emission of 21.0 %. But according to population and proportion Regular gasoline performs very well in terms of emission. They exhibit a 73.2% proportion in low emission and 81.0 % rate in very low emission with a population of 4066 and 311, respectively. It shows this dominating proportion rate among the other variable because of their large sample size of 28,733 which is the majority portion of the observation and that is 66.5%. This uneven

sample size makes harder conclusive decision for a data scientist. Though they have a dominating sample size they show good performance in terms of emission.

Electricity vehicle shows extraordinary and conspicuous performance in CO_2 emission. They perform very well in CO_2 emission. Though their sample size is small all of the vehicles of them are in the Ultra-Low emission category which is very conclusive. Government should increase the subsidies to car manufacturers to increase their manufacturing products that can impact the total environmental issues.

Diesel and Midgrade vehicles are not good for the environment. There are not any samples of vehicles of these primary fuel types that are included in Ultra-Low emission and Very-low emission. Though 303 vehicles with diesel can be included in the Low emission category which is only 5.5% of the total proportion, on the other hand, Midgrade Gasoline vehicles cannot be included in the Low emission category. But among 130 vehicles of Midgrade Gasoline vehicles, 124 vehicles are included in the standard emission category but there are no vehicles in the gross polluter category.

In terms of vehicle type, the maximum number of vehicles is “Unknown” with a total sample size of 19,730 with a 45.07 % proportion. Among these cars, they have not been considered environment-friendly vehicles as they maintain a dominant proportion of 86.8% and 78.2% in the Polluter and Gross Polluter categories respectively. On the other hand, hatchback vehicles type is environment-friendly as they exhibit 38.0%, 33.3%, and 32.8% proportion rates in Ultra-Low emission, very-low emission, and low emission categories respectively though they have a small sample size with 5,070 observations and are 11.7% of the total sample. There are only 1 and 47 samples labeled in the Gross polluter and Polluters category which are 0.1% and 0.8% respectively.

Correlation Coefficient with different variables

Unlike in primary fuel type, in Table 5, there is a total of 42,917 samples are used in making the correlation coefficient between different characteristics of different variables. It helps to visualize the

relationship between two variables and helps the policymaker, manufacturer, and customer to make a conclusive decision.

Tailpipe CO_2 emission shows a strong positive and negative relationship with all the other variables. Like it shows strong positive with annual petroleum consumption in a barrel and classified CO_2 tailpipe emission with .9885 and .8894. It shows a strong negative relationship with a combined GPM (mile per gallon) emission of -.9184. Annual petroleum consumption shows some strong positive correlation with Tailpipe CO_2 emission and strong negative relation with combined GPM (mile per gallon) emission with .9885, -.9050 respectively. Displacement of the engine shows strong positive relation with engine cylinder number .9046 and does not show any strong negative correlation with any variable other than combined GPM (mile per gallon) emission with -.7327. Manufacturer (make_id) does not show any strong correlation with any of the other variables. All of their relations are weak and random.

Chi-squared test for homogeneity and Independence:

From Table 6, without any statistical approach, row percentage and column percentage show a visibly uneven distribution within the given variables. It depicts that two variables can never be dependent on each other and the homogeneity is unclear. But it needs a statistical approach to get a conclusion(Plackett, 1983). Devore (2016) states that the null hypothesis claims that two variables, emission category, and vehicle type, are independent of one another. The alternative hypothesis states that these two variables are dependent on each other. To test the hypothesis, a contingency table was made as Table 7 with the essential information of Table 6. Table 7 shows observed values, expected values, and chi-square values with sample numbers of 43,177 and a degree of freedom of 15. Expected value and chi-square contributions are calculated under this formula.

Expected Value:

$$\hat{e}_{ij} = \frac{n_i n_j}{n}$$

i = Total row

j = Total column

n = Total population

Chi-Square Statistics Formula:

$$X = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

According to the following formula, the chi-square value is 9406.05 and the p-value is <0.001 with the degree of freedom 15. All the data in the expected values are higher than 5, so there is not any constraint to applying the chi-square value. Chi-square supersedes the critical value because of the over-inflation of the large sample size. The p-value is <0.001 which can be rejected by any significant value which rejects the null hypothesis of their homogeneity and independence. But due to the large sample size with these two variables, it cannot be concluded that that hypothesis is rejected so they are the non-homogenous and dependent variable. Additional research is required to get a decisive result and better decision.

Multicollinearity

Determining the emission of CO_2 by a different variable of the vehicle is tough work to do as the variable are very close to each other and sometimes they show slight dependence on each other which is familiar to Car Industry. But these dependencies and relationships with the variables cause multicollinearity in the regression. It is very hard to avoid any multicollinearity in a regression model. The variance of factor (VIF) and tolerance are used to measure this. Generally, VIF higher than 5 shows an indication of multicollinearity. Adjusted R^2 value and multicollinearity of each variable help to remove

the unnecessary variables from the regression model which may cause deviation from a proper prediction.

In the trimmed regression model-limited variables- no null, it shows a higher proportion of multicollinearity, so it is hard to eliminate the higher level of multicollinearity from that model. Sometimes categorical variable shows different and unsynchronized relation with their corresponding independent variable. And categorical data shouldn't indicate with the That's why eliminating categorical data from the features of the independent variable may stable the R^2 value and help the model to get rid of overfitting.

Regression Model for decisive prognosis:

This is the final regression model to estimate the emission of CO_2

$$CO_2 = 249.69920 + 6.04887\ EngineDisplacement + 1.18174\ Volume + 5.64859\ Cylinder + \epsilon \quad (3)$$

In the final regression model, there are 3 independent variables which are displ, cylinders, and volumes. All these variables help to determine the CO_2 emission. Among all the variables, these independent variables are important and all the categorical variables are eliminated from the model. Categorical data without Boolean values become pointless. The categorical data to numerical IDs confuses the data set as they behave like an ordinal variable. It causes disturbance and inconsistency in the regression model(Kass, 1980).

Table 8 has shown some comparison of different regression models regarding their corresponding characteristics like Variance Inflation, R^2 value, and adjusted R^2 value. The originally planned model with NULLs removed explains 98.89% of the variance in the dependent variable by variance in the independent variable it shows overfitting in their regression model. Overfitting is unwanted in a regression model because it hinders the model to give precise and accurate results (Dietterich, n.d.). All the model shows the same R^2 and adjusted R^2 value. It shows that a different and

wide range of variables cannot impact the model so all the models are robust. Variation Inflation of each variable is close to optimal in your model. But our final model is superior to other models for 2 reasons. 1) Moderate R^2 value and 2) Minimum number of independent variables.

A moderate R^2 value of 0.6632 means they have not overfitted and under fitted like another following model. Moderate R2 value helps the dataset to train well and approximately accurate predictions. A minimum number of datasets will give the autorotational flexibility to execute the conclusive decision and help them to avoid overfitting the dataset. However, this model has some multicollinearity problems which may cause an error in the output. The model is not impeccable, it has some constraints and limitations but it may show and predict a better result in terms of CO_2 emission.

Discussion

Weakness and strengths of the study

From the dataset, it has worked with a dataset with N=43,177 with various models of vehicles manufactured by different companies. There is a large range of variables to get a clear picture of the entire population of the vehicle. Many results are conclusive for the government, environmental activists, and customers with environmental awareness. The results show some variables that are directly connected with the emission of CO_2. So, the dataset may help the authority to restrict the manufacturer and customer to change their decision to make less pollution in the environment.

But there are some major weaknesses of the dataset which have constrained its impact to solve the real world. Normal distribution in sample data is taken without any test and verification as a sample size below 30 can't follow the central limit theorem. The sale of the total car is 15 million in 2021. (New Auto Sales up in 2021, but Long Way Before Full Recovery, n.d.). but our dataset contains only 43,177 samples of data and these data are taken about 40 years. It is a very wide range of time and with a small range of the sample. Many of the sample vehicles are not available in the market and their mechanical tools, engines and all other deteriorate pieces of stuff became a long ago. So, it is hard to make an

impactful insight with the existing data. If it becomes expensive to make a statistic for a large population, then they need to focus to make the same size of samples in different states. Another limitation of the data is, some vehicles cannot count unrounded data, so there is a lot of data becomes unused for it. If "comb08" can be turned into continuous data from a discrete dataset, the problems may solve immediately. There are also some issues regarding the unit used to measure tailpipe CO_2 emission. Two different tests, the EPA tests, and the EPA emission factor are used to measure the same feature. It damages the consistency of the dataset. In the dataset, the weights of each variable are skewed, and hard to get a conclusion and which will result in underfitting when it will be applied to any prognosis model. Ex- there are Premium gasoline and regular gasoline fuel typed vehicles are counted for 12,801 and 28,733 but on the other hand, natural gas and electricity fuel typed vehicles are counted for 60 and 257. So, some variables are not equally dispersed and are skewed distributed. That's why if we could make a sample size to understand the impact of various variables, then it will be easy to avoid skewness, overfitting, and underfitting by arranging a selective dataset arrangement for modeling a perfect prognosis model.

Finding and interpretation:

Initial hypothesis was that, no correlation is intertwined between the nine-associate variables. These 9 variables are, annual primary-fuel petroleum consumption in barrels, vehicle type, vehicle manufacturer, make, primary fuel type, engine cylinders, combined luggage and passenger volume in cubic feet, , transmission type, and engine displacement. But the hypothesis is dismissed by their low p-value of <0.001 within 95% of confidence level.

Among these 9 variables, only three variables; engine displacement, engine cylinders, and combined volume are only used to make a regression model with a coefficient of 6.04887, 5.64859 and 1.18174. It maintains a R^2 value of .6632 which makes the constraint for overfitting and underfitting. Multicollinearity and categorical variables are another reason for reducing degree of freedom.

Increasing number of degrees of freedom can cause multicollinearity and reducing data can make an extra emphasis on the existing variable to get better prediction and analysis. Also reducing number of variables make the adjusted R^2 value and R^2 value same.

References

.

Global transport CO2 emissions breakdown 2020 | Statista. (n.d.). Retrieved September 10, 2022, from https://www.statista.com/statistics/1185535/transport-carbon-dioxide-emissions-breakdown/

Dietterich, T. (n.d.). *Overfitting and Undercomputing in Machine Learning*.

Kass, G. v. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *29*(2), 119–127. https://doi.org/10.2307/2986296

New Auto Sales up in 2021, but Long Way Before Full Recovery. (n.d.). Retrieved October 17, 2022, from https://www.usnews.com/news/business/articles/2022-01-04/new-auto-sales-up-in-2021-but-long-way-before-full-recovery

Plackett, R. L. (1983). Karl Pearson and the Chi-Squared Test. *International Statistical Review / Revue Internationale de Statistique*, *51*(1), 59. https://doi.org/10.2307/1402731

Zhang, L., Long, R., Chen, H., & Geng, J. (2019). A review of China's road traffic carbon emissions. *Journal of Cleaner Production*, *207*, 569–581. https://doi.org/10.1016/J.JCLEPRO.2018.10.003

Figure 1

Histogram of CO2 Emissions

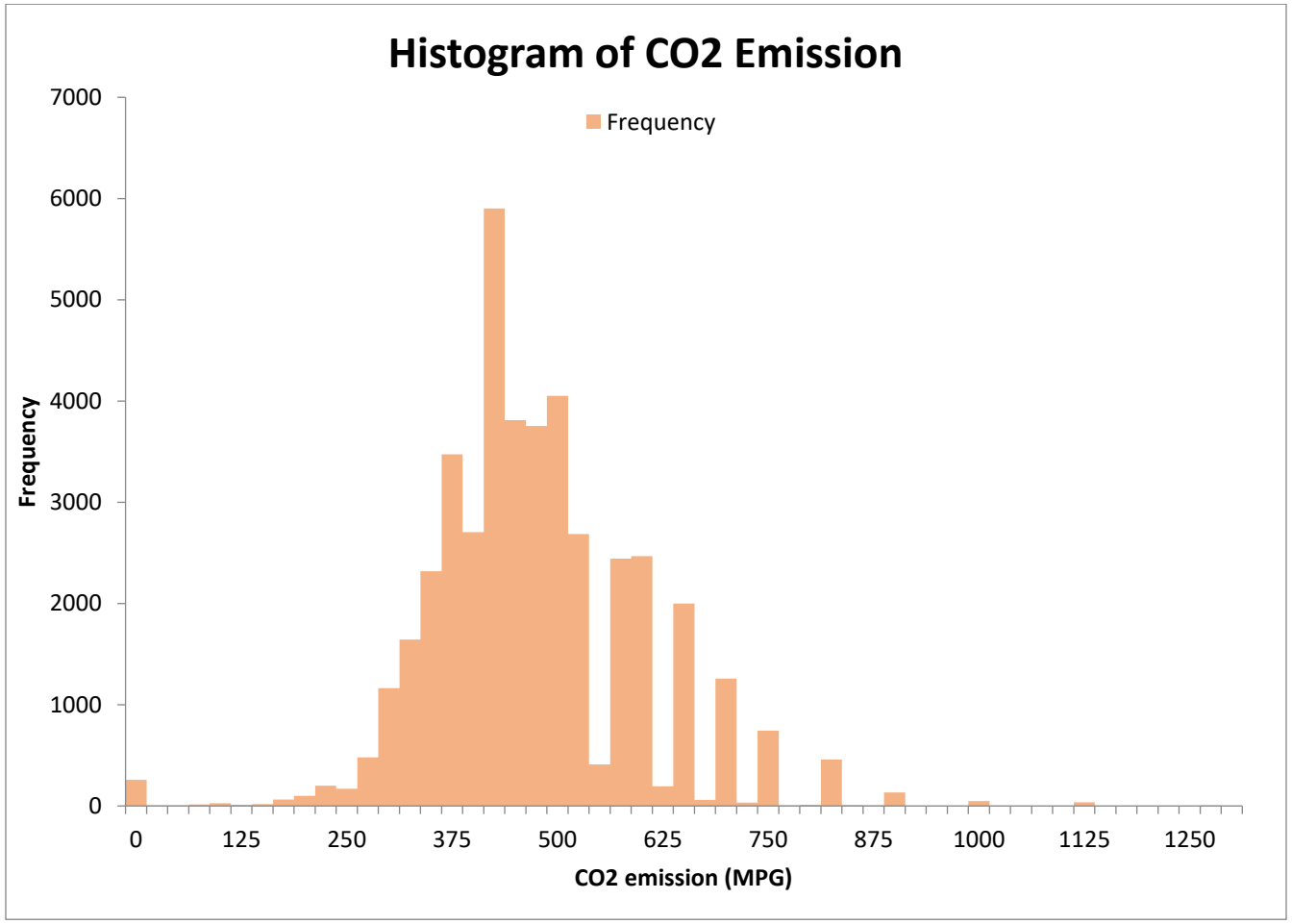

Figure 2

Histogram of Annual petroleum consumption

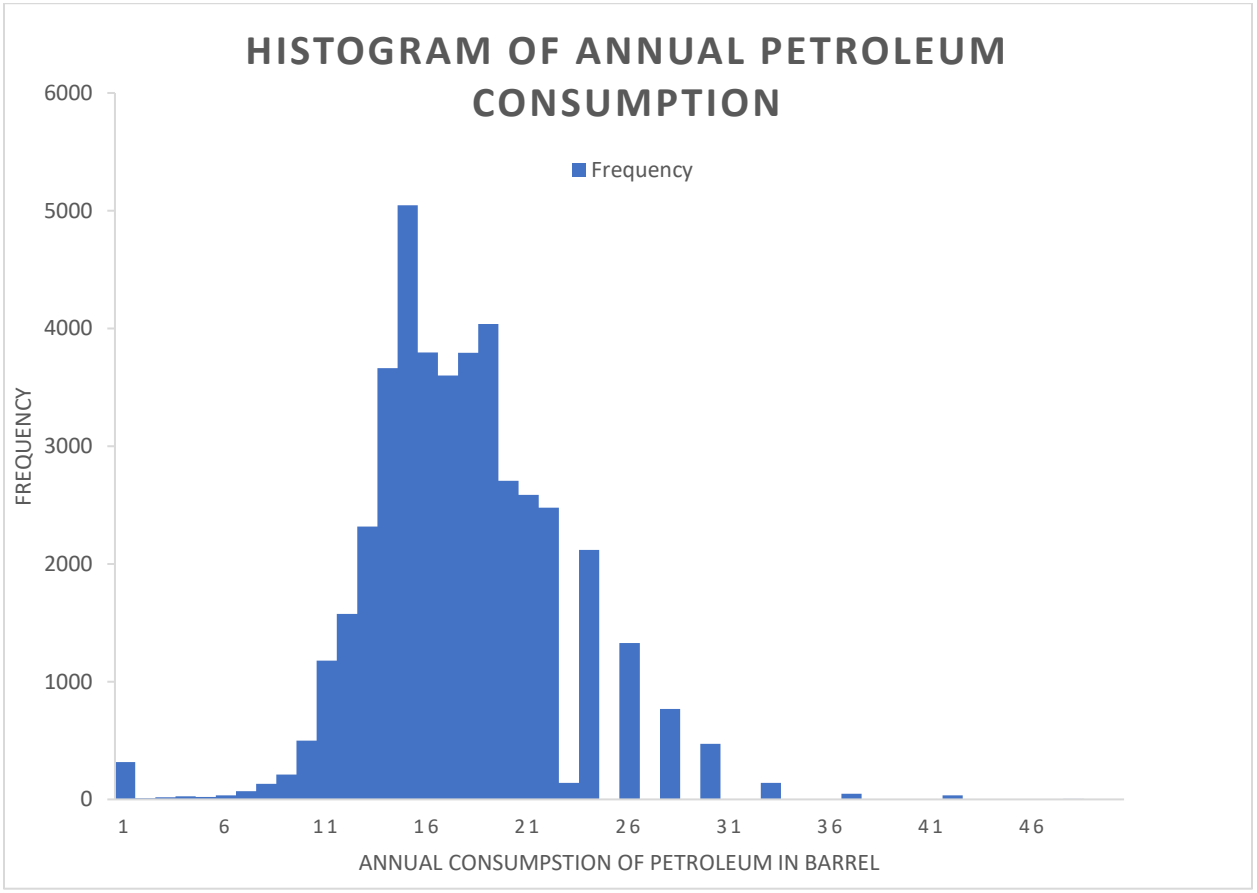

Figure 3

Box plot CO2 emission vs engine cylinder

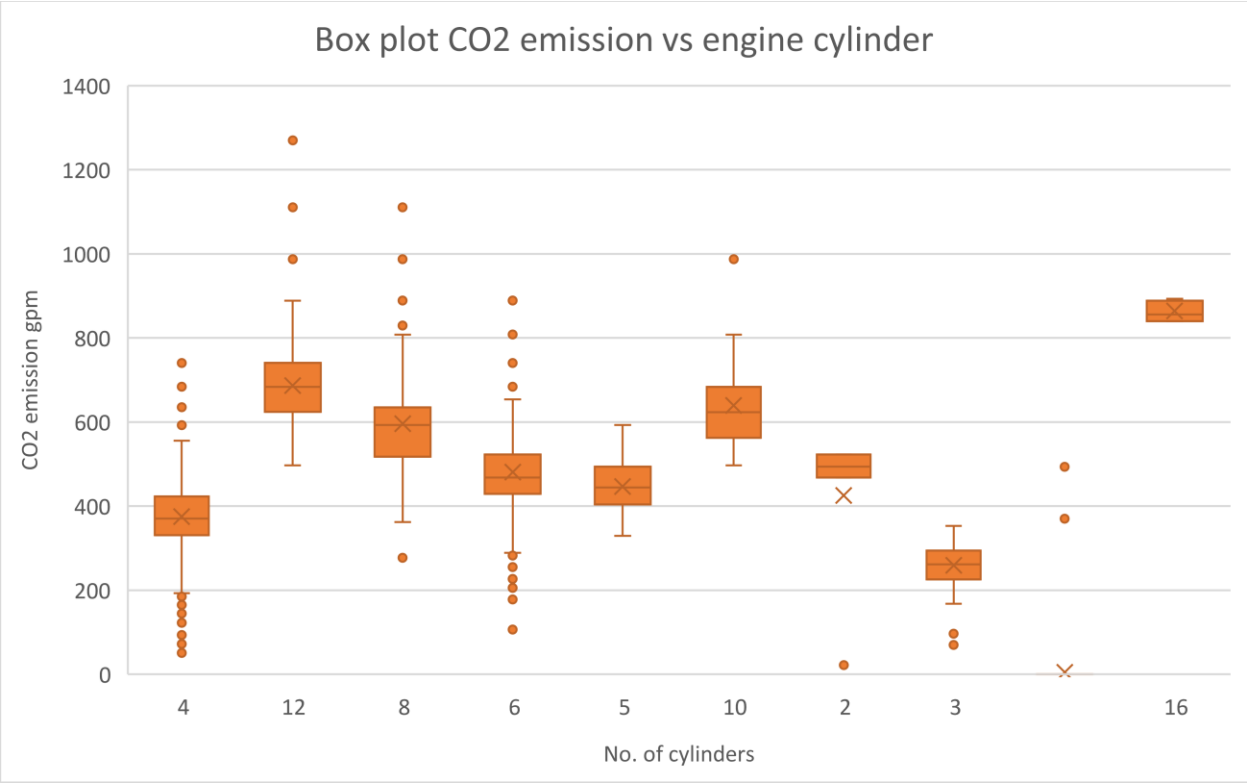

Figure 4

Boxplot annual consumption of petroleum vs engine cylinder

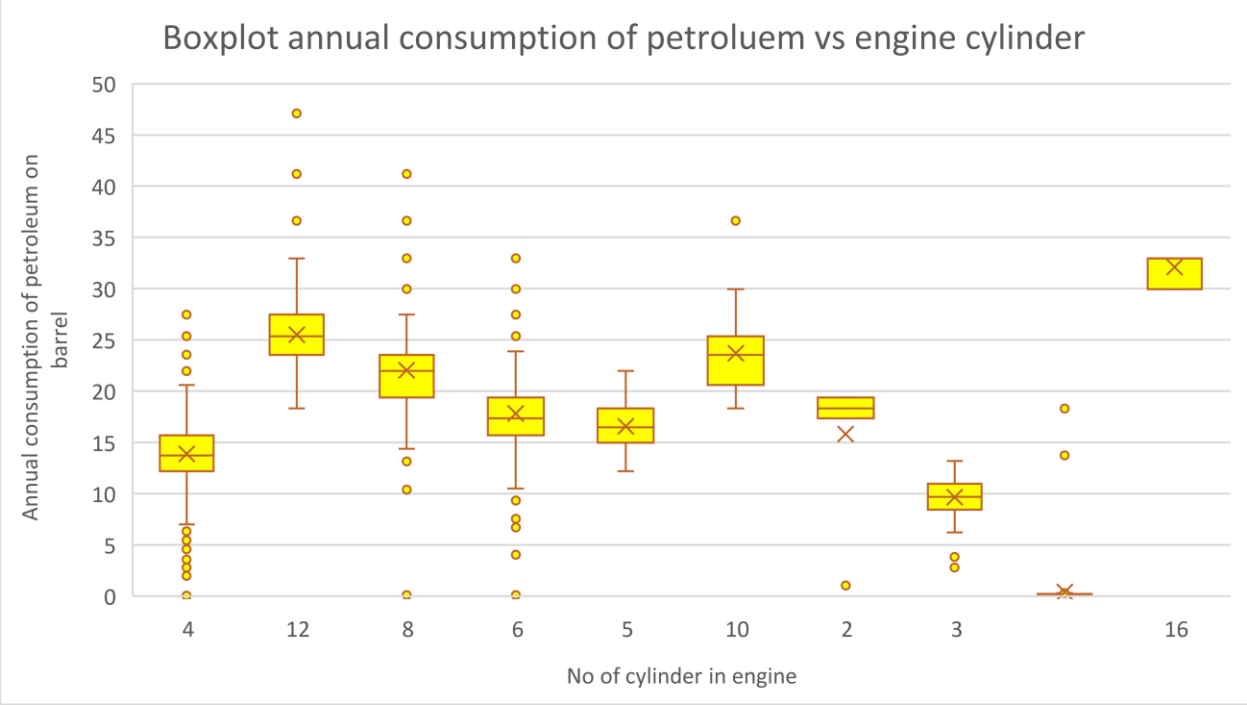

Figure 5

Boxplot of CO_2 emission vs different groups of fuel types

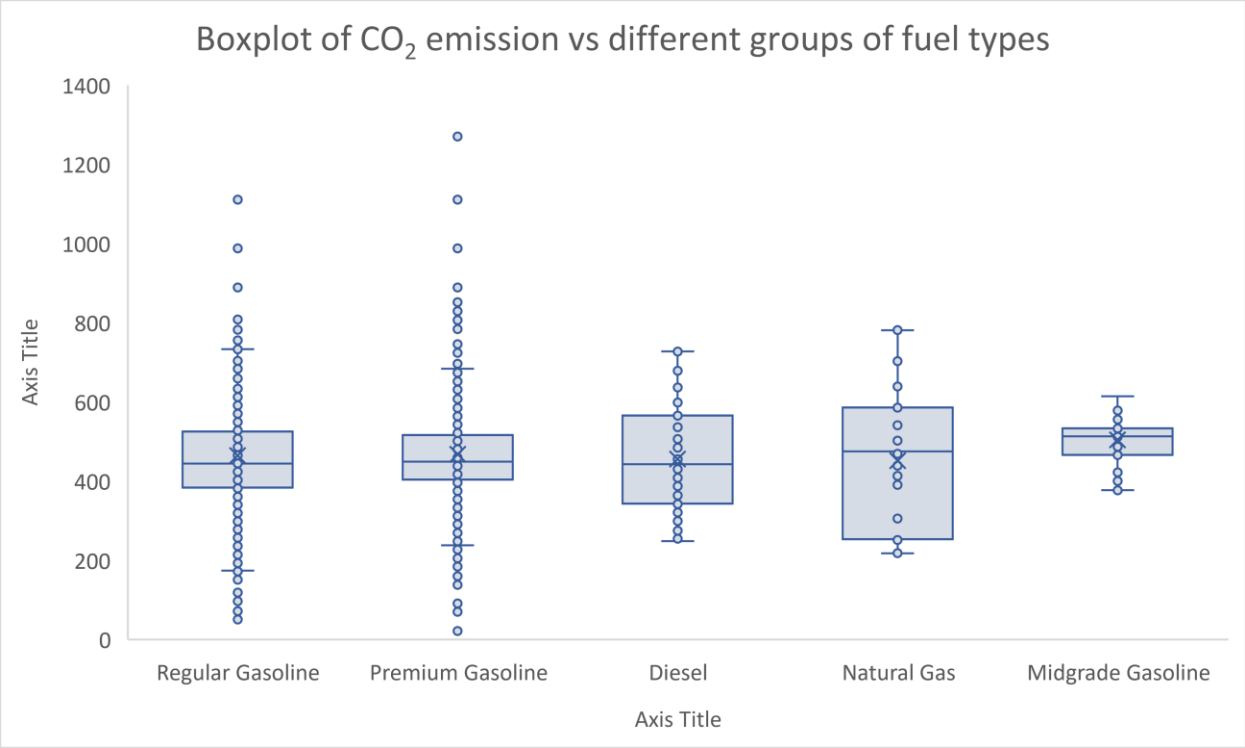

Figure 6

Histogram of Normal distribution of CO_2 emission

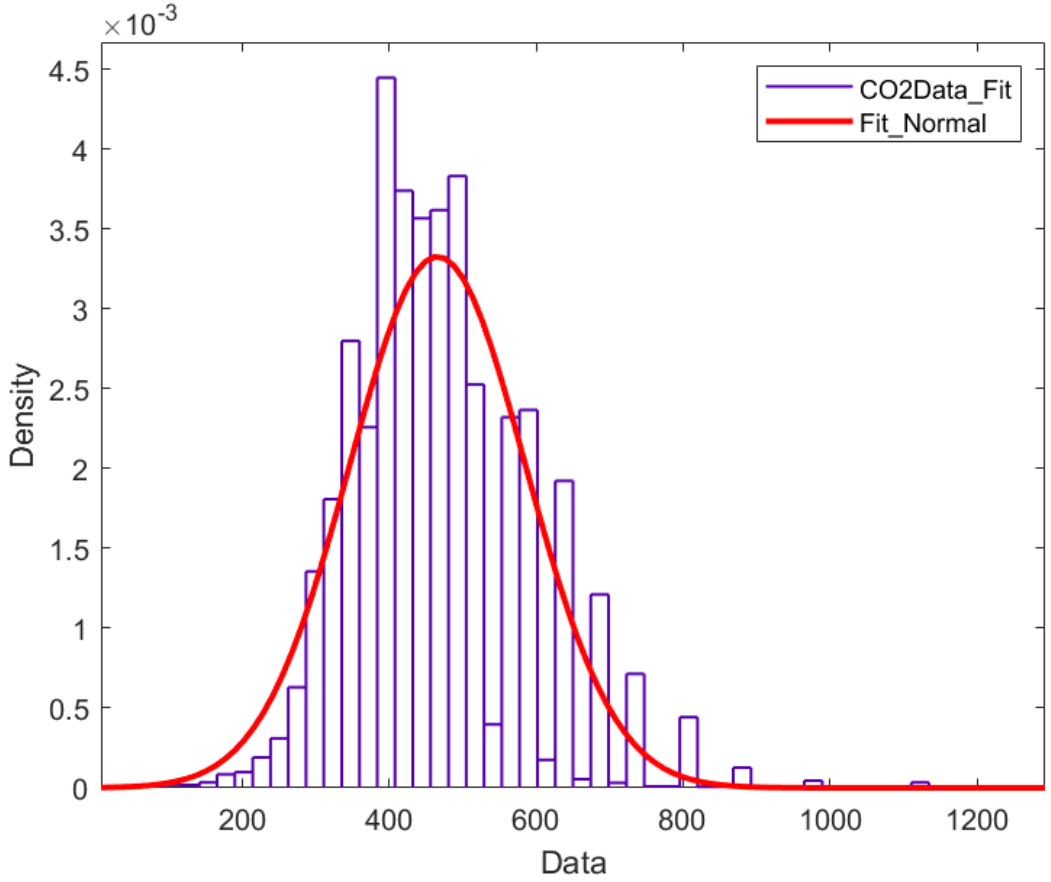

Figure 7

Column chart of 43,177 sample vehicles motor by primary fuel type

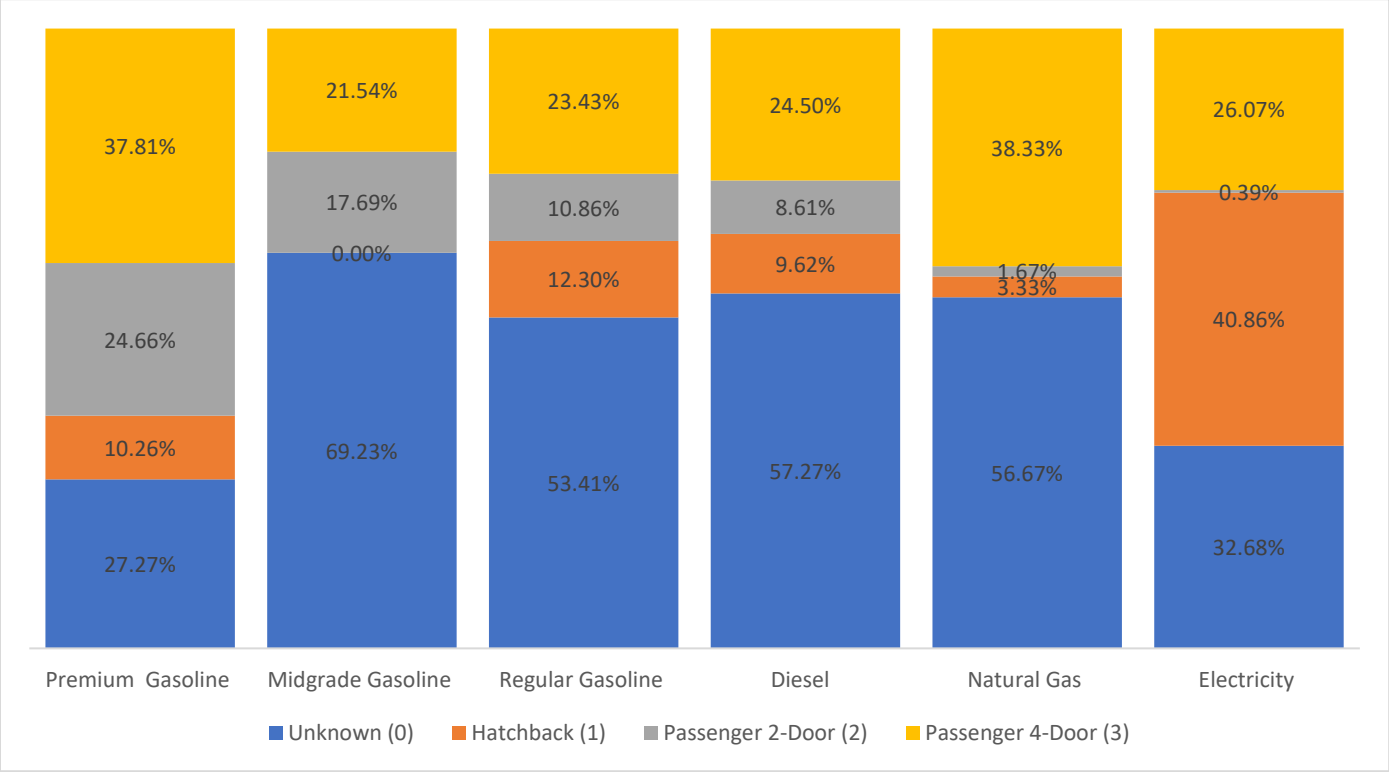

Table 1

Descriptive statistics analysis

	barrels08	**co2TailpipeGpm**	**displ**	**Volume**
Median	16.48	444.36	2.8	86
Sum	740622.26	19980883.99	20396.1	2889792
Mean	17.15	462.77	3.24	66.93
Standard Deviation	4.66	124.77	1.42	69.04
Sample variance	21.74	15568.06	2.02	4766.75
Skewness	0.37	0.42	0.61	0.65

Table 2

Change of Descriptive Statistics without Blank/null values

	barrels08	Barrel08 (No-0)	Change (%)	co2TailpipeGpm	co2TailpipeGpm (No-0)	Change (%)	displ	displ (No-0)	Change (%)	Volume	Volume (No-0)	Change (%)
Median	16.48	16.4805	0.00%	444.36	444.35	0.00%	2.8	3	7.14%	86	86	0.00%
Sum	740622.26	740566.15	-0.01%	19980883.99	19980884	0.00%	20396.1	69338.5	239.96%	2889792	2870500	-0.67%
Mean	17.15	17.25457	0.61%	462.77	465.5378	0.60%	3.24	3.22249849	-0.54%	66.93	66.88024231	-0.07%
Standard Deviation	4.66	4.4883303	-3.68%	124.77	119.8801	-3.92%	1.42	1.388263464	-2.23%	69.04	69.12331331	0.12%
Sample variance	21.74	20.144639	-7.34%	15568.06	14371.25	-7.69%	2.02	1.927185875	-4.59%	4766.75	4778.032444	0.24%
Skewness	0.37	0.6699891	81.08%	0.42	0.748666	78.25%	0.61	0.721131387	18.22%	0.65	0.651778137	0.27%

Table 3

Bivariate Frequency table of 43,177 Sample Vehicle Models by Primary Fuel Type

	Population		Premium Gasoline		Midgrade Gasoline		Regular Gasoline		Diesel		Natural Gas		Electricity		
	***N*(%)**		***n*(%)**		***n*(%)**		***n*(%)**		***n*(%)**		***n*(%)**		***n*(%)**		
Variable	**(N=43,177)**		**(n=12,801)**		**(n=130)**		**(n=28,733)**		**(n=1,196)**		**(n=60)**		**(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

	Population		**Ultra-Low Emission**		**Very-Low Emission**		**Low Emission**		**Standard**		**Polluter**		**Gross Polluter**		
	***N*(%)**		***n*(%)**		***n*(%)**		***n*(%)**		***n*(%)**		***n*(%)**		***n*(%)**		
Variable	**(N=43,177)**		**(n=321)**		**(n=384)**		**(n=5,556)**		**(n=29,543)**		**(n=5,899)**		**(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 5

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	-0.9184
prifueltype	-.1128	-.1084	.1234	.0710	-.2149	-.2181	.0498	-.0340	-.0874	1.0000	0.9885

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

Table 6

Contingency table between emission category and vehicle type:

		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
	Row Pct	0.07	9.57	12.21	78.15	-
	Col Pct	0.02	2.21	1.50	5.84	-
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
	Row Pct	32.76	12.65	41.29	13.30	-
	Col Pct	35.90	10.99	19.14	3.75	-
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
	Row Pct	0.80	5.75	6.68	86.78	-
	Col Pct	0.93	5.30	3.29	25.95	-
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
	Row Pct	9.99	17.58	29.85	42.58	-
	Col Pct	58.22	81.22	73.60	63.76	-
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
	Row Pct	38.01	2.18	31.46	28.35	-
	Col Pct	2.41	0.11	0.84	0.46	-
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
	Row Pct	33.33	2.86	50.78	13.02	-
	Col Pct	2.52	0.17	1.63	0.25	-
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

Table 7

Chi-Square calculation process

Observed Value	Hatchback	Passenger 2-Door	Passenger 4-Door	Unknown	Total
Gross polluter	1	141	180	1152	1474
Low emission	1820	703	2294	739	5556
Polluter	47	339	394	5119	5899
standard	2952	5193	8819	12579	29543
ultra-low emission	122	7	101	91	321
very-low-emission	128	11	195	50	384
Total	5070	6394	11983	19730	43177
Expected Value	Hatchback	Passenger 2-Door	Passenger 4-Door	Unknown	Total
Gross polluter	173.08	218.28	409.08	673.55	1473.99
Low emission	652.41	822.78	1542	2538.8	5555.99
Polluter	692.68	873.57	1637.2	2695.6	5899.05
standard	3469	4375	8199.1	13500	29543.10
ultra-low emission	37.693	47.536	89.088	146.68	321.00
very-low-emission	45.091	56.866	106.57	175.47	384.00
Total	5069.95	6394.03	11983.04	19730.10	43177
Chi- square contribution	Hatchback	Passenger 2-Door	Passenger 4-Door	Unknown	Total
Gross polluter	171.09	27.36	128.28	339.86	666.59
Low emission	2089.59	17.44	366.73	1275.91	3749.67
Polluter	601.87	327.12	944.02	2178.69	4051.70
standard	77.05	152.94	46.87	62.83	339.69
ultra-low emission	188.57	34.57	1.59	21.14	245.86
very-low-emission	152.45	36.99	73.38	89.72	352.53
p-value	<.001	Degree of Freedom	15	Chi-Square Value	9406.05

Table 8

Comparison of different regression model

Regression models	Displ	Cylinders	volume				
Trimmed reg model-LV	6.04887	5.64859	1.18174	R-sqr	0.6632	Adj R-sqr	0.6332
Trimmed reg model-LV (0 Null)	5.97938	6.00619	1.02134	R-sqr	0.5902	Adj R-sqr	0.5902
LR-org. planned model	7.29952	6.47947	2.44999	R-sqr	0.9829	Adj R-sqr	0.9828
LR-org. planned model (0-null)	6.86506	6.4264	1.16937	R-sqr	0.9889	Adj R-sqr	0.9889
trimmed reg mod- removed cat.	6.88532	5.76127	1.27869	R-sqr	0.9812	Adj R-sqr	0.9812
trimmed reg mod-removed cat (0 Null)	6.73623	6.11311	1.02167	R-sqr	0.9886	Adj R-sqr	0.9886

Peer Feedback Form

Reviewer:	Itzel Cruz	Date Paper Received	10/28/2022
Peer:	S M Sultan Mahmud Rahat	Date Feedback Provided:	10/19/2022

Criteria	Example of "Meets or Exceeds" Standards	Comments and/or Recommendation	Reference: Related area, page number, paragraph, or editorial remarks in paper
APA-7 elements, formats, citations, references, and structure. 20%	☐ Submission follows all requirements of the APA Publication Manual 7th for a professional work ☐ fewer than three individual and minor errors, inclusive of professional grammar and writing style. As noted in the assignment instructions,	• One citation in the Introduction paragraph needs to be fixed. • Remove extra space	• Page 4 • Page 7 • Page 17

Criteria	Example of "Meets or Exceeds" Standards	Comments and/or Recommendation	Reference: Related area, page number, paragraph, or editorial remarks in paper
	☐ the required number of quality and relevant references (six) are used.	before Visualization and explanation.. • Needs at least 6 references – only see 3 on paper • Missing list of equations.	
Abstract: Brief, comprehensive, summary of the contents of the paper that is accurate, non-evaluative, coherent and readable. 5%	☐ Submission clearly and concisely presents the problem under investigation, ☐ data sources and pertinent characteristics, essential feature of the study methodology, basic findings, and discussion, and conclusion. ☐ All requirements as listed in the assignment instructions are included. ☐ Statistical terminology and mathematical symbols and formulae are used appropriately and show a high-level of professional understanding.	Not included yet.	
Introduction: Includes problem statement, hypothesis, aims, and objective, in a compelling manner 5%	☐ Submission clearly and concisely presents the introduction with context, introduction with a professional and ☐ statistically-appropriate level of understanding. All requirements as listed in the assignment instructions are included. ☐ Statistical terminology and mathematical symbols and formulae are used appropriately and show a high-level of professional understanding.	• Background and hypothesis paragraph should be split into two. • Generalized Object Formula paragraph contains all requirements listed in the assignment.	• Page 4 • Page 4
Method: Outlines data collection and instrumentation, data characteristics, procedures, and measures and covariates, associations, data diagnostics, and analytic strategy 25%	☐ Submission succinctly and with appropriate detail describes the methodology outlining data creation, descriptive statistics, associations, and characteristics. ☐ All requirements as listed in the assignment instructions are included. ☐ Statistical terminology and mathematical symbols and formulae are used appropriately and show a high-level of professional understanding.	• Missing data pre-processing paragraph • Statistical terminology are used appropriately and great use of graphs to explain the key concepts.	• Page 6 • Page 6
Results: Includes statistics and data	☐ Submission clearly, concisely and with appropriate detail describes the results of the	• Number 2nd equation to (2) • Include that "Categorical	• Page 10 • Page 14

Criteria	Example of "Meets or Exceeds" Standards	Comments and/or Recommendation	Reference: Related area, page number, paragraph, or editorial remarks in paper
analysis, inclusive of information detailing the statistical and data-analytic methods used, inferential statistics, and complex data analyses 25%	analysis. All requirements as listed in the assignment instructions are included. ☐ Statistical terminology and mathematical symbols and formulae are used appropriately and show a high-level of professional understanding.	variables have not been converted to c-1 indicator variables" (from office hours) • Rest of section has a clear and precise explanation	
Discussion: Reports strengths and problems related to data and/or statistical assumptions, support of original hypothesis, interpretation, generalizability, and implications 15%	☐ Submission clearly and concisely presents conclusions, recommendations, strengths, and limitations. All requirements as listed in the assignment instructions are included. ☐ Statistical terminology and mathematical symbols and formulae are used appropriately and show a high-level of professional understanding.	• Submission clearly and concisely presents conclusions, recommendations, strengths, and limitations. All requirements as listed in the assignment instructions are included.	