**Predictive Modeling of Tailpipe CO2 Emissions for Early Vehicle Design**

Sheshma Jaganathan

Manufacturer of XYZ

June 23, 2025

**Abstract**

The objective of this analysis is to statistically investigate the association of primary fuel tailpipe carbon dioxide (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. As an automobile manufacturer, our objective is to develop a predictive model for evaluating tailpipe CO2 emissions during the early engineering design phase. This model will help to improve the efficiency of our product development by allowing us to evaluate the emissions impact of key design decisions. The hypothesis is that CO2 emissions can be significantly predicted by vehicle characteristics like petroleum consumption, engine displacement and number of cylinders. A secondary data analysis was conducted on a dataset of 43177 vehicles obtained from the U.S. Environmental Protection Agency (EPA) through FuelEconomy.gov with 138 different manufacturers. After cleaning data and addressing multicollinearity, the final linear regression model included only the most statistically significant predictors like annual fuel consumption, combined miles per gallon (MPG), engine displacement, number of cylinders and volume for CO2 tailpipe emissions. All these variables were found to be statistically significant with a p-value less than 0.0001, indicating strong evidence against the null hypothesis. Moreover, an of 0.98 indicates how well our model explains the variations in CO2 emissions. Overall, the predictors have a narrow 95 % confidence interval confirming the strong relationship between predictors and CO2 tailpipe emission. These statistical significance supports our predictive model and can be confidently used during the vehicle engineering design phase to estimate and reduce emissions to build a cleaner and more efficient vehicle for the future.

**Contents**

Abstract………………………………………………………………………………………………………………………………………………2

List of Tables……………………………………………………………………………………………………………………………………….4

List of Equations………………………………………………………………………………………………………………………………….5

List of Figures………………………………………………………………………………………………………………………………………6

Introduction………………………………………………………………………………………………………………………………………..7

Generalized Object Formula……………………………………………………………………………………………………7

Method……………………………………………………………………………………………………………………………………………….8

Data Collection And Pre-processing…………………………………………………………………………………………8

Population Characteristics……………………………………………………………………………………………………….9

Descriptive Statistics………………………………………………………………………………………………………………10

Bivariate Frequency Analysis……………………………………………………………………………………………….…11

Association of Emission Category by Fuel Type and Other Characteristics………………………………12

Results…………………………………………………………………………………………………………………………………………………13

Probability Density Analysis……………………………………………………………………………………………………13

Pearson Correlation Coefficient Analysis…………………………………………………………………………………14

Chi-Square Test for Homogeneity and Independence……………………………………………………………….15

Issue of Multicollinearity and Overfitting…………………………………………………………………………………16

Initial Regression Model…………………………………………………………………………………………………………..17

Justification of Final Regression Model……………………………………………………………………………………18

Final Regression Model……………………………………………………………………………………………………………19

Discussions……………………………………………………………………………………………………………………………………………20

Conclusion………………………………………………………………………………………………………………………………20

Strengths and Weaknesses of the Study……………………………………………………………………………………21

References……………………………………………………………………………………………………………………………………………23

**List Of Tables**

Table 1 Descriptive Statistical Analysis ……………………………………………………………………………………

Table 2 Characteristics of 43177 Vehicle Models by Primary Fuel Type…………………………………………..12

Table 3 Association of Emission Category by Fuel Type and Other Characteristics……………………………16

Table 4 Pearson Correlation Coefficients (N=43,177)………………………………………………………………………18

Table 5 Chi-Square Table for Emission Category and Vehicle Type……………………………………………………20

Table 6 Two-way Contingency Table for Emission Category and Vehicle Type………………………………21

**List Of Equations**

Equation 1 Generalized Object Formula……………………………………………………………………………………

Equation 2 Probability Density Function of CO2 Emissions………………………………………………………..

Equation 3 Initial regression Model…………………………………………………………………………….

Equation 4 Final Regression Model…………………………………………………………………………………………….

**List Of Figures**

Figure 1: Boxplot of CO2 Tailpipe Emission(gm/mile)…………………………………………………………

Figure 2: Probability Density Function of CO2 Emission in Grams per Mile……………………..

Figure 3: Positive correlation between CO2Tailpipe and barrels08…………………………………………………

Figure 4: Negative correlation between CO2Tailpipe and comb08…………………………………………………

**Introduction**

A growing concern for our global environment is the increasing levels of CO2 and fuel consumption of automobiles. The transportation sector is one of the major sources of greenhouse gas emissions, contributing up to 28 % of total emissions in the US only (Environmental Protection Agency, 2023). There is a need for more stricter emission standards in order to reduce pollution and meet the global climate goals. A better understanding of the link between automobile fuel consumption and CO2 emissions from tailpipes would provide useful insights for policymakers to make regulations, to help consumers to make better choices, and manufacturers to understand more about their designs to eliminate further negative environmental impacts.

The study addresses how various predictors can influence the tailpipe CO2 emissions and fuel consumption of vehicles. The study analyzed dataset with variables including vehicle manufacturer, make, engine displacement, combined miles-per-gallon for the primary fuel type, engine cylinders, combined luggage and passenger volume in cubic feet, vehicle type, transmission type, and primary fuel type to understand the relationship between tailpipe CO2 emissions per mile and the annual petroleum consumption. The purpose of the study is to build a prediction model to estimate tailpipe CO2 emissions in the early engineering design stage of vehicle development. For secondary data analysis, we are testing the hypothesis using multiple linear regression. The null hypothesis is that none of the vehicle characteristics has a statistically significant impact on tailpipe CO2 emissions and the parameter estimates are essentially zero. In contrast, the alternative hypothesis is that at least one of these variables can influence emissions and its parameter estimates are significantly different from zero. If the variables like petroleum consumption, engine displacement and MPG are statistically significant( p-values <0.001), we reject the null hypothesis. This confirms that these predictors have a significant role in evaluating emissions and should be considered for future design of vehicles. This model can be used as a decision making tool during the vehicle design phase to improves design and efficiency and reduces emissions for a cleaner environment.

**Generalized Object Formula**

To understand how different variables can affect the tailpipe CO2 emission, we use a generalized objective formula using multiple linear regression. The helps us to determine how each variable contributes to CO2 emissions and thus make predictions that can be used in the design of more efficient vehicles. This formula helps us to see how multiple variables like fuel consumption(barrels08), combined fuel economy(comb08), number of cylinders, vehicle manufacturer(make\_id), engine displacement(displ), fuel type, transmission type can affect the CO2 Emission. The regression model is given as in Equation 1

CO2=++++++++ (1)

Where,

is the intercept.

through are the regression coefficients of each variable.

is the error term associated with each regression coefficient (epsilon factor)

This regression model supports the hypothesis that the vehicle characteristics can significantly affect the tailpipe CO2 emission.

**Method**

**Data Collection and Pre-processing**

The data was downloaded for the study from the FuelEconomy.gov website, which provides fuel economy information for U.S. vehicle models from year 1984 to 2020 (U.S. Department of Energy,2020). The raw data was in CSV format, which was opened in Microsoft Excel, where it was converted into a table format. The finalized Excel table was imported into MATLAB for distribution curve fitting and then imported into SAS for statistical analysis. The system used for research was a LENOVO Think Station P330 Tower with an Intel Xeon E-2186 6 Core Processor with vPro,64 GB DDR 2666MHz ECC memory, RAID 1, 1.0 Terabyte M.2 PCIe Opal SSD storage. The system was having the operating system Windows 10 Pro for Workstations. The software used in the analysis included MATLAB R2020b for predictive modelling and distribution curve fitting, MS Excel, SAS for statistical analysis, and Tableau 2020.3 for data visualization and exploration.

**Population Characteristics**

The data collected involves a study of 43177 vehicles with 138 different manufacturers. Inorder to study the link between Tailpipe CO₂ emissions in grams/mile(co2TailpipeGpm) and Annual barrels of petroleum used(barrels08), a total of 54 variables were evaluated among 138 different manufacturers that vary in specifications like number of Cylinders (cylinders) ,fuel Costs (fuelCost08) ,City MPG (cityE, cityUF, cityCD),Highway MPG (highway08, highwayCD, highwayE, highwayUF) ,Combined MPG (combE, combinedCD, combinedUF),Electric Range (range, rangeCity, rangeHwy),youSaveSpend – Distribution of cost savings or overspending compared to average.

The categorical variables estimated involved Fuel Type (fuelType, fuelType1) , Drive Type (drive) – AWD, FWD, RWD counts, Vehicle Class (VClass) – SUVs, sedans, etc, Transmission Type (transtype) – Manual vs automatic, Vehicle Type (vehtype\_name) – Electric, hybrid, conventional, Make (make)-Number of models per manufacturer, Year (year) – Distribution of vehicles over time.

**Descriptive statistics**

By statistically analyzing the relationship between the variables, we can gain insights into how fuel efficiency and other vehicle characteristics affect overall CO2 emissions. Table 1 gives a summary of descriptive statistics for key variables of the vehicle dataset, including metrics like CO2 emission (grams per mile), annual barrel consumption(barrels08), combined MPG(comb08), engine displacement in liters (displ) and volume, for which the total data count, mean, standard deviation , minimum and maximum values are provided. On average, vehicles emit approximately 462.77 grams of CO2 per mile with a standard deviation of 124.77. The emission range is from 0 to 1269.57 grams per mile. Average fuel consumption is around 17.15 barrels annually, with a standard deviation of 4.66. Combined fuel economy averages about 20.85 miles per gallon (MPG), with high skewness indicating the presence of outliers. The engine displacement shows an average of 3.29 liters and vehicle volume has a mean of 66.93 and a high standard deviation of 69.041, indicating variation among different vehicles. The zero value for engine displacement and volume in Table 1 indicates null values or missing values in the dataset.

Figure 1 shows the boxplot for CO2 tailpipe emission (grams per mile) and number of cylinders, indicating the distribution of emissions within each cylinder category. Dataset contains cylinders ranging from 2 to 16. The 2-cylinder has not been included due to less amount of data. Vehicles with 3-cylinders and 4-cylinders show low CO2 emissions with an average of 259.33 and 374.66 g/mi, respectively. The vehicles with less cylinders are usually compact or hybrid, and they emit a low level of CO2 compared to vehicles with more cylinders. The 4-cylinder has a lower whisker ranging from 261.38 g/mi to a higher whisker up to 555.43 g/mi. The mean is greater than the median, showing a right-skewed distribution. There are outliers present on both ends of the whiskers indicating variability of emission, as most 4-cylinder vehicles have lower to moderate emissions, with a few higher emissions in the vehicles with 4-cylinder. The 5-cylinder vehicle shows less variability in CO2 emissions with an average of 446.49 g/mi.

In contrast, the vehicles with 8 or more cylinders have a higher mean, indicating high fuel consumption.

As the cylinder count increases, the interquartile range (IQR) of the boxplot also becomes wider, indicating more variability in CO2 emissions. The presence of outliers on the higher side of the whisker shows larger variability within the group and can skew the data to right. We can conclude that the greater the number of cylinders, more is the CO2 emission. We must focus on developing a more fuel-efficient vehicle by optimizing the vehicle's engine design in order to reduce CO2 emissions.

**Bivariate frequency analysis**

The bivariate frequency in Table 2 shows a statistically significant association between vehicle characteristics like vehicle type and fuel type *(p* < 0.001). Each column from Table 2 adds to 100%, allowing for the comparison of how each subgroup contributes to the overall outcome. The total population size is 43,177 vehicles, with the independent variable categorized by vehicle type-Unknown, hatchback, passenger 2-door, and passenger 4-door. Also, the variable transmission types include automatic and manual. The dependent variable consists of different fuel types- Midrange gasoline, regular gasoline, diesel, Natural gas, and electricity.

In vehicle type, almost half (45.7%) of all vehicles are of Unknown types. Regular gasoline is widely used, with 53.4% of the total, and diesel is even more common at 57.3%. Their use of mid-range gasoline is also quite high (69.2 %).

Hatchbacks make up only 11.7% of the population of vehicle types, yet they consume 12.3% of regular gasoline and 9.6% of diesel in comparison to their other vehicle types. However, their use of electricity is quite high at 40.9%, which exceeds their population share. This suggests that small vehicles like hatchbacks, with electric ones, are the top common choice of vehicle. Diesel is consumed to a greater extent by large vehicles like trucks, while electricity or standard gasoline is consumed by smaller vehicles like hatchbacks. This means that different vehicles use different fuel, and this coincidence is not just random but is strongly related to each other.

Passenger 2-door vehicles make up about 14.8% of all vehicle types, but they consume 24.7% of all premium gasoline, more than expected. On the other hand, they consume only 8.6% of diesel and 0.4 % of electricity, which strongly suggests these vehicles aren’t fuel efficient and are built for speed and style.

In the transmission type, out of 43177 vehicles, about 70% (30,210) have automatic transmission, and the remaining 30 % (12,956) the manual transmission. Vehicles with automatic transmissions mostly use mid-range gasoline, natural gas, or electricity (100%). They make up a majority of premium gasoline users (73.5 %). On the other hand, manual transmission vehicles, which are less common nowadays, mostly use diesel (35.4%), and regular gasoline vehicles in manual transmission use gasoline, natural gas, and electricity. The p-value for all these proportions is low ( *p* < 0.001), which indicates a strong association between transmission type and fuel choice, representing the difference between our proportions is statistically significant at 95 % confidence interval.

**Association of Emissions Category by Fuel Type and Other Characteristics**

Table 3 shows the association of Emission categories by fuel type, Vehicle Type, and Transmission type. The emissions are categorized into Ultra low, Very low, Low standard, Polluter, and Gross Polluter. Out of the dataset of 43,177 vehicles, the standard emission group makes up 68.42% (*n*=29,543) of the population. On the other hand, a total of 6261 vehicles (14.5%) make up ultra low, low, and very low categories, showing low emissions. Regular gasoline vehicles, which are 66.5% of the total population, fall to 67% of gross polluters and 73.9% polluter group. Diesel vehicles, which make up 2.8 % of the population, account for 4.4 % of polluters and 5.5 % of low emissions. A study on on-road CO2 emissions found that gasoline and diesel vehicles are the main contributors of CO2 emissions (Yan et al., 2024). Though electric vehicles make up just 0.6% of the population, it make up 80.1% of the ultra-low emission group. Hence, encouraging the use of electric vehicles can be the cleanest fuel option for our environment.

In vehicle types, Unknown vehicles make up 45.7 % of the population, contribute to 86.8 % of polluters, and 78.2 % of gross polluters. In contrast, small vehicles like hatchbacks, which make up 11.7% of the overall vehicle population, have ultra-low and low emissions of 38% and 33.3%, respectively. Hence, the use of small vehicles can significantly reduce CO2 emissions. For Passenger 4 doors, which accounts for 27.8 % of the population, has 31.5 % ultra low emission and only 12.2% gross polluter.

In transmission type, automatic dominates with 70 % of the total population, with only 30% in manual. On average, automatic vehicles emit more CO2 compared to the manual vehicles. These associations of CO2 emission between fuel type, vehicle type, and transmission type are statistically significant (*p <* 0.0001), suggesting that the observed differences are not likely due to chance but rather represent significant associations.

**Results**

**Probability Density Analysis**

The probability density function (PDF) of CO2 Tailpipe emissions is important in understanding how well the normal distribution curve matches with the actual distribution of the data. The shape of a bar graph in figure 2 is similar to the bell-shaped curve of the normal distribution. The mean for the normal distribution curve is 465.538 g/mile. This represents the peak of the curve of PDF is unimodal. The standard deviation of 119.88 shows the spread, which is the amount of variability in the CO2 emission data. A larger standard deviation indicates a wider spread of CO2 emission values from the mean. The data is normally distributed with a slight right skew (tail to the right), which means that there are more vehicles with higher emissions. The Probability density function (PDF) for CO2 emission, assuming a normal distribution, is given by as follows (Devore, 2016)

= (2)

= the probability density function of CO2 emission

=mean

= standard deviation

**Pearson Correlation Coefficient Analysis**

Correlation describes the strength and direction of the linear relationship between two variables, ranging from -1 to +1. A correlation value approaching +1 demonstrates that both variables increase together, showing strong positive relationship. When the correlation value approaches -1, it shows a strong negative relationship between variables, so that one variable's increase leads to the other's decrease. When the correlation approaches zero, it demonstrates there is minimal or no relationship between two variables, according to Bruce & Bruce (2017). As a general rule, correlations are considered weak if ∣r∣<0.5, moderate if between 0.5 and 0.8, and strong if above 0.8 (Devore, 2016, p.529). Table 4 shows the Pearson correlation coefficient between different predictors to understand the relationship between variables.

There is a strong positive correlation of 0.9885 between CO2 emission (CO2 tailpipe) and fuel consumption(barrels08). Figure 3 shows a positive correlation between CO2 emissions and fuel consumption. This means, vehicles that consume more fuel emit more CO2. Similarly, there is a correlation of 0.9046 between cylinders and engine size(displ). This strong positive correlation indicates that the larger the size of the engine, the greater the number of cylinders.

As seen in Figure 4, there is a strong negative correlation of -0.9184 between CO2 emission (co2Tailpipe) and miles per gallon(comb08), which shows vehicles with better fuel efficiency tend to emit less CO2. The correlation of CO2 consumption with vehicle manufacturer(make\_id), combined volume (volume), and vehicle type (Vehtype) is -0.2157, -0.4323, and -0.3626, respectively. Thus, we get a valuable insight that the vehicle manufacturer, interior size, or vehicle type does not predict CO2 emissions, and there is a weak relationship between these variables. From table 4, the categorical variable make\_id shows a weak correlation (r < 0.2) with all the variables. Pearson correlation may not be a good choice to examine relationships between categorical and quantitative variables since it needs numeric data; hence, different analytical method like the Chi-Square test can be used for such cases.

**Chi-Square Test for Homogeneity and Independence**

The table 5 shows there are six emission categories and four Vehicle types. We aim to see if there is an association between vehicle type and emission level. The chi-squared test of independence is used to determine whether the two categorical variables, Vehicle type and emission, are statistically independent. If there is no impact of vehicle type on the emission level, we say that the two variables are independent. The null hypothesis states that the emission category is not dependent on vehicle type and the alternate hypothesis states that the emission is associated with vehicle type.

From the table, the standard emission category makes 68.42% of the total Vehicle type, where 81.22 % is passenger 2 door and 73.6 % is passenger 4 door. Similarly, 78.15 % of gross polluters and 86.5% of polluters are Unknown vehicle type. If emissions and vehicle type were independent, these proportions would be more balanced, but the table shows disproportionality. This is strong evidence against independence. The Chi-square test statistic, as given in the formula, is calculated by dividing the square of the difference between the observed frequency and the expected value by the expected value for each observation (Devore, 2016). The summation of all chi-square values was calculated to be 9,406.0459. The degree of freedom was found to be 15. Further, with a large test statistic and low p-value (*p*<0.001), we reject the null hypothesis. We conclude there is an association between vehicle type and emission level.

The chi-squared statistic for each cell (*i,j)* is calculated by using the formula

=

Where *i*= Number of rows

*j*=Number of columns

*n*=total population

=Observed values

= is the estimated expected count of each cell (*i,j*)

Similarly, the Chi-square test of homogeneity is used to compare the distribution of one categorical variable across different groups. It compares observed count to expected count using a two-way contingency table as in Table 6, and the chi-squared statistic is calculated for each cell using the above formula. The Null hypothesis states that all vehicle types have the same distribution of emissions, and the alternative hypothesis states that the distribution of emissions differs across various Vehicle types. If there is a significant difference, we conclude the group doesn’t share the same distribution of emissions. The result shows a significant difference, as the total chi-square value calculated is 9406.0459. Also, the p-value is less than 0.001, so the hypothesis of homogeneity is rejected in favor of the alternative hypothesis (Devore, 2016), and we can conclude that the distribution of emissions differs across various Vehicle types. However, with 15 degrees of freedom, a low p-value, and a very high Chi-square value, further research should be done to conclude the test of homogeneity.

**Issues of Multicollinearity and Overfitting**

As Kutner et al. (2004) explains, if two or more of the independent variables in the regression model are strongly correlated, it becomes increasingly difficult to estimate the unique impact of each individual independent variable. This is known as multicollinearity, and can leads to an unstable model as small changes in the data can result in significant changes in regression coefficients and p-values. “As a result, the standard errors for the regression coefficients are inflated, increasing the possibility of Type II errors” (Kutner et al., 2004). Even with a strong fit to the model, multicollinearity can weaken the statistical significance of relevant variables, making the model unreliable (Midi et al., 2010). In the original model, VIF (Variation Inflation Factor) exceeds 5 for many of the predictors, suggesting potential multicollinearity (Devore, 2016). Furthermore, there are predictors within the model that are highly correlated (correlation is greater than 0.7, as shown in Table 3) and categorical variables incorrectly coded as numeric. This makes the model more complex and puts it at a higher risk of overfitting and severe multicollinearity at higher value, which can decrease the predictive ability of the model. Overfitting happens when a model not only captures the underlying pattern in the data but also the random noise and outliers, which can reduce its reliability when used with new data (Ying, 2019), leading to a poor predictive performance of model.

While it is not possible to get rid of multicollinearity in a model altogether, it can be mitigated by removing certain predictor variables as long as we do not omit relevant variables**.** One method to do this is to remove predictors that have a Variance Inflation Factor (VIF) above 10, as it is considered severe multicollinearity (Devore, 2016). Additionally, instead of converting all categorical variables into dummy variables, the model can be improved if we are excluding poorly coded categorical variables. Thus, reducing the complexity and improving the predictability of the model.

**Initial Regression Model**

The initial regression model aimed to predict CO2 emissions, where the predictor variables included petroleum consumption (barrels08), Combined MPG (comb08), engine displacement(displ), and cylinder count. The model also includes categorical variables like Vehtype, emissioncat, trantype\_id, prifuel, and make\_id. The value of the model is 0.9828, which means that 98.28 % of the variation in CO2 emissions can be explained by the predictor variables of the model.

Additionally, categorical (qualitative) variables were incorrectly coded as discrete numeric values instead of being coded as dummy variables(indicators).To use these categorical variables in the regression model, a categorical variable with possible C categories requires the use of C-1 indicator variables (Devore, 2016). But including too many indicator variables can increase the complexity of the model, as in the case of make\_id. When categorical variables are coded as numeric variables, it can introduce noise in the model, which can result in overfitting and multicollinearity issues. The initial model has too many predictors than required to explain the dependent output variable, which can result in unstable or overdetermined model.

Although a high value seems to be desirable in a model, the combination of highly correlated predictors and improperly handled categorical variables in the model can cause multicollinearity and overfitting. Devore (2016) explains, alone does not guarantee a good model, particularly if it has multicollinearity (>0.9) which inflates both and VIF, reducing the statistical power and reliability of the model. The regression equation of the initial model with estimated parameters is given as in Equation (3)

= 79.55324 + 20.74377 (barrels08) – 2.53562 (comb08) + 2.17223 (displ) + 2.26299 (cylinders)

+ 0.0033 (make\_id) - 0.00779 (volume) – 0.41356 (vehtype) + 0.77998(trantype\_id)

+ 12.87660 (emissioncat) +2.99025 (prifueltype)+ (3)

is the error term

**Justification of the final regression model**

The main objective of the study is to develop a predictive model for estimating CO2 emissions during the early design phase of vehicle development. This brings up a critical question of how much multicollinearity is acceptable in the model. Whether multicollinearity is an issue or not will depends on the purpose of the model.In case the model is primarily aimed at prediction or at forecasting unseen data, multicollinearity is less of a problem since the focus is on the overall accuracy of our model rather than analyzing the relationship between variables. However, multicollinearity can be an issue when the models purpose is to explain how each predictor variable is correlated to the dependent variable, and to make conclusions about the population. Strong correlation of individual predictors can make the model less reliable when used for inferential purposes. The final model, as given in Equation (4), supports the research objective of making accurate predictions by keeping the most important and relevant variables associated with CO2 emissions. It reduces the complexity of the initial regression model and maintains the accuracy and fit, which is useful for making design decisions in the early design phase.

**Final Regression model**

In the final model, five quantitative predictor variables are used: barrels08, comb08, displ, cylinder, and volume. The variables retained are strong predictors of CO2 emissions, which reduces the complexity of the model compared to the initial model. The coefficient of determination of the final regression model is high(=0.9812) which is same as the initial model, which implies that the final model is useful for forecasting.The adjusted value is also 0.9812, indicating the model is not overfitted by irrelevant predictors and has strong explanatory power.

The categorical variables encoded as numeric in the original model have been excluded from the final model, as converting these categorical variables into dummy variables can increase the model complexity and risk of overfitting (Kutner et al., 2004). All the variables in the final regression model were statistically significant (p-value <0.0001) and has a VIF that is ranging 1.28 to 6.9, which was lower than the initial regression model. This indicates our final model has moderate to acceptable multicollinearity.

The new model is more reliable for drawing conclusions and forecasting CO2 emissions during the early design phase. The regression equation for the final model is as given in Equation 4

= 123.29415 + 22.3662 (barrels08) - 2.7891 (comb08) + 3.20908 (displ) + 0.61407 (cylinders)

-0.01339 (volume) + (4)

**Discussions**

**Conclusion**

The objective of our study was to determine whether vehicle characteristics can be used to predict CO2 Emissions. After conducting exploratory analysis and regression modelling, we found that the hypothesis was mostly supported. For each predictor in the model, hypothesis testing was performed to test the statistical significance. The null hypothesis was that vehicle characteristics did not have an impact on CO2 emissions and the alternate hypothesis was that the variables had a statistically significant impact on pollution levels. Predictors like petroleum consumption, combined fuel economy, engine displacement, number of cylinders, and interior volume showed statistically significant correlation with CO2 emission, while categorical variables like make\_id, Vehicle type, and primary fuel type had a weak correlation with CO2 emission. The initial model showed issues of multicollinearity, overfitting, and had poorly coded categorical variables, which were mitigated in the final regression model.

We were able to successfully develop a predictive model to predict CO2 emissions output in the preliminary engineering design in this study. The final regression model is given as in Equation 4. In the regression equation, each coefficient shows how CO2 emissions (in grams per mile) change when a specific variable changes, while keeping all other variables constant. For example, the coefficient for petroleum consumption is around 22.37, which shows that for each barrel of fuel consumed by vehicle per year, its CO2 emissions will rise by 22 grams per mile. The coefficient for the number of cylinders is 0.61407, which implies that for each additional number of engine cylinders in a vehicle, we can expect emissions to be about 0.61 grams/mile higher. Also, the negative value for the coefficient implies that emissions reduces by 2.79 g/mile for every additional MPG in fuel efficiency. For every liter increase in engine displacement, CO2 emissions increase by 3.21 g/mile. Additionally, for every cubic foot increase in interior volume, CO₂ emissions decreased by 0.013 g/mile, respectively. In addition, all the p-values of these predictors are below than 0.001, which reveals that each of these coefficients were statistically significant for the contribution to our model, which can be adopted by engineers and industries for designing low emission vehicles. The final model and its findings can be used to forecast CO2 emissions in the early design phase of automobiles.

**Strengths and Weaknesses of the Study**

The major strength of the study is its large sample size (N=43177), which reduces the influence of outliers and lowers the chance for type two errors. Also, larger dataset makes us more confident that our estimates are near the true population, with narrow confidence interval. With a large dataset, decision-making can be more solid than assumptions. This helps manufacturers, policymakers, and researchers to make more accurate and confident predictions. But, as Devore (2016) points out, large datasets can also amplify very small differences from the null hypothesis, making them appear statistically significant.

The information gathered from 1984 to 2020 has a number of constraints that might impact the reliability of our analysis. Many of the vehicle models are outdated and no longer useful for new market trends. Outdated data can result in bias and give a wrong interpretation if not analysed carefully. According to Kim et al. (2014), data loses its value over time, particularly when used in dynamic sectors such as transportation. Another concern is that the dataset has a high number of null and missing values in many columns, including the Unknown category of vehicle type, which totals 19730 entries. Missing data can skew statistical analysis and lead to wrong conclusions if not handled properly, as Van Den Broeck et al. (2005) point out. This results in incorrect p-values and a higher chance of type 1 errors. Next, there were differences in the way tailpipe CO2 was measured. For 2013 and beyond, the calculation of tailpipe CO2 is based on EPA air quality testing. For previous years, CO2 is approximated using an emission factor from the EPA (Environmental Protection Agency, 2010). This non-uniform way of testing may lead to errors, especially when we want to compare the whole dataset. There are also other constraints due to missing interior volume data on a few types of vehicle categories, like two-seater vehicles. Also, missing unrounded MPG values for some vehicles reduce the precision of fuel economy comparisons. Finally, some of the qualitative variables are not correctly coded as categorical variables in the initial regression model, which reduces their accuracy and statistical power. In the Final regression model, multicollinearity is not fully eliminated but is reduced by eliminating these poorly coded categorical variables.

The implications of future research and study of the dataset are extensive, as it helps researchers to not only identify gaps but may in fact lead to asking more questions to find patterns and understand complex issues better. A better predictive model can be built, which can be applied in the study of environmental science, transportation, and by policymakers, to make better decisions and policy evaluation. Automotive and energy industries can use the findings to design innovative and vehicles. Better product performance can lead to better designs and thus improve customer satisfaction and lower pollution, making our environment cleaner and healthier for everyone.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

**References**

Bruce, P., & Bruce, A. (2017). *Practical Statistics for data scientists: 50 Essential Concepts*. “O’Reilly Media, Inc.”

Devore, J. L. (2016). Probability and statistics for engineering and the sciences (9th ed.). Cengage Learning.

Kim, G., Trimi, S., & Chung, J. (2014). Big-data applications in the government sector. *Communications of the ACM*, *57*(3), 78–85. <https://doi.org/10.1145/2500873>

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). *Applied Linear Statistical Models* (5th ed.). McGraw-Hill/Irwin.

Midi, H., Sarkar, S., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, *13*(3), 253–267. <https://doi.org/10.1080/09720502.2010.10700699>

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

U.S. Environmental Protection Agency. (2023). *Inventory of U.S. greenhouse gas emissions and sinks: 1990-2021*. <https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks>

Van Den Broeck, J., Cunningham, S. A., Eeckels, R., & Herbst, K. (2005). Data cleaning: Detecting, diagnosing, and editing data abnormalities. *PLoS Medicine*, *2*(10), e267. <https://doi.org/10.1371/journal.pmed.0020267>

Vanderbilt, M. C. (2023). *USD-ADS500A Research Project* [GitHub repository]. GitHub. <https://github.com/mcvanderbilt/USD-ADS500A/tree/main/Research_Project>

Yan, L., Zhang, Q., Zheng, B., & He, K. (2024). Modeling fuel-, vehicle-type-, and age-specific CO2 emissions from global on-road vehicles in 1970–2020. *Earth System Science Data*, *16*(10), 4497–4509. <https://doi.org/10.5194/essd-16-4497-2024>

Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics Conference Series*, *1168*, 022022. <https://doi.org/10.1088/1742-6596/1168/2/022022>

**Table 1**

*Descriptive Statistical Analysis*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | co2Tailpipe(g/mile) | barrels08 | comb08 | displ | volume |
|  | |  |  |  |  |  |
| Mean | | 462.7677599 | 17.1532 | 20.84591 | 3.28676779 | 66.93052 |
| Standard Error | | 0.600482725 | 0.022441 | 0.03958 | 0.006550395 | 0.332269 |
| Median | | 444.35 | 16.4805 | 20 | 3 | 86 |
| Mode | | 493.7222222 | 18.31167 | 18 | 2 | 0 |
| Standard Deviation | | 124.7733169 | 4.662937 | 8.224245 | 1.357023083 | 69.04173 |
| Sample Variance | | 15568.05631 | 21.74298 | 67.63821 | 1.841511649 | 4766.76 |
| Kurtosis | | 2.072877205 | 2.117522 | 64.74742 | -0.491912129 | -0.03132 |
| Skewness | | 0.417488592 | 0.36887 | 6.350721 | 0.658093667 | 0.647535 |
| Range | | 1269.571429 | 47.02714 | 134 | 8.4 | 538 |
| Minimum | | 0 | 0.06 | 7 | 0 | 0 |
| Maximum | | 1269.571429 | 47.08714 | 141 | 8.4 | 538 |
| Sum | 19980460.8 | | 740606.6 | 900043 | 141061.5 | 2889792 |
| Count | 43176 | | 43176 | 43176 | 42918 | 43176 |
|  |  | |  |  |  |  |

*Note*. The dataset includes missing and null values

**Table 2**

*Characteristics of 43177 Vehicle Models by Primary Fuel Type*

A close-up of a calculator

AI-generated content may be incorrect.

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

**Table 3**

*Association of Emission Category by Fuel Type and Other Characteristics*

A close-up of a table

AI-generated content may be incorrect.

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

**Table 4**

*Pearson Correlation Coefficients (N=43,177)*

A screenshot of a computer

AI-generated content may be incorrect.

*Note*. All Correlation Values resulted in *a p-*value < 0.0001

**Table 5**

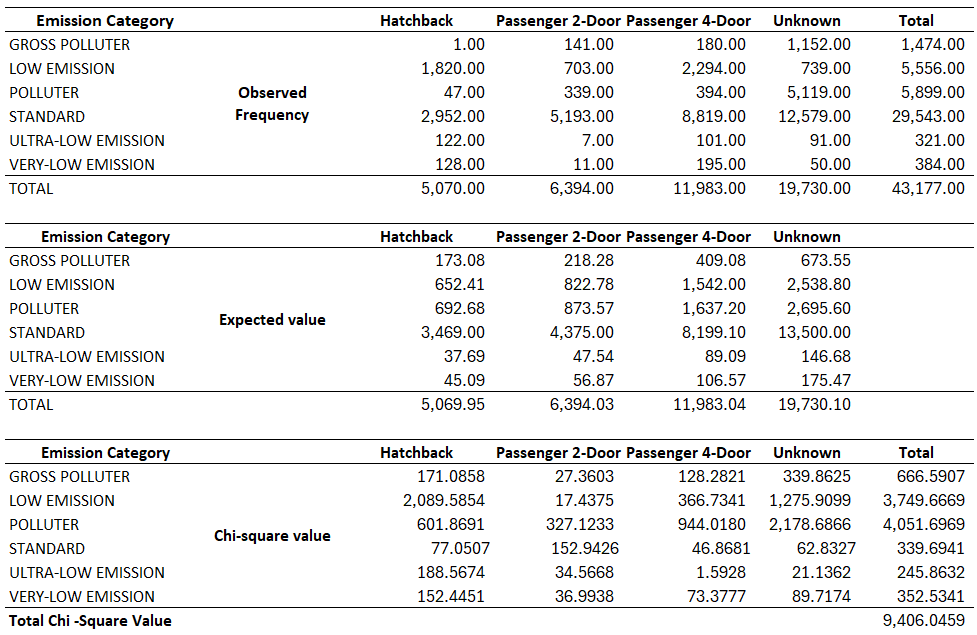
*Chi-Square Table for Emission Category and Vehicle Type*

A screenshot of a table

AI-generated content may be incorrect.

**Table 6**

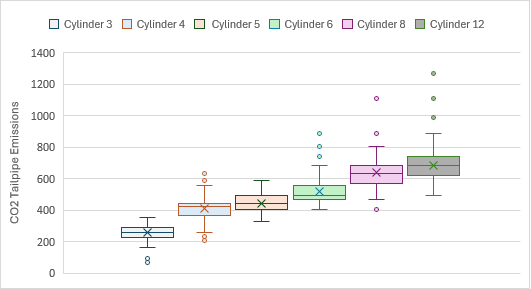
*Two-way Contingency Table for Emission Category and Vehicle Type*



*Note.* Total Chi-square value = 9,406.0459, degrees of freedom 15*, p-*value <0.001.

**Figure 1**

*Boxplot of CO2 Tailpipe Emission and Number of Cylinders*



**Figure 2**

*Probability Density Function of CO2 Emission in Grams per Mile*

A graph of a normal distribution

AI-generated content may be incorrect.

*Note.* Adapted from USD-ADS500A Research Project by M. C. Vanderbilt, 2023, GitHub

**Figure 3:**

Positive correlation between CO2Tailpipe and barrels08

**Figure 4:**

Negative correlation between CO2Tailpipe and comb08