Predicting CO2 emissions from Vehicles

IEOR 242A Final Project Report

Yasmin Graham, Rohini Tamarana, Haakon Tveiten, Yiyang Zhao, Jiacheng Zheng

May 2024

Motivation and Goal

The automotive market is rapidly evolving with the increasing prevalence of electric vehicles (EVs). Despite this trend, fuel-based vehicles still comprised over 80% of the global market in 2023 [1]. This stark reality underscores the urgency and importance of our project aimed at understanding and addressing carbon dioxide (CO2) emissions from vehicles. By predicting CO2 emissions, identifying low-emission vehicle characteristics, and formulating practical recommendations for manufacturers and policymakers, this project aspires to catalyze substantial progress toward more sustainable and environmentally friendly transportation.

1. Predicting CO2 Emissions

The first goal is to develop predictive models for vehicle CO2 emissions using historical data. This approach will enable accurate forecasting of future trends, providing a data-driven foundation for decision-making. Policymakers and manufacturers can leverage these predictions to anticipate market shifts, identify the most pressing areas for improvement, and design effective regulations and products that align with sustainability targets. Robust predictive models not only offer a clear picture of the current emissions landscape but also illuminate potential pathways toward a cleaner future.

2. Analyzing Low-Emissions Vehicle Characteristics

The second aspect of the project involves analyzing vehicle characteristics to identify those that are essential for minimizing emissions. By scrutinizing technical features and their correlation with emission levels, we can pinpoint the design elements that significantly contribute to lower emissions. This analysis informs manufacturers on how to prioritize and innovate vehicle designs that align with stringent environmental standards. It also empowers consumers to make greener choices when purchasing vehicles by highlighting models that embody effective emissions reduction strategies.

3. Developing Recommendations for Reducing Emissions

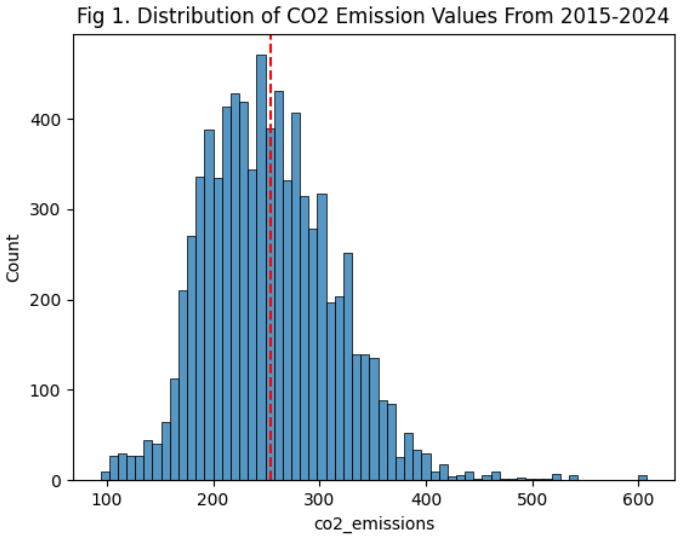
Finally, the project focuses on developing practical recommendations for reducing CO2 emissions from vehicles. These guidelines are tailored to meet the needs of both manufacturers and legislative bodies. For manufacturers, the recommendations emphasize implementing more efficient technologies, transitioning to sustainable materials, and prioritizing low-emission designs. For legislators, they provide a framework for establishing clear, impactful standards that promote rapid adoption of greener practices. Through such targeted recommendations, stakeholders can work in unison to reduce the environmental impact of vehicles and build a market conducive to sustainable innovation.

Data

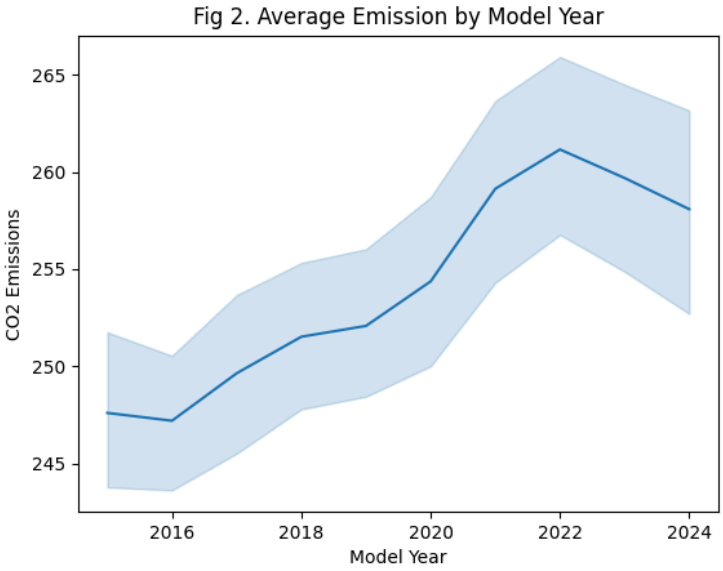
The data we used for this project came from an open dataset from the Government of Canada that looked at the fuel consumption rating and estimated carbon dioxide emissions for different car brands from 1995 to 2024 [2]. The fuel consumption ratings were found while each car was in a 5-cycle fuel consumption test. The 5 cycle testing simulates trips under city and highway conditions, cold temperature operation, air conditioner use and higher speeds that result in more rapid acceleration and braking. The estimated carbon dioxide emissions were

found by measuring the emissions from the vehicle's tailpipe during the testing period. When thinking through the applications of this project, we thought of which model years of cars we currently see most often on the road. To that end, rather than looking at the full dataset available, we looked at data from 2015-2024. The dataset provides information on the model year, make, model type for each brand, the vehicle class, the transmission and fuel type, the engine size, the number of cylinders, the fuel consumption under city, highway and combined setting, the estimated CO2 emissions, the co2 rating and the smog rating. With such a variety of variables, we were able to separate them into categorical variables and numerical variables.

While the dataset was mostly clean, some of the CO2 and smog ratings were missing values. In order to account for this, we took two approaches: 1) use the dataset as is but disregard these variables as they are probably a calculation from the co2 emissions values; 2) incorporate the fact that data is missing into how the model analyzes the data. To create a training and testing dataset, we did a random 80-20 split of the data.



Prior to building our models, we first visualized the data. As we were interested in the CO2 emissions as our dependent variable, we looked at this relative to the other variables, both categorical and numerical. Figure 1, on the right, shows the distribution of all CO2 emissions from 2015-2024 as well as the mean value of 253.6, indicated by the red line. From the distribution we saw that most of the CO2 emission values are between 100 and 500. There are three values outside of this range. Figure 2, on the left, shows the average CO2 emissions relative to the model year of the vehicle. One interesting observation from this figure is that the more recent car model years had higher emissions relative to the older models. This is surprising because as technology has improved and as a society we have become more interested in reducing fossil fuel and CO2 emissions, you would expect the inverse relationship to be evident. The other data visualizations and key observations can be found in appendices A-F.



Analytics Model

To analyze the data we built four different models: linear regression, logistic regression and CART. The results from each model can be found below.

Linear Regression Model

Our initial approach for building the linear regression model to predict CO2 emissions was to use all the variables except for the CO2 and smog ratings, as they had missing values and we assumed the ratings were a direct calculation from the CO2 emissions, and so as numerical variables would be highly correlated with the CO2 emission values. The process for finding the final model was done in an interactive manner; in total we went through five iterations. With each iteration, we looked at the p-value and the variance inflation factor (VIF) value, which is a measure to detect multicollinearity in the regression variables, and we dropped any variables whose VIF was greater than five. The first model produced using all the variables and resulted in an R^2 value of 0.997. Such a high R^2 indicated that the model was overfit to the training data and that there may be multicollinearity across the variables. The top half of the summary tables for all the iterations, showing the R^2 , can be found in the appendices O - T. The code will show the full tables.

After the fifth iteration, we found that the best model for predicting CO2 emissions was one built on the number of cylinders, the make of the car, the fuel type and the vehicle class. These variables were selected because they had a low VIF and a p-value close to 0. Initially, we thought the combined fuel consumption was a significant variable, however, we determined that this fuel usage indicator was overpowering the model and resulting in R^2 values still close to 1. Thus for the final model it was removed. This final model had an R^2 value of 0.821. In addition to looking at the R^2 as a performance metric, we also looked at the OSR^2, which was 0.807, and the Root Mean Square Error (RMSE), which was 25.45. For the second approach we added on from the fifth model from the previous approach and took into account the missing values for CO2 and smog rating as categorical variables. This model resulted in an R^2 of 0.909, an OSR^2 of 0.907 and a RMSE of 17.65. The addition of the missing rating data did improve the model.

Logistic Regression Model

To predict the CO2 emissions using a logistic regression model, we used the median of CO2 emissions (249 g/km) as threshold, in order to create a binary variable. This binary variable, which became the new dependent variable, was called ‘Above’ and indicated car models that have CO2 emissions higher than 249 g/km as 1 and otherwise as 0. After testing with different combinations of all independent variables, observing the p-value for each variable and the R^2 value for the whole model, the summary table of our first model is shown in Appendix G. However, the variable “fuel_type[T.N]” still had a large p-value, so we removed it individually and conducted the logistic regression model again to have our final logistic regression model shown in Appendix H.

From the summary table (Appendix H), we had fuel type E, X, and Z, engine size, and cylinders as independent variables. All variables had the p-value equal to 0 which indicated their significance. Fuel type E, X, and Z had negative coefficients which implied that as these variables increased, the log-odds of CO2 emissions being higher than 249 decreased. In other words, these variables are associated with lower probabilities of exceeding the CO2 emissions threshold of 249. Meanwhile, variable “engine size” and “cylinders” had positive coefficients which showed these variables were associated with higher probabilities of exceeding the CO2 emissions threshold of 249. Also, we could compare between fuel type E, X, and Z. Since fuel type E’s coefficient had the largest absolute value among those three, fuel type E had the strongest association with lower CO2 emissions compared to fuel type X and Z. Thus, we would recommend using fuel type E to reduce CO2 emissions if applicable.

The performance metrics of the logistic regression model were the R^2 value and accuracy. The R^2 value of the model was 0.5462 which means only 54.62% of the variability of the dependent variable had been explained by the independent variables. After testing the logistic regression model with the test dataset, we acquired the accuracy of this model was 0.8646 (Appendix I).

CART Model

In developing a decision tree model to predict the CO2 emissions based on vehicle characteristics we took an iterative design approach. Our primary goal was to identify the most influential variables for predicting emissions. Initially, we removed certain variables that were either irrelevant or overly predictive. Specifically, the "id-variable" was eliminated due to its irrelevance, and both "co2_rating" and "smog_rating" were removed since they are directly derived from CO2 emissions. We also excluded "model" to prevent overfitting and maintain generalizability and "Combined MPG" was dropped in favor of "Combined L/100 km," as these variables are measures of the same attribute but expressed in different units. We also decided to remove the "model_year" because it is not a factor controlled by the car manufacturers. If consumers are interested in identifying which years typically have low emissions, they can refer to the plots of average emissions per year. (Appendix J).

During our analysis, we discovered that fuel consumption variables, particularly "Combined L/100 km," were the most significant predictors of emissions. This aligns with the expectation that a vehicle's fuel efficiency strongly influences its emissions levels. To explore further, we iteratively removed variables related to fuel usage: "Combined L/100 km," (Appendix K) followed by "City L/100 km" (Appendix L), and finally "Highway L/100 km" (Appendix M). Subsequently, we developed a model that predicts CO2 emissions using non-fuel-related variables.

Initially, our model achieved a RMSE of 23.6 and an R^2 value of 0.83. However, the model's complexity made it difficult to interpret, so we pruned it to simplify it. After the pruning we had a way more interpretable model, and the model mostly retained its performance metrics (RMSE of 28.17 and R^2 of 0.76). In this refined model (Appendix N), "engine size" emerged as the primary determinant of emissions levels, with "number of cylinders" and "transmission type" also contributing to the predictions. Having a bigger engine size or more cylinders proved to increase the likelihood of higher emissions and interestingly having a AV transmission (continuously variable transmission) proved to reduce the likelihood of high emissions.

Model Comparison

Model Type	Performance Metrics	Values
Linear Regression	R^2, OSR^2 and RMSE	0.821, 0.807, & 25.45
Linear Regression with Categorical Missing Data	R^2, OSR^2 and RMSE	0.909, 0.907 & 17.65
Logistic Regression	R^2 and Accuracy	0.5462 & 0.8646
CART Model	R^2 and RMSE	0.76 & 28.17

From the comparison of the performance metrics we can see that our models are performing reasonably well even when not using the fuel economy variables. The fact that bigger engine size or a higher number of cylinders generally increase emissions is a natural find, however the type of transmission the car has also impacts the likelihood of higher or lower emissions. Judging from our performance metrics we are generally confident in our results.

Extend analysis

To extend the analysis we could experiment with interaction terms for our variables. For instance engine size with the number of cylinders might produce predictions that are better than what we currently have. We could also look into incorporating more datasets with new variables that can impact emissions. We could explore other modeling techniques like neural networks or gradient boosting machines.

Potential Impact

Our project aims to predict CO2 emissions and identify low-emission vehicle characteristics to inform manufacturers, policymakers, and consumers about more sustainable practices. The models we have developed allow consumers to identify vehicles that produce lower emissions, empowering them to make environmentally friendly purchasing decisions. Automotive manufacturers can use the predictive models to innovate and prioritize designs that align with stringent environmental standards, thus advancing the industry's sustainability goals.

The insights collected from our predictive models provide valuable data and perspectives for policymakers to develop regulations that align with climate goals. By setting clear and impactful standards, legislators can facilitate the adoption of more efficient vehicles, ultimately reducing transportation's environmental footprint. Additionally, the project highlights specific vehicle features and technical factors that contribute to emissions, enabling stakeholders to design comprehensive strategies for sustainability in the future.

However, the model's predictive power also presents potential challenges. If not applied with caution, it could encourage "greenwashing," wherein manufacturers market vehicles as environmentally friendly despite marginal improvements in emissions. Thus, robust verification and transparent methodologies are crucial to ensure the models genuinely drive impactful sustainability initiatives.

Further expanding the scope of the project could involve using more extensive datasets to refine the models, enabling manufacturers to better understand the specific characteristics that affect emissions. By including a wider range of vehicles and technical features, the project can yield more precise recommendations and help shape a more sustainable automotive industry.

Code Link

https://deepnote.com/workspace/rohini-bf1e-3f9de230-e8e0-4859-a2c4-c208458afd41/project/Predicting-CO2-Emissions-8d187ae3-23b2-422f-912e-2c9c5d762e04/notebook/Final%20Project%20code-39d9ce621ada438f95ff28f52df21c25

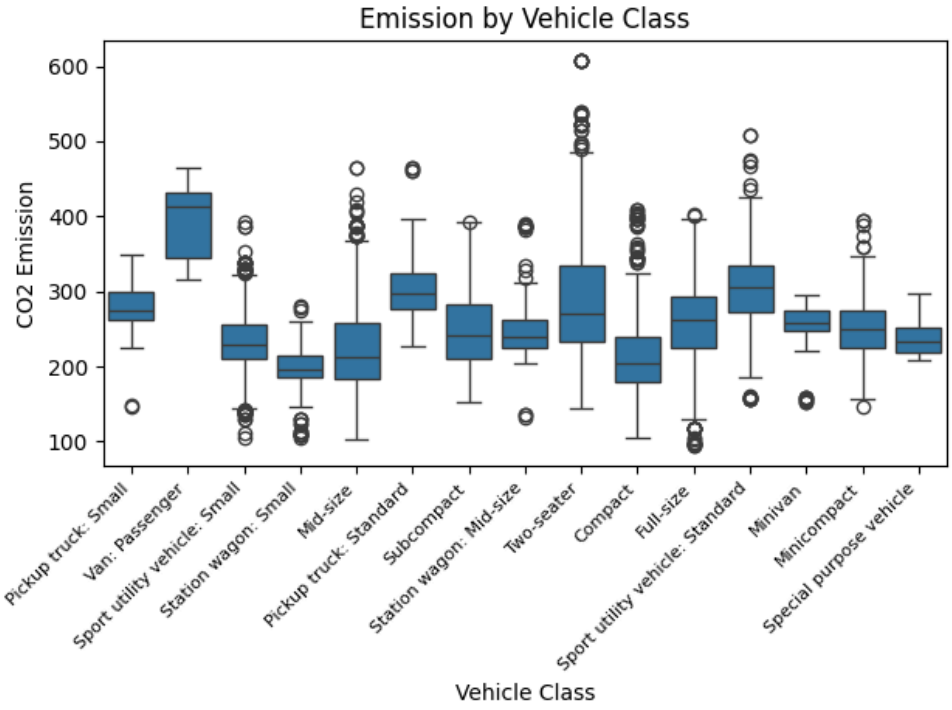
In order to reproduce the results simply run the code one code block at a time.

Reference

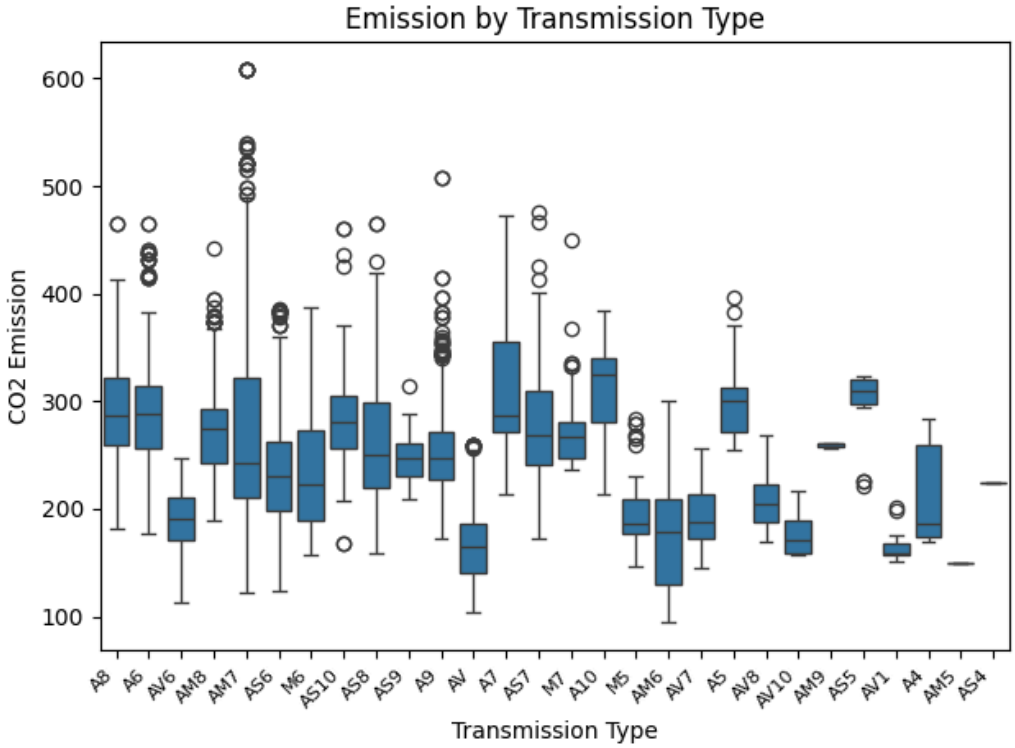
1. Statista. “Global Share of CO2 Emissions from Fossil Fuel and Cement.” https://www.statista.com/statistics/1129656/global-share-of-co2-emissions-from-fossil-fuel-and-cement/.
2. Government of Canada. “Fuel Consumption Ratings.” *Open Government Portal.* https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64.
3. Statista. “Estimated Share of CO2 Emissions in the Transportation Sector.” https://www.statista.com/chart/30890/estimated-share-of-co2-emissions-in-the-transportation-sector/.

Appendix

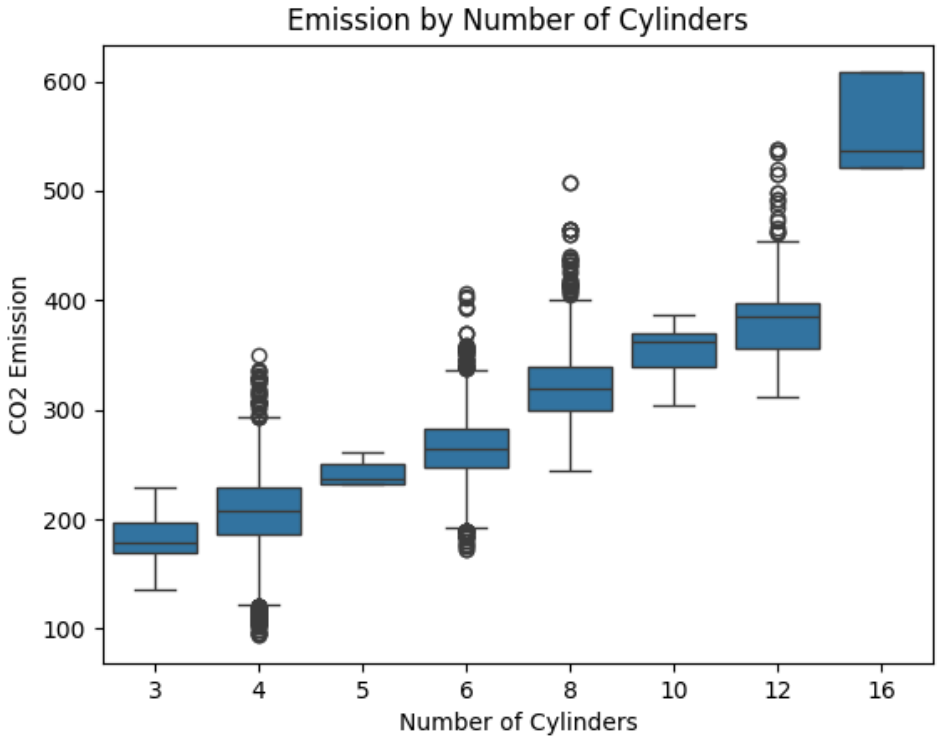
A. CO2 emission values based on the vehicle class across all car brands. Of the different vehicle classes, the station wagon: small, has the lowest median CO2 emission value.



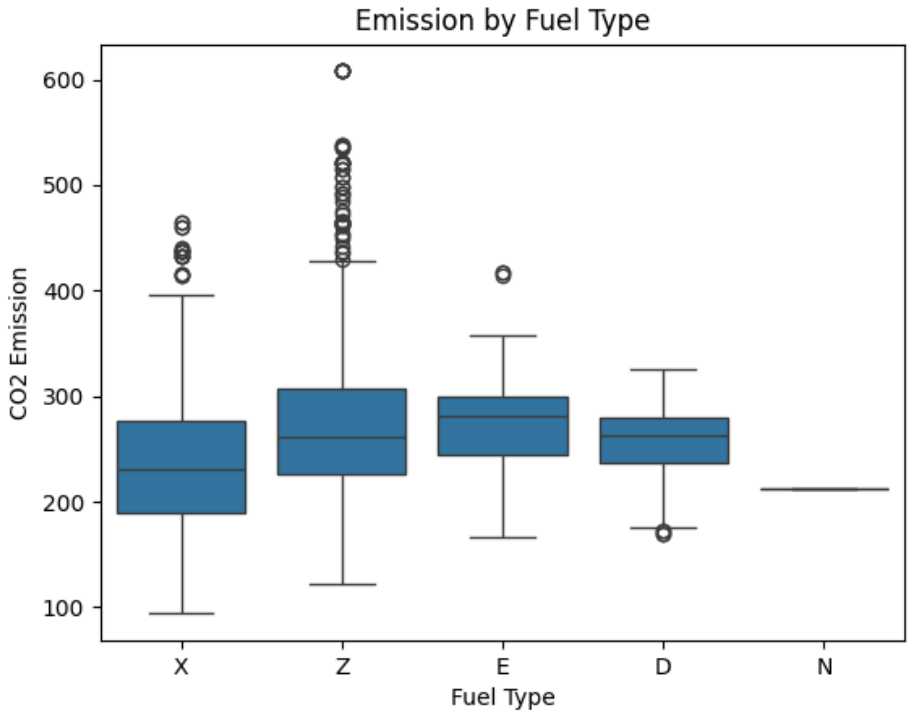
B. CO2 emission values for the individual transmission types. The transmission type with the widest distribution of CO2 emissions was AM7.



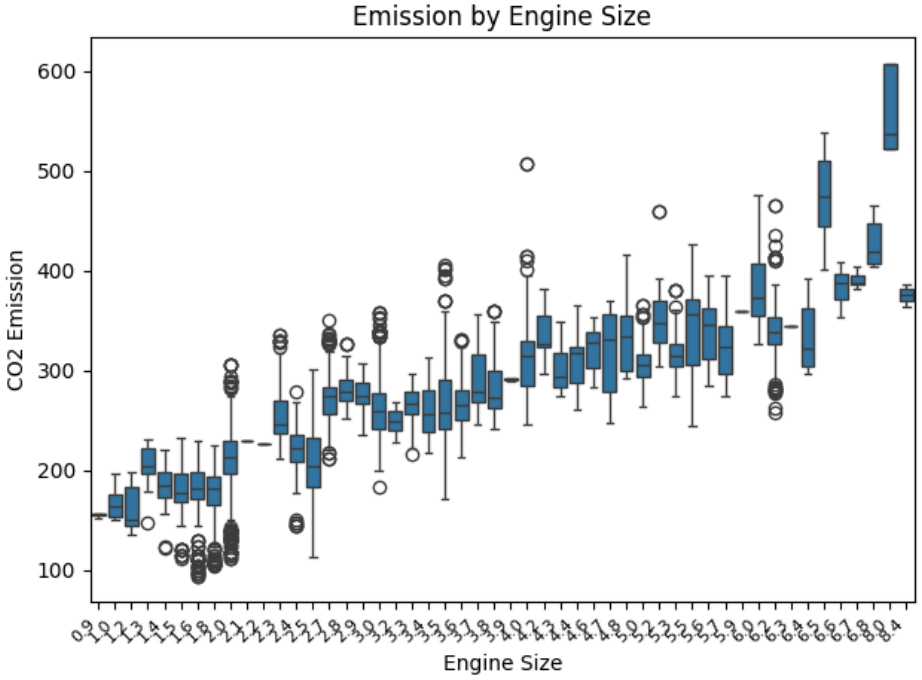
C. CO2 emission values based on the number of cylinders. There is a positive linear relationship between the CO2 emissions and the number of cylinders. As the number of cylinders increases, the CO2 emissions also increase.



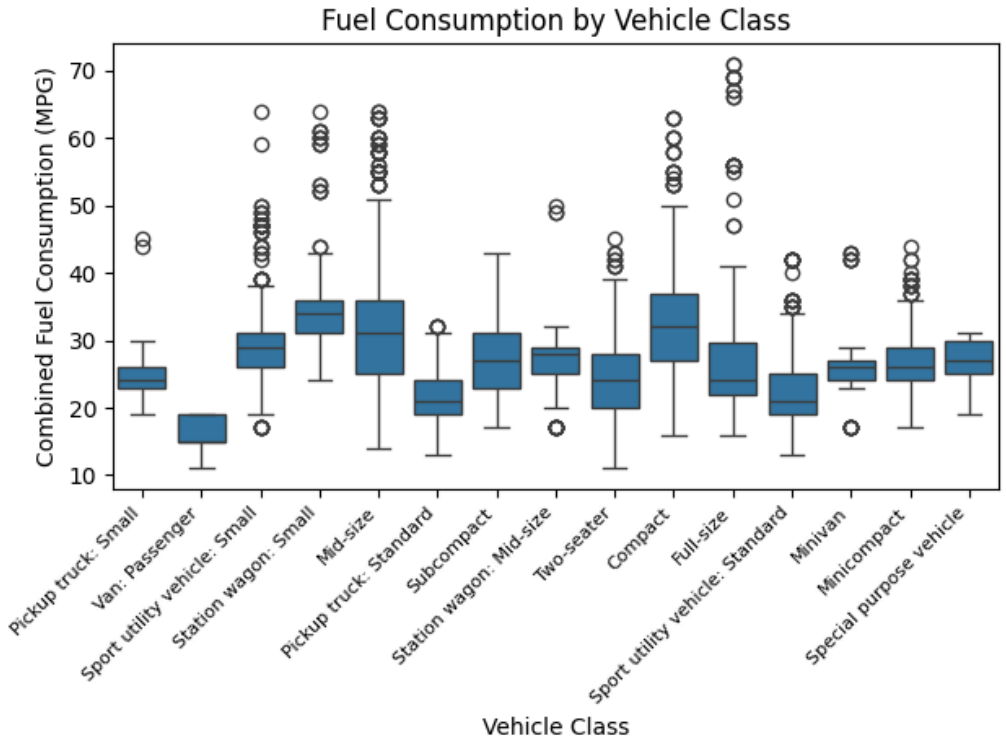
D. CO2 emission values for the individual fuel types. The median CO2 emission values for the different fuel types are very similar.



E. CO2 emission values based on the engine size. There is a linear relationship between the emission and the engine size whereas the engine size increases, more CO2 is produced. The highest CO2 reading comes from vehicles with an engine size of 8.4.



F. Combined fuel consumption relative to the vehicle class. A smaller fuel consumption value indicates that the vehicle is very efficient with fuel usage. The lower combined fuel consumption ratings were attributed to the pick up truck: standard, the sport utility vehicle: standard and the van: passenger.



G. The summary table of our final logistic regression model. Variable “fuel_type[T.N]” has a large p-value.

```
         Iterations: 35
                           Logit Regression Results
==============================================================================
Dep. Variable:                  Above   No. Observations:                 7944
Model:                          Logit   Df Residuals:                     7937
Method:                           MLE   Df Model:                            6
Date:                Thu, 09 May 2024   Pseudo R-squ.:                  0.5467
Time:                        02:19:08   Log-Likelihood:                -2495.9
converged:                      False   LL-Null:                       -5506.1
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept         -7.4513      0.267    -27.868      0.000      -7.975      -6.927
fuel_type[T.E]    -2.1388      0.275     -7.768      0.000      -2.679      -1.599
fuel_type[T.N]   -32.3666   2.62e+06  -1.23e-05      1.000   -5.14e+06    5.14e+06
fuel_type[T.X]    -0.9961      0.192     -5.177      0.000      -1.373      -0.619
fuel_type[T.Z]    -1.0491      0.191     -5.496      0.000      -1.423      -0.675
engine_size        2.1013      0.128     16.443      0.000       1.851       2.352
cylinders          0.4488      0.078      5.729      0.000       0.295       0.602
==============================================================================
```

H. The summary table of our final logistic regression model.

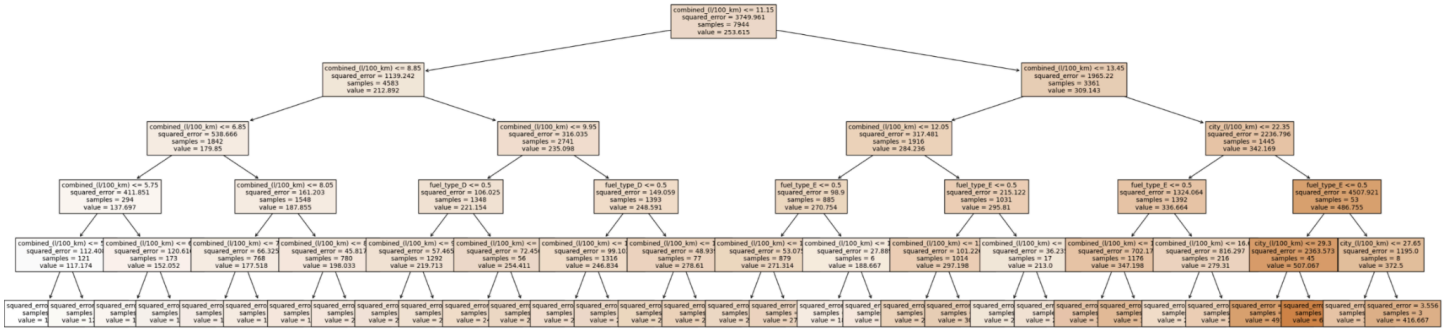
```
Optimization terminated successfully.
         Current function value: 0.314549
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:                  Above   No. Observations:                 7944
Model:                          Logit   Df Residuals:                     7938
Method:                           MLE   Df Model:                            5
Date:                Thu, 09 May 2024   Pseudo R-squ.:                  0.5462
Time:                        02:19:14   Log-Likelihood:                -2498.8
converged:                       True   LL-Null:                       -5506.1
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      -7.4778      0.267    -28.016      0.000      -8.001      -6.955
E              -2.1045      0.275     -7.665      0.000      -2.643      -1.566
X              -0.9641      0.191     -5.037      0.000      -1.339      -0.589
Z              -1.0180      0.190     -5.364      0.000      -1.390      -0.646
engine_size     2.0949      0.128     16.415      0.000       1.845       2.345
cylinders       0.4512      0.078      5.765      0.000       0.298       0.605
==============================================================================
```

I. Accuracy of our logistic regression model with test dataset.

```
Confusion Matrix :
 [[3275  667]
 [ 409 3593]]
Logistic Regression Test Accuracy: 0.8646
```

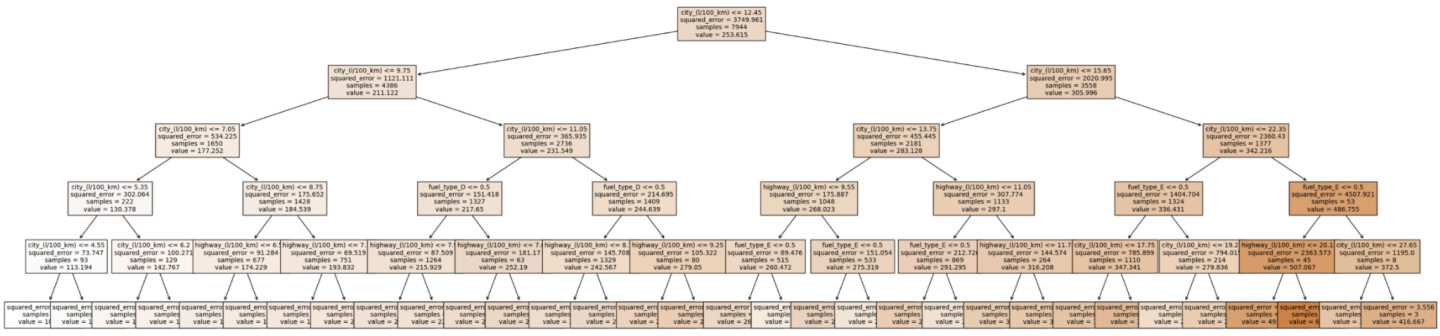
J. First iteration of CART

```
Mean Squared Error: 68.59322069051532
Root Mean Squared Error: 8.282102431781155
Mean Absolute Error: 5.296716279425655
R^2 Score: 0.979513821215842
```



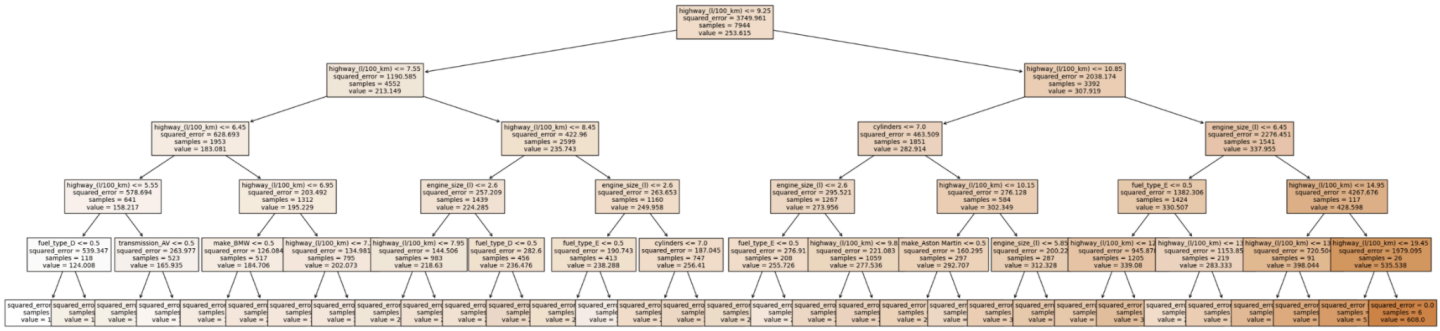
K. Second iteration of CART

```
Mean Squared Error: 114.41744253537483
Root Mean Squared Error: 10.696608926915802
Mean Absolute Error: 7.633267132114627
R^2 Score: 0.965827873947171
```



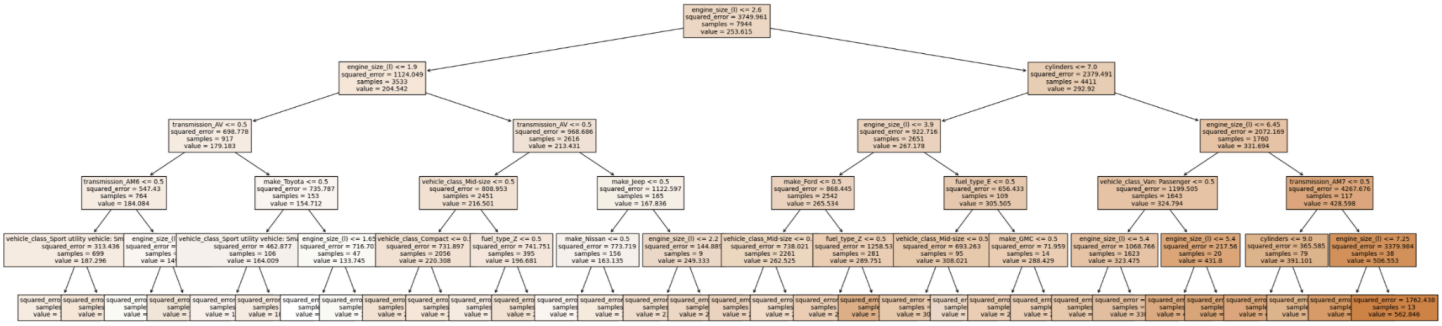
L. Third iteration of CART

```
Mean Squared Error: 247.7276058139052
Root Mean Squared Error: 15.739364847855366
Mean Absolute Error: 11.02661984580609
R^2 Score: 0.9260132128017016
```



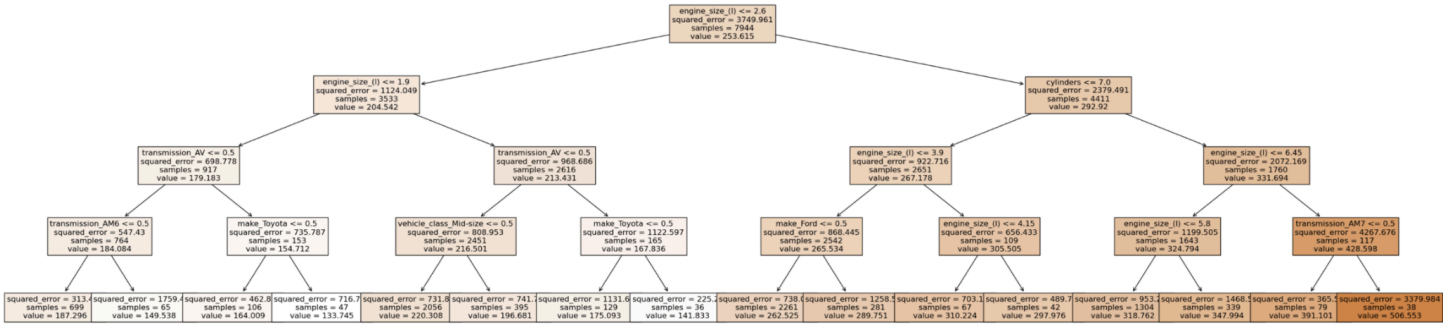
M. Fourth iteration of CART

```
Mean Squared Error: 692.8135329919099
Root Mean Squared Error: 26.32135127594915
Mean Absolute Error: 20.008283928125316
R^2 Score: 0.7930830225191781
```



N. Fifth iteration of CART

```
Pre-Pruned Tree - RMSE: 26.29577389275738, MAE: 19.998698494879953, R^2 Score: 0.7934849641492498
Post-Pruned Tree with optimal ccp_alpha - RMSE: 28.17680064451186, MAE: 21.6235125752302, R^2 Score: 0.7628827649742108, Optimal ccp_alpha: {'ccp_alpha': 0.001}
```



O. Model 1 for linear regression

OLS Regression Results

Dep. Variable:	co2_emissions	R-squared:	0.997
Model:	OLS	Adj. R-squared:	0.997
Method:	Least Squares	F-statistic:	2.477e+04
Date:	Thu, 09 May 2024	Prob (F-statistic):	0.00
Time:	05:37:04	Log-Likelihood:	-21422.
No. Observations:	7944	AIC:	4.303e+04
Df Residuals:	7851	BIC:	4.368e+04
Df Model:	92		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	59.5760	1.262	47.192	0.000	57.101	62.051
make[T.Alfa Romeo]	1.2897	0.776	1.662	0.097	-0.232	2.811
make[T.Aston Martin]	2.1861	0.744	2.938	0.003	0.728	3.645
make[T.Audi]	0.4720	0.530	0.890	0.373	-0.567	1.511
make[T.BMW]	0.2431	0.512	0.475	0.635	-0.760	1.246
make[T.Bentley]	3.2350	0.702	4.606	0.000	1.858	4.612

P. Model 2 for linear regression

OLS Regression Results

Dep. Variable:	co2_emissions	R-squared:	0.997
Model:	OLS	Adj. R-squared:	0.996
Method:	Least Squares	F-statistic:	2.470e+04
Date:	Thu, 09 May 2024	Prob (F-statistic):	0.00
Time:	05:36:52	Log-Likelihood:	-21477.
No. Observations:	7944	AIC:	4.314e+04
Df Residuals:	7852	BIC:	4.378e+04
Df Model:	91		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	60.2614	1.269	47.477	0.000	57.773	62.750
make[T.Alfa Romeo]	1.0895	0.781	1.395	0.163	-0.442	2.621
make[T.Aston Martin]	2.2064	0.749	2.946	0.003	0.738	3.675
make[T.Audi]	0.4978	0.534	0.933	0.351	-0.548	1.544
make[T.BMW]	0.3478	0.515	0.676	0.499	-0.662	1.357
make[T.Bentley]	3.3714	0.707	4.768	0.000	1.985	4.757

Q. Model 3 for linear regression

OLS Regression Results

Dep. Variable:	co2_emissions	R-squared:	0.984
Model:	OLS	Adj. R-squared:	0.984
Method:	Least Squares	F-statistic:	5359.
Date:	Thu, 09 May 2024	Prob (F-statistic):	0.00
Time:	05:36:40	Log-Likelihood:	-27540.
No. Observations:	7944	AIC:	5.526e+04
Df Residuals:	7853	BIC:	5.590e+04
Df Model:	90		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	149.3399	2.474	60.357	0.000	144.490	154.190
make[T.Alfa Romeo]	9.3024	1.672	5.563	0.000	6.024	12.580
make[T.Aston Martin]	7.0031	1.606	4.361	0.000	3.856	10.151
make[T.Audi]	7.0724	1.142	6.194	0.000	4.834	9.310
make[T.BMW]	5.9071	1.102	5.359	0.000	3.746	8.068
make[T.Bentley]	25.4126	1.490	17.051	0.000	22.491	28.334

R. Model 4 for linear regression

OLS Regression Results

Dep. Variable:	co2_emissions	R-squared:	0.983
Model:	OLS	Adj. R-squared:	0.983
Method:	Least Squares	F-statistic:	5071.
Date:	Thu, 09 May 2024	Prob (F-statistic):	0.00
Time:	05:36:30	Log-Likelihood:	-27800.
No. Observations:	7944	AIC:	5.578e+04
Df Residuals:	7854	BIC:	5.641e+04
Df Model:	89		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	146.6950	2.554	57.440	0.000	141.689	151.701
make[T.Alfa Romeo]	6.8097	1.724	3.949	0.000	3.430	10.190
make[T.Aston Martin]	-0.1399	1.628	-0.086	0.931	-3.331	3.051
make[T.Audi]	4.4201	1.174	3.766	0.000	2.119	6.721
make[T.BMW]	3.5143	1.134	3.099	0.002	1.292	5.737
make[T.Bentley]	20.4378	1.524	13.412	0.000	17.451	23.425

S. Model 5 for linear regression

OLS Regression Results

Dep. Variable:	co2_emissions	R-squared:	0.821
Model:	OLS	Adj. R-squared:	0.819
Method:	Least Squares	F-statistic:	601.7
Date:	Thu, 09 May 2024	Prob (F-statistic):	0.00
Time:	05:54:03	Log-Likelihood:	-37131.
No. Observations:	7944	AIC:	7.438e+04
Df Residuals:	7883	BIC:	7.481e+04
Df Model:	60		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	71.8720	3.918	18.344	0.000	64.192	79.552
make[T.Alfa Romeo]	14.5536	5.224	2.786	0.005	4.313	24.794
make[T.Aston Martin]	-8.7814	4.941	-1.777	0.076	-18.466	0.904
make[T.Audi]	12.6959	3.482	3.646	0.000	5.870	19.522
make[T.BMW]	9.6081	3.331	2.884	0.004	3.078	16.138
make[T.Bentley]	25.3010	4.664	5.424	0.000	16.157	34.445

T. Model 6 for linear regression

OLS Regression Results

Dep. Variable:	co2_emissions	R-squared:	0.909
Model:	OLS	Adj. R-squared:	0.909
Method:	Least Squares	F-statistic:	1025.
Date:	Thu, 09 May 2024	Prob (F-statistic):	0.00
Time:	05:54:27	Log-Likelihood:	-34421.
No. Observations:	7944	AIC:	6.900e+04
Df Residuals:	7866	BIC:	6.954e+04
Df Model:	77		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	152.1527	3.083	49.351	0.000	146.109	158.196
make[T.Alfa Romeo]	8.3940	3.730	2.251	0.024	1.083	15.705
make[T.Aston Martin]	1.3289	3.526	0.377	0.706	-5.582	8.240
make[T.Audi]	8.3172	2.488	3.343	0.001	3.441	13.194
make[T.BMW]	5.2957	2.377	2.228	0.026	0.636	9.956
make[T.Bentley]	13.3660	3.351	3.989	0.000	6.797	19.935