Describing the linear correlation between Fuel Consumption Comb (miles per gallon) and the CO2 Emissions(g/km)

1. Introduction

This study develops predictive models to estimate CO2 emissions based on various vehicle attributes. An in-depth analysis of the chosen dataset was conducted, leading to the creation of four prediction models. These models describe the linear correlation between Fuel Consumption Combined (miles per gallon) and CO2 Emissions (g/km). Further analysis of the models was performed to evaluate their performance. Based on the results, several business applications were suggested, highlighting the models' potential for optimizing operations and aligning with sustainability goals.

2. Materials and Methods

❖ Dataset description and summary

The dataset utilized in this study comprises comprehensive information on various vehicle attributes, encompassing 11 columns. These columns include unique identification numbers (UUID), car manufacturers (Make), models, vehicle classes, engine sizes, cylinders, fuel types, different fuel consumption metrics, and CO2 emissions. The data types are categorized as follows: objects for UUID, Make, Model, Vehicle Class, Transmission, and Fuel Type; floats for Engine Size and various Fuel Consumptions; and integers for Cylinders and CO2 Emissions. The dataset contains a total of 7,385 data points, providing a robust foundation for the analysis and modeling conducted in this research.

To familiarize with the dataset, the initial rows, final rows, data types, number of rows and columns, and column names were examined. The names and counts of manufacturers ('Make') and vehicle classes were listed to gain further insight. Additional information about the data frame was extracted to understand its structure better. Basic statistical characteristics of each numerical feature in the dataset were also reviewed.

The primary findings indicate significant variability in sample sizes between manufacturers and vehicle classes. Consequently, a smaller number of cars in certain vehicle classes leads to less reliable outcomes.

❖ Technical setup

In the technical setup, the necessary libraries were imported into Google Colab, and the basic variables were defined.

❖ Statistics

The most relevant descriptive statistics identified were for fuel consumption and CO2 emissions. The minimum and maximum values, the range (difference between minimum and maximum values), mean, median, standard deviation and variance were calculated for these variables.

❖ Data cleaning process

The data cleaning process involved removing unnecessary columns to focus on relevant variables for the study. The dataset was thoroughly checked for missing values, and none were found, ensuring completeness. Additionally, certain columns were renamed to enhance clarity and consistency in the analysis.

❖ Exploratory data analysis (Correlation matrix)

The relevant numeric attributes selected for the correlation analysis included Engine Size (L), Cylinders, Fuel Consumption City (L/100 km), Fuel Consumption Highway (L/100 km), Fuel Consumption Combined (L/100 km), Fuel Consumption Combined (mpg), and CO2 Emissions (g/km).

❖ Prediction models

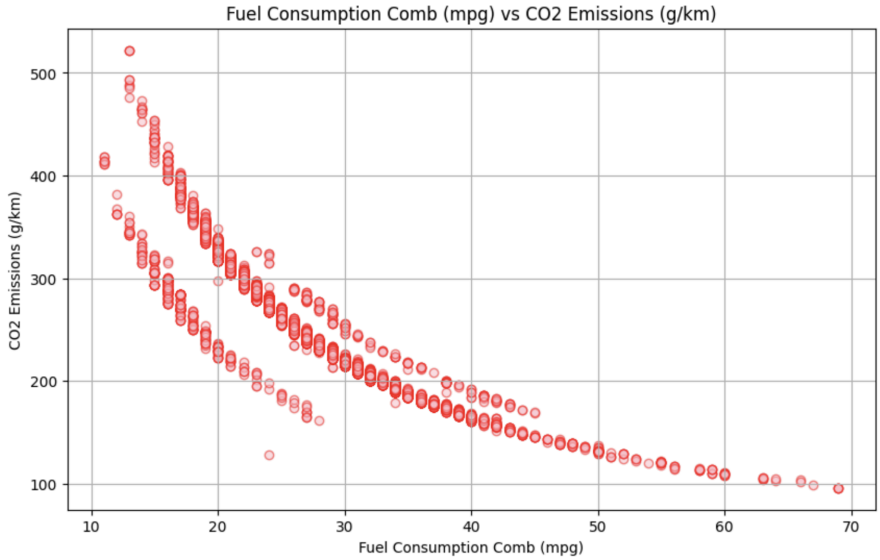
Four prediction models were developed in this study. One model utilized a single independent variable, while the other three models incorporated multiple independent variables.

3. Results

a. Exploratory data analysis (Correlation matrix)

- Correlation between Fuel Consumption Comb (mpg) and CO2 Emissions (g/km)

I created a scatter plot to see the nature of the relationship between the two attributes. The correlation coefficient is -0.91 which shows a strong negative correlation, meaning that as fuel consumption (in mpg) increases, CO2 emissions decrease significantly. This is expected since higher fuel efficiency (higher mpg) generally results in lower emissions.

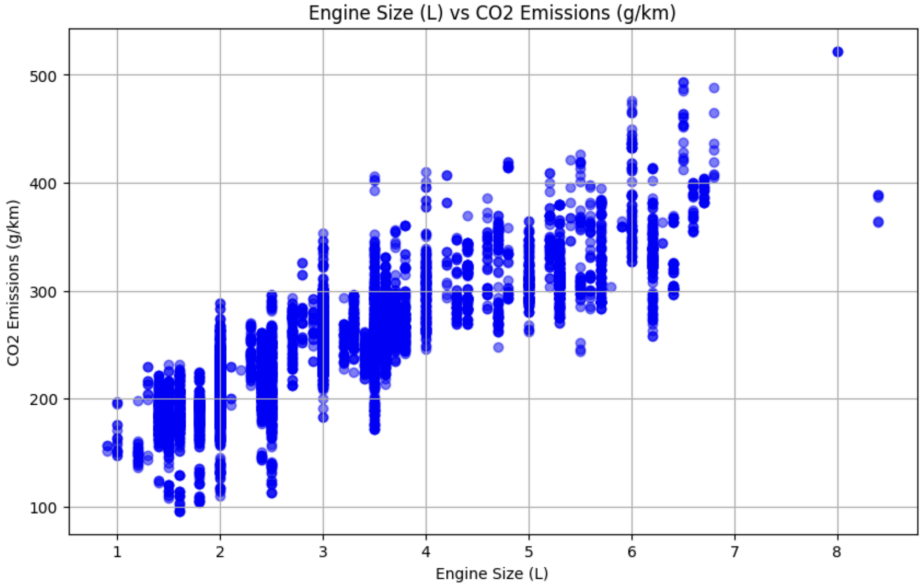



```
Correlation coefficient between Fuel Consumption Comb (mpg) and CO2 Emissions (g/km): -0.91
```

Fig.1 Correlation between Fuel Consumption Comb (mpg) and CO2 Emissions (g/km).

- Correlation between Engine Size (L) and CO2 Emissions (g/km)

I also used a scatter plot. The correlation coefficient is 0.85, therefore it is a positive correlation. This means that the attributes increase together. Therefore, larger engines produce more emissions.

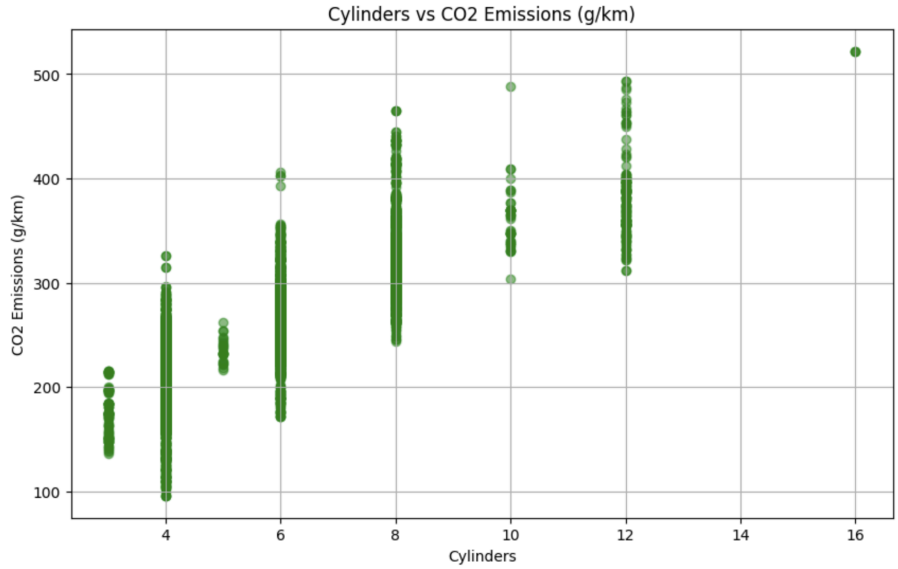



```
Correlation coefficient between Engine Size (L) and CO2 Emissions (g/km): 0.85
```

Fig. 2 Correlation between Engine Size (L) and CO2 Emissions (g/km).

- Correlation between Cylinders and CO2 Emissions (g/km)

I used a scatter plot. The correlation coefficient is 0.83, therefore it is a positive correlation. The result shows that vehicles with more cylinders tend to have higher CO2 emissions.

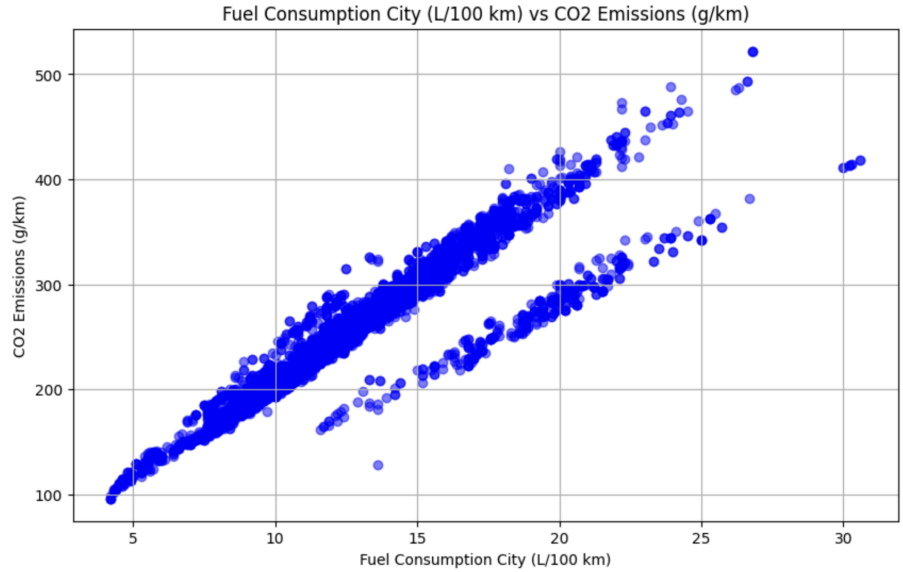



```
Correlation coefficient between Cylinders and CO2 Emissions (g/km): 0.83
```

Fig. 3 Correlation between Cylinders and CO2 Emissions (g/km).

- Correlation between Fuel Consumption City (L/100 km) and CO2 Emissions (g/km)

I used a scatter plot. The correlation coefficient is 0.92, which shows a strong positive correlation. This means that higher fuel consumption in the city leads to higher CO2 emissions.

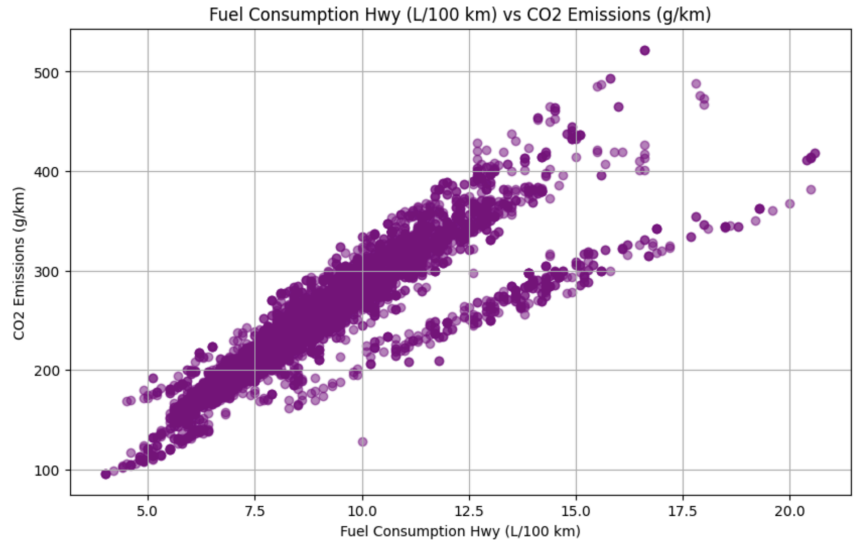



```
Correlation coefficient between Fuel Consumption City (L/100 km) and CO2 Emissions (g/km): 0.92
```

Fig. 4 Correlation between Fuel Consumption City (L/100 km) and CO2 Emissions (g/km).

- Correlation between Fuel Consumption Hwy (L/100 km) and CO2 Emissions (g/km)

I used a scatter plot to understand the data. The correlation coefficient is 0.88 (positive correlation), which shows that higher highway fuel consumption is also associated with higher CO2 emissions.

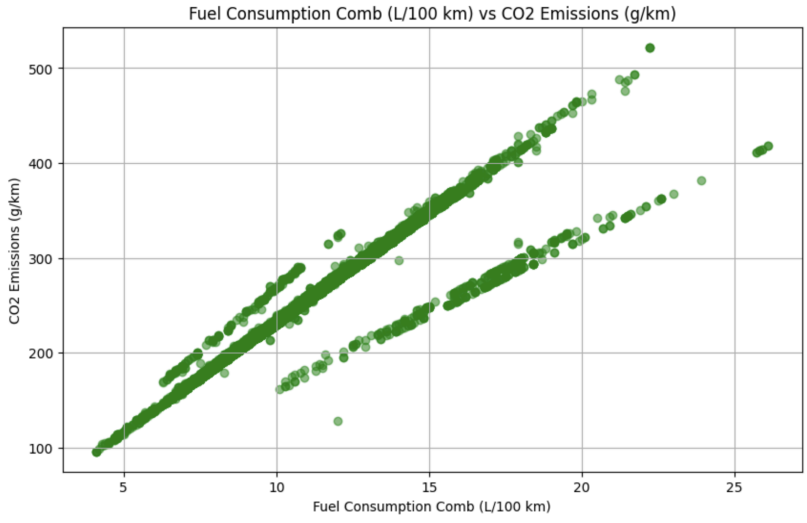



```
Correlation coefficient between Fuel Consumption Hwy (L/100 km) and CO2 Emissions (g/km): 0.88
```

Fig. 5 Correlation between Fuel Consumption Hwy (L/100 km) and CO2 Emissions (g/km).

- Correlation between Fuel Consumption Comb (L/100 km) and CO2 Emissions (g/km)

I used a scatter plot to explore the data. The correlation coefficient is 0.92 (strong positive) shows that overall higher fuel consumption is strongly associated with higher CO2 emissions.




```
Correlation coefficient between Fuel Consumption Comb (L/100 km) and CO2 Emissions (g/km): 0.92
```

Fig. 6 Correlation between Fuel Consumption Comb (L/100 km) and CO2 Emissions (g/km).

- The relationship between Fuel consumption(L/100) and CO2 emissions

The following plot shows the relationship between fuel consumption and CO2 emissions, with data points colored by different fuel types. The points are colored based on different fuel types labeled X, Z, D, E, and N. This allows for a comparison of CO2 emissions across various fuel types. The plot indicates that different fuel types have distinct patterns of CO2 emissions relative to fuel consumption. Some fuel types may produce more CO2 emissions at the same level of fuel consumption compared to others.

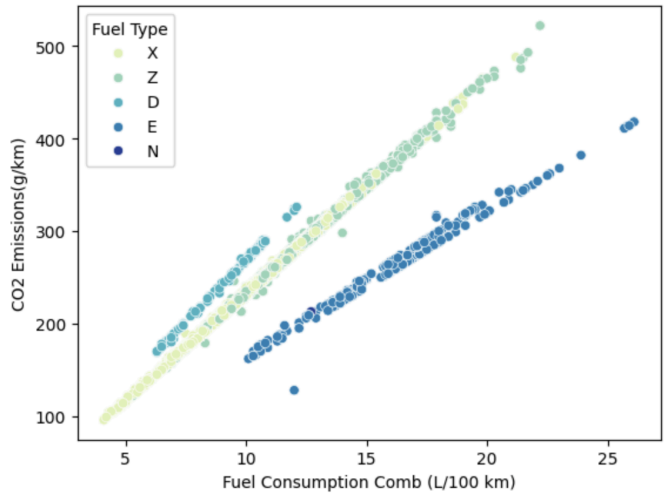


Fig 7. CO2 emissions based on Fuel Type and Fuel consumption Combined (L/100 km).

- Pair plot

A pair plot was done to depict pairwise relationships between fuel consumption and CO2 emissions. Results show a clear negative correlation between fuel consumption (mpg) and CO2 emissions (g/km). As fuel consumption increases (more mpg), CO2 emissions decrease. This is evident from the downward trend in the scatter plots.

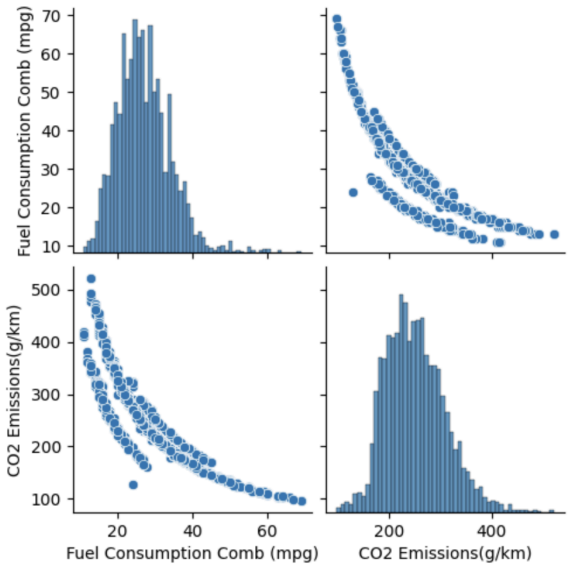


Fig 8. Pait plot. The histogram on the diagonal shows the distribution of the combined fuel consumption (mpg) and indicates how the fuel consumption data is spread across different mpg values. The scatter plots in the off-diagonal show the relationships between the combined fuel consumption and CO2 emissions. The histogram on the diagonal shows the distribution of CO2 emissions (g/km) and provides insights into how CO2 emissions are distributed in the dataset. The scatter plots in the off-diagonal depict the relationship between CO2 emissions and combined fuel consumption.

Causality: We need a more in-depth analysis to explain causality between any of the variables. We can see correlations between consumption and higher CO2 emissions but do not know the other potential factors that might influence the CO2 levels. An understanding of the underlying mechanisms is needed to draw the causality.

Why do I see this much clustering in most of the scatter plots?
Most of the clustering is due to the fuel type. But, the vehicle type can also be a significant factor that causes the clusters. It influences the relationship between fuel consumption and CO2 emissions. Then, vehicle class, engine size, and the number of cylinders should also explain the clustering.

b. Prediction Models

1. OLS Model:

For the first model prediction, we used Fuel Consumption Combined as the independent variable and CO2 Emissions as the dependent variable. The prediction model indicates that CO2 emissions increase with higher fuel consumption, demonstrating that CO2 emissions depend on fuel consumption. Further details about the model are provided in the following paragraphs.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.843
Model:	OLS	Adj. R-squared:	0.843
Method:	Least Squares	F-statistic:	3.959e+04
Date:	Wed, 31 Jul 2024	Prob (F-statistic):	0.00
Time:	08:46:30	Log-Likelihood:	-33697.
No. Observations:	7385	AIC:	6.740e+04
Df Residuals:	7383	BIC:	6.741e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	46.7632	1.059	44.142	0.000	44.686	48.840
x1	18.5713	0.093	198.968	0.000	18.388	18.754

Omnibus:	3592.018	Durbin-Watson:	2.018
Prob(Omnibus):	0.000	Jarque-Bera (JB):	22309.895
Skew:	-2.290	Prob(JB):	0.00
Kurtosis:	10.178	Cond. No.	44.9

Fig. 9 OLS Model.

☐ Explanation

The output provided is from an Ordinary Least Squares (OLS) regression analysis, which models the relationship between a dependent variable (CO2 emissions) and an independent variable (fuel consumption). Here are the key components of the analysis:

Date type of analysis: Indicates when the analysis was conducted and the method used.
Number of observations: Refers to the total data points included in the analysis.
Df residuals: Calculated as the number of observations minus the number of estimated parameters.
Df Model: For this model, "Df Model: 1" indicates it includes one predictor variable.

Regression Statistics

- R-squared (0.843): This statistic represents the proportion of variance in the dependent variable (CO2 emissions) explained by the independent variable (fuel consumption). An R-squared of 0.843 means that 84.3% of the variance in CO2 emissions is explained by fuel consumption.
- Adjusted R-squared (0.843): This is also 0.843, which helps compare models with different numbers of predictors. In this case, it does not provide additional insight as we only have one predictor.

Model Significance

- F-statistic: Measures the overall significance of the model.
- Prob (F-statistic) (0.00): This p-value indicates the model is highly significant.

Covariance Type

- Nonrobust = Standard errors are calculated assuming homoscedasticity (constant variance of the errors). If errors exhibit heteroscedasticity (variable variance), these standard errors may be biased. Robust standard errors can be used to address this.

Coefficients

- coef: Estimated coefficients for the regression equation.
- std err: Standard errors of the coefficients.
- t-statistic: Ratio of each coefficient to its standard error.
- P>|t|: p-value for the t-statistic, testing the null hypothesis that the coefficient is zero. In this case, both p-values are 0.000, indicating that the coefficients are statistically significant.

Diagnostic statistics indicate that while the regression model is significant and explains 84.3% of the variance in the dependent variable, further diagnostic tests and potentially

remedial measures may be needed to address any issues identified by the diagnostic statistics.

In summary, the model demonstrates high explanatory power, but it is essential to perform additional diagnostic tests to ensure the robustness of the results and address any potential issues such as heteroscedasticity.

2. Multivariate Model 1

The independent variables chosen for this model are vehicle class, engine size, cylinders, and fuel consumption combined (L/100 km), with CO2 emissions as the dependent variable. Feature engineering was employed to create a list of features to include in the model. Subsequently, the feature matrix (X) and the target variable (y) were created. The data were then split into training and testing sets.

A linear regression model was created and trained using the training data. The trained model was used to make predictions on the test data. The performance of the model was evaluated using Mean Squared Error (MSE) and R-squared (R^2) metrics.

The MSE, which measures the average squared difference between the actual and predicted CO2 emissions, was 343.16. The R^2 value, which indicates the proportion of the variance in the dependent variable that is predictable from the independent variables, was 0.90. This suggests that 90% of the variance in CO2 emissions can be explained by the model.

A plot was constructed to compare the actual versus the predicted CO2 emissions based on the model. The plot shows that the predicted CO2 emissions closely follow the actual CO2 emissions, with most data points lying near the trend line. The strong alignment between the actual and predicted values, along with the high R^2 value, indicates that the model is performing well in predicting CO2 emissions.

In conclusion, the linear regression model effectively predicts CO2 emissions based on the given features, demonstrating strong predictive power with an R^2 value of 0.90 and a relatively low MSE of 343.16.

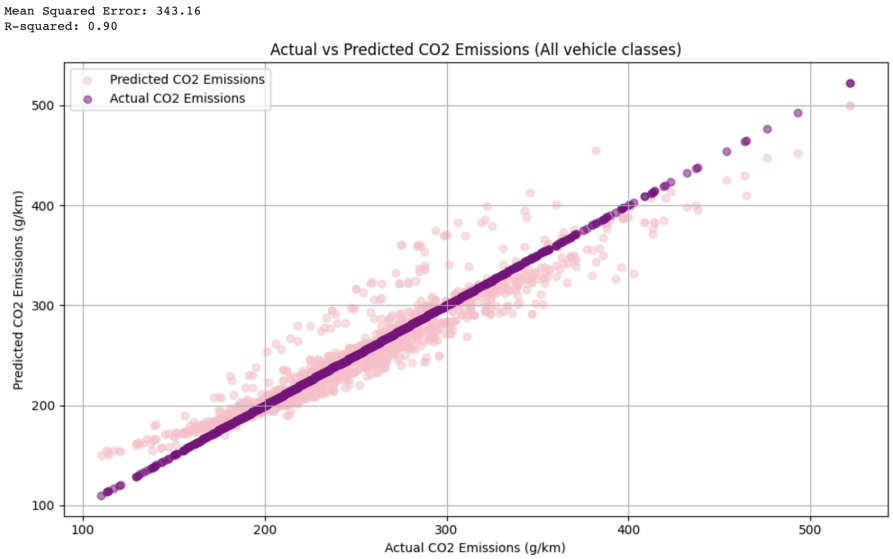


Fig. 10 Multivariate model 1. Each point represents an observation with its actual and predicted CO2 emissions. The trend line shows the ideal scenario where actual values equal predicted values.

- Multivariate Model 2

The independent variables chosen for this model are vehicle class, engine size, cylinders, and fuel consumption combined (L/100 km), with CO2 emissions as the dependent variable. Feature engineering was employed to create a list of features to include in the model. Subsequently, the feature matrix (X) and the target variable (y) were created. The data were then split into training and testing sets.

A linear regression model was created and trained using the training data. The trained model was used to make predictions on the test data. The performance of the model was evaluated using Mean Squared Error (MSE) and R-squared (R^2) metrics.

The MSE, which measures the average squared difference between the actual and predicted CO2 emissions, was 451.34. The R^2 value, which indicates the proportion of the variance in the dependent variable that is predictable from the independent variables, was 0.83. This suggests that 83% of the variance in CO2 emissions can be explained by the model.

A plot was constructed to compare the actual versus the predicted CO2 emissions based on the model. The plot shows that the predicted CO2 emissions closely follow the actual CO2 emissions, with most data points lying near the trend line. The strong alignment between the actual and predicted values, along with the high R^2 value, indicates that the model is performing well in predicting CO2 emissions.

In conclusion, the linear regression model effectively predicts CO2 emissions based on the given features, demonstrating strong predictive power with an R^2 value of 0.83 and a relatively low MSE of 451.34.

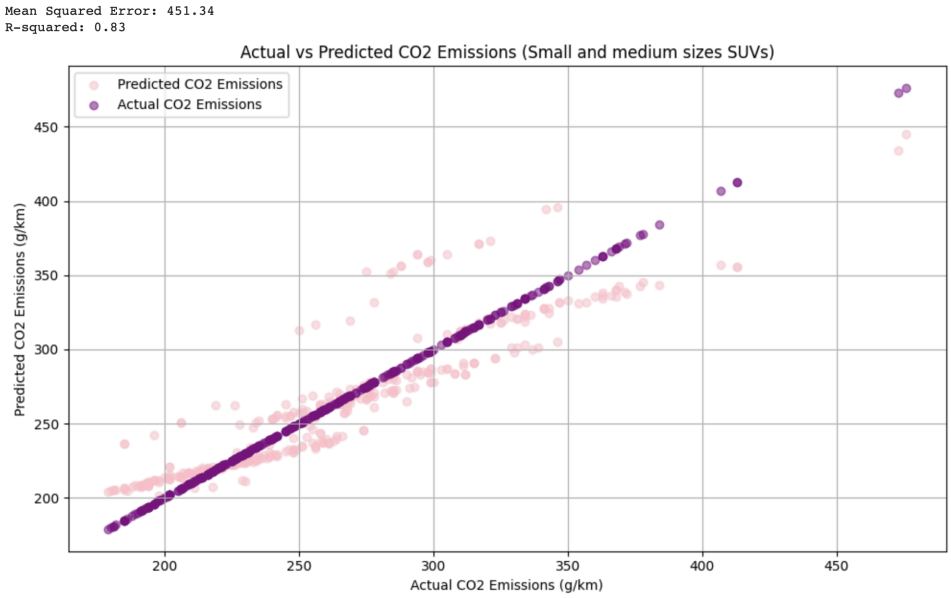


Fig. 11 Multivariate Model 2. Each point represents an observation with its actual and predicted CO2 emissions. The trend line shows the ideal scenario where actual values equal predicted values.

- Multivariate Model 3

The independent variables chosen for this model are engine size, cylinders, fuel consumption combined (L/100 km), and fuel type, with CO2 emissions as the dependent variable. Feature engineering was employed to create a list of features to include in the model. Subsequently, the feature matrix (X) and the target variable (y) were created. The data were then split into training and testing sets.

A linear regression model was created and trained using the training data. The trained model was used to make predictions on the test data. The performance of the model was evaluated using Mean Squared Error (MSE) and R-squared (R^2) metrics.

The MSE, which measures the average squared difference between the actual and predicted CO2 emissions, was 24.13. The R^2 value, which indicates the proportion of the variance in the dependent variable that is predictable from the independent variables, was 0.99. This suggests that 99% of the variance in CO2 emissions can be explained by the model.

A plot was constructed to compare the actual versus the predicted CO2 emissions based on the model. The plot shows that the predicted CO2 emissions closely follow the

actual CO2 emissions, with most data points lying near the trend line. The strong alignment between the actual and predicted values, along with the high R^2 value, indicates that the model is performing exceptionally well in predicting CO2 emissions.

In conclusion, the linear regression model effectively predicts CO2 emissions based on the given features, demonstrating very strong predictive power with an R^2 value of 0.99 and a very low MSE of 24.13.

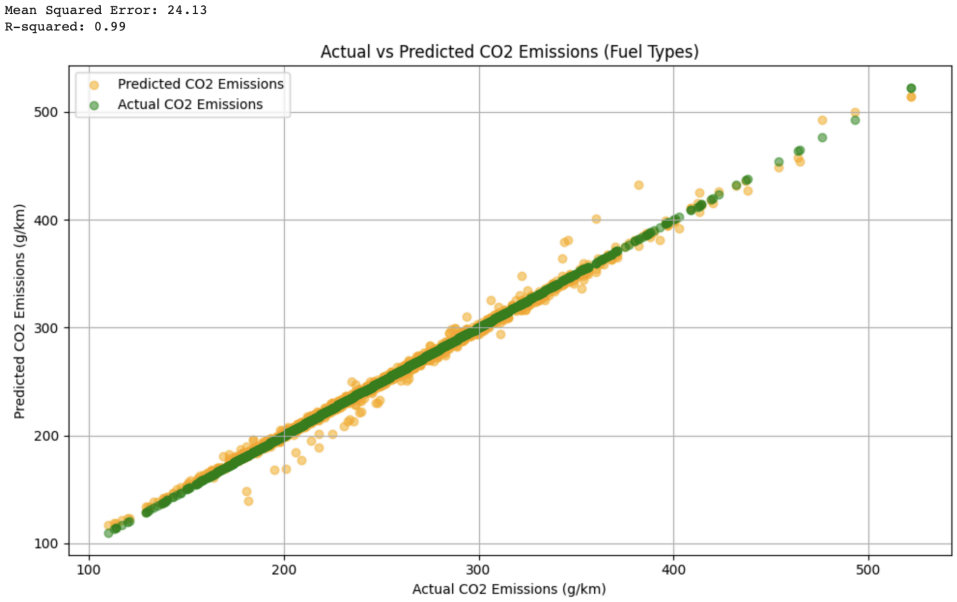


Fig. 11 Multivariate Model 3. Each point represents an observation with its actual and predicted CO2 emissions. The trend line shows the ideal scenario where actual values equal predicted values.

4. Discussion

Business application for the prediction models

One of the business applications of this prediction model is highly relevant for logistics companies aiming to optimize their operations with sustainability in mind. By predicting CO2 emissions based on fuel consumption, supply chain companies that prioritize Sustainable Development Goals (SDGs) can leverage this model to reduce their environmental impact. This model can aid in monitoring and managing emissions more effectively, aligning with their sustainability objectives. Additionally, government agencies can use the model for regulatory compliance and policy-making, while consumer goods companies and retailers can enhance their green logistics strategies. NGOs focused on environmental protection can also benefit by using the model to advocate for lower emissions in the transportation and logistics sectors. Overall, this prediction model serves as a vital tool for organizations committed to sustainable practice.

Another business application can be a car-sharing startup. This newly formed company aims to extract relevant information from the given dataset to make strategic business decisions. Specifically, they need to determine which cars to use for their car-sharing service, which operates solely within cities.

From previous research, they have identified that compact and mid-size vehicles are best suited to their customers' needs based on factors such as car size, available seats, and trunk space. A second, more comprehensive customer research study has confirmed these findings. Additionally, this extended research revealed a strong customer preference for SUVs, driven by emotional appeal rather than practical considerations.

As part of its strategic decision-making, the startup aims to understand the business implications of incorporating small-to-standard-sized SUVs into its fleet compared to sticking with compact and mid-sized vehicles. They are particularly interested in the following comparisons:

- Differences in CO2 emissions between these vehicle types.
- Differences in overall fuel consumption.
- Variations in fuel consumption when comparing different fuel types.

By leveraging the prediction models, the startup can make informed decisions on fleet composition that align with both customer preferences and environmental sustainability goals. This approach will not only enhance customer satisfaction but also contribute to the startup's sustainability objectives by minimizing CO2 emissions and optimizing fuel consumption.

5. Conclusion

In conclusion, the dataset used in this study has proven to be a valuable resource for developing accurate prediction models. The prediction models demonstrated high levels of precision, effectively capturing the relationships between the various vehicle attributes and CO2 emissions. The results of this study highlight the strong linear correlation between Fuel Consumption Combined (miles per gallon) and CO2 Emissions (g/km). These findings underscore the potential of the dataset and the developed models to provide insightful predictions that can inform and optimize strategies for reducing vehicular CO2 emissions.

6. Acknowledgments

For this study, I collaborated with Marius Nickig, whose contributions and insights were invaluable.

Link: https://colab.research.google.com/drive/1QJYkeq8WvZkA7FzUzmgVRYX30K7zFuEW?usp=sharing