

CARBON EMISSION PREDICTION

TEAM MEMBERS:-

1)Varun Mangla-(23MCA062)

Contribution:

- Conducted data preprocessing and exploratory data analysis (EDA).
- Built and trained the **Multiple Linear Regression** model to predict CO₂ emissions.

2)Ashish Kumar Agrawal(23MCA015)

Contribution:

- Handled model evaluation, calculating performance metrics such as MAE, MSE, RMSE, and R-squared.
- Drafted the sections on Experimental Results, Discussion, Conclusions, and Future Work for the report.

DETAILED REPORT

A) Background, Problem, and Objectives

Carbon emissions are a major contributor to global warming and climate change. Given that vehicles are a significant source of CO₂ emissions, accurately predicting these emissions based on vehicle characteristics can guide policy and personal choices to mitigate environmental impacts.

The primary challenge is to predict CO₂ emissions accurately using vehicle characteristics, such as engine size, fuel consumption, and number of cylinders.

OBJECTIVE:-

To develop a predictive model that estimates CO₂ emissions based on vehicle features using regression analysis, thereby assisting in understanding and reducing automobile emissions.

B) Data

#Data Preprocessing

Engine Size(L)	Cylinders	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	Fuel Consumption Comb (mpg)	CO2 Emissions(g/km)
0	2.0	4	9.9	6.7	8.5	33
1	2.4	4	11.2	7.7	9.6	29
2	1.5	4	6.0	5.8	5.9	48
3	3.5	6	12.7	9.1	11.1	25
4	3.5	6	12.1	8.7	10.6	27
...
7380	2.0	4	10.7	7.7	9.4	30
7381	2.0	4	11.2	8.3	9.9	29
7382	2.0	4	11.7	8.6	10.3	27
7383	2.0	4	11.2	8.3	9.9	29
7384	2.0	4	12.2	8.7	10.7	26

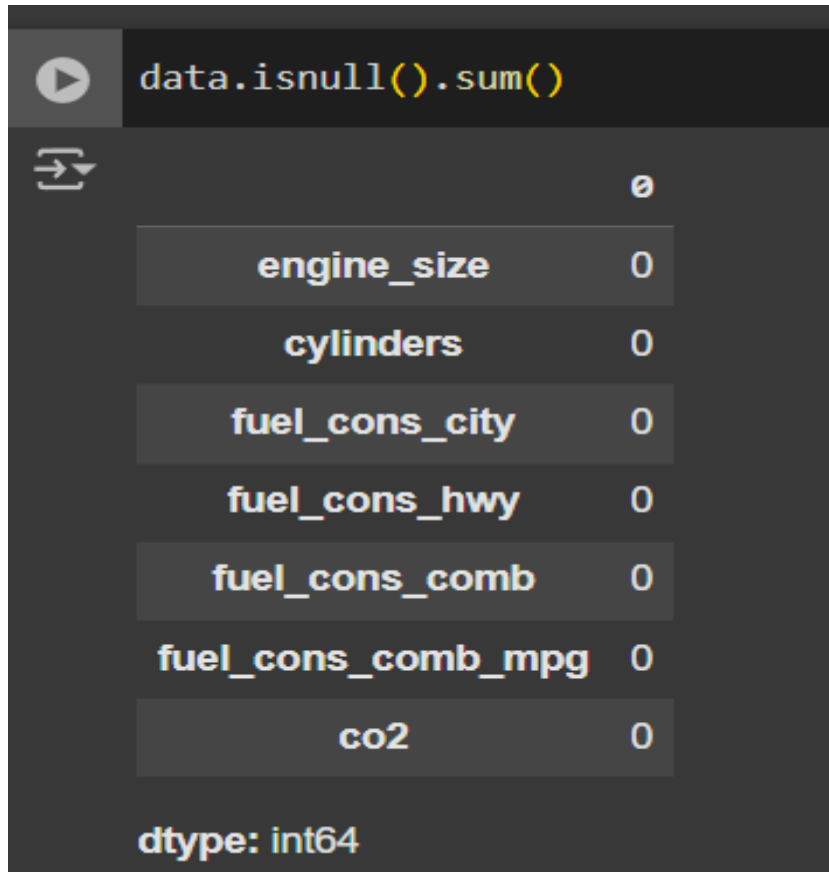
7385 rows x 7 columns

The dataset consists of 7,385 entries and 7 attributes representing vehicle characteristics. The columns include:

1. Engine Size(L): Engine size in liters.
2. Cylinders: Number of cylinders in the engine.
3. Fuel Consumption City (L/100 km): Fuel consumption in city driving conditions (liters per 100 kilometers).
4. Fuel Consumption Hwy (L/100 km): Fuel consumption on highways (liters per 100 kilometers).
5. Fuel Consumption Comb (L/100 km): Combined (city and highway) fuel consumption (liters per 100 kilometers).

6. Fuel Consumption Comb (mpg): Combined fuel consumption in miles per gallon.
7. CO2 Emissions (g/km): Target variable representing CO₂ emissions in grams per kilometer.

After loading the dataset, the **isnull().sum()** method was used to check for missing values. The dataset was found to be clean, with no missing values in any columns, simplifying preprocessing.



```
data.isnull().sum()
```

	0
engine_size	0
cylinders	0
fuel_cons_city	0
fuel_cons_hwy	0
fuel_cons_comb	0
fuel_cons_comb_mpg	0
co2	0

dtype: int64

#Exploratory Data Analysis

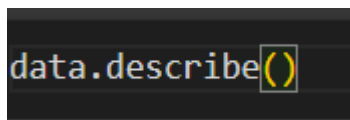
1)Data Overview

The dataset contains 7385 entries and 7 columns, with no missing values. The columns include details about vehicle specifications, such as **engine_size** (float), **cylinders** (integer), and fuel consumption values for **city**, **highway**, and **combined** (all float), as well as combined fuel efficiency in **MPG** (integer) and **CO2 emissions** (integer).

```
print(data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7385 entries, 0 to 7384
Data columns (total 7 columns):
#   Column                Non-Null Count  Dtype  
---  -
0   engine_size           7385 non-null  float64
1   cylinders              7385 non-null  int64   
2   fuel_cons_city        7385 non-null  float64
3   fuel_cons_hwy         7385 non-null  float64
4   fuel_cons_comb        7385 non-null  float64
5   fuel_cons_comb_mpg    7385 non-null  int64   
6   co2                   7385 non-null  int64   
dtypes: float64(4), int64(3)
memory usage: 404.0 KB
None
```

2)Statistical Summary of the dataset



	Engine Size(L)	Cylinders	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	Fuel Consumption Comb (mpg)	CO2 Emissions(g/km)
count	7385.000000	7385.000000	7385.000000	7385.000000	7385.000000	7385.000000	7385.000000
mean	3.160068	5.615030	12.556534	9.041706	10.975071	27.481652	250.584699
std	1.354170	1.828307	3.500274	2.224456	2.892506	7.231879	58.512679
min	0.900000	3.000000	4.200000	4.000000	4.100000	11.000000	96.000000
25%	2.000000	4.000000	10.100000	7.500000	8.900000	22.000000	208.000000
50%	3.000000	6.000000	12.100000	8.700000	10.600000	27.000000	246.000000
75%	3.700000	6.000000	14.600000	10.200000	12.600000	32.000000	288.000000
max	8.400000	16.000000	30.600000	20.600000	26.100000	69.000000	522.000000

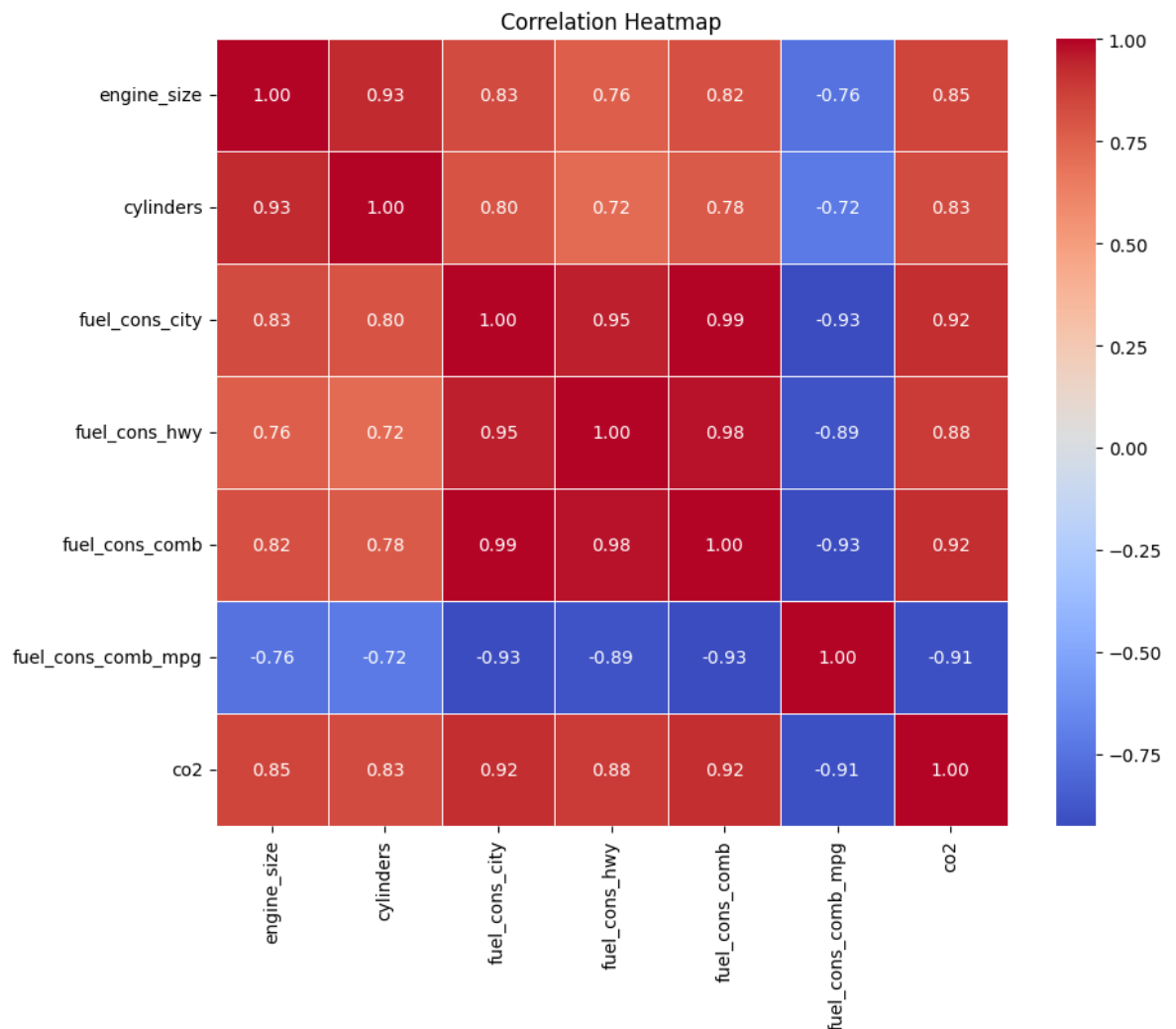
Using describe(), the following insights were obtained:

1. **Engine Size (engine_size):** Ranges from 1.0 to 8.4 liters, with a mean of around 3.3 liters. This distribution suggests a variety of vehicle types, from compact to larger-engine vehicles.
2. **Cylinders (cylinders):** The cylinder count varies from 3 to 16, with a mean of approximately 6. This feature helps capture engine design complexity, affecting fuel consumption and emissions.
3. **fuel_cons_city** ranges from 4.6 to 30.9 L/100 km, with a mean of 13.3.
4. **fuel_cons_hwy** ranges from 4.9 to 20.5 L/100 km, with a mean of 9.4.
5. **fuel_cons_comb** is typically used as a single metric for fuel consumption and has a mean of 11.1 L/100 km.
6. **CO₂ Emissions (CO2_emissions):** The target variable has a mean of 256 g/km, with values ranging from 96 to 522 g/km. This wide range reflects differences in fuel efficiency and engine size across vehicle types.

3)Correlation Analysis

Correlation analysis reveals relationships between features and the target variable, helping identify influential predictors and potential feature redundancies.

- **fuel_cons_comb** and **engine_size** exhibit strong positive correlations with **CO2_emissions**, indicating that vehicles with larger engines and higher fuel consumption produce more emissions. These features are likely crucial for the prediction model.
- **fuel_cons_city**, **fuel_cons_hwy**, and **fuel_cons_comb** are highly correlated with each other (above 0.8), suggesting redundancy. Among these, **fuel_cons_comb** serves as a representative feature, summarizing fuel consumption in both city and highway scenarios.
- **Cylinders**: Also shows a moderate positive correlation with **CO2_emissions**. This is consistent with the idea that more cylinders usually mean a larger engine, leading to higher fuel consumption and emissions.



Based on the correlation analysis, **fuel_cons_comb** was selected as the primary fuel consumption metric, while **fuel_cons_city** and **fuel_cons_hwy** were excluded to prevent multicollinearity.

C)METHODOLOGY

Feature Scaling

The features were standardized using StandardScaler, which normalizes data to improve model performance, particularly for algorithms sensitive to scale differences.

Data Splitting

The dataset was split into training and testing sets (80-20 split).

Model Choice

Multiple Linear Regression was selected as a suitable method due to its interpretability and effectiveness in understanding linear relationships between features and the target variable.

D)Experimental Results and Discussion

After training the Multiple Linear Regression model on the training set, predictions were made on the test set.

Evaluation Metrics

- **Mean Absolute Error (MAE): 13.56 :-**This indicates that, on average, the model's predictions are off by about 13.56 units from the actual values. MAE gives a sense of the average magnitude of errors without considering their direction.
- **Mean Squared Error (MSE): 421.59:-**The MSE is a measure of the average squared difference between the predicted and actual values. A higher value suggests larger errors, and in this case, 421.59 reflects a somewhat higher level of error.
- **Root Mean Squared Error (RMSE): 20.53:-**RMSE is the square root of MSE, providing an error metric that is in the same unit as the target variable. It indicates that, on average, the model's predictions are off by 20.53 units.
- **R-squared (R^2): 0.88:-** R^2 represents the proportion of variance in the dependent variable that is explained by the model. A value of 0.88 suggests that the model explains 88% of the variability in the target, indicating a good fit.

Mean Absolute Error: 13.56
Mean Squared Error: 421.59
Root Mean Squared Error: 20.53
R-squared: 0.88

E) Conclusions and Future Work

Conclusions

The linear regression model effectively predicted CO₂ emissions, highlighting key relationships between features and emissions. Notably, engine size and fuel consumption had strong positive correlations with emissions, suggesting larger engines and higher fuel usage increase emissions.

Future Work

Future work could explore complex models such as neural networks or random forests, which might capture non-linear relationships. Additionally, incorporating more diverse datasets could improve model generalization.

F) BIBLIOGRAPHY

1) <https://www.kaggle.com/>

This website provides a wide range of datasets. The **CO₂ Emissions Dataset** on **Kaggle** contains vehicle data such as engine size, fuel consumption, and CO₂ emissions, useful for analyzing the relationship between vehicle specifications and environmental impact. This dataset helps in predicting CO₂ emissions based on car attributes.

2) <https://scikit-learn.org/>

Scikit-learn is a popular Python library for machine learning, providing tools for model building and evaluation. It includes functions to calculate performance metrics like **MAE**, **MSE**, **RMSE**, and **R-squared**, which were used to evaluate the model's predictions in this analysis.