PREDICTIVE MODELING OF CO2 EMISSIONS FROM VEHICLES

Group B

TABLE OF CONTENTS

01.INTRODUCTION

CO_2 emissions are a significant contributor to greenhouse gas emissions, which are known to be a primary driver of global warming and climate change. The transportation industry is a major source of CO_2 emissions, with vehicles powered by internal combustion engines being the primary culprits. The amount of CO_2 emitted by a vehicle depends on various factors such as the make and model of the vehicle, its fuel efficiency, driving conditions, and technology advancements.

MAIN OBJECTIVES

01

EDA

To identify factors that affect the CO2 emissions from vehicles.

02

MODEL

To construct a model to predict the CO2 emissions from vehicles based on the most influential predictors.

03

DATA PRODUCT

To design a website to accurately predict the CO2 emissions from vehicles when the relevant factors are provided.

02

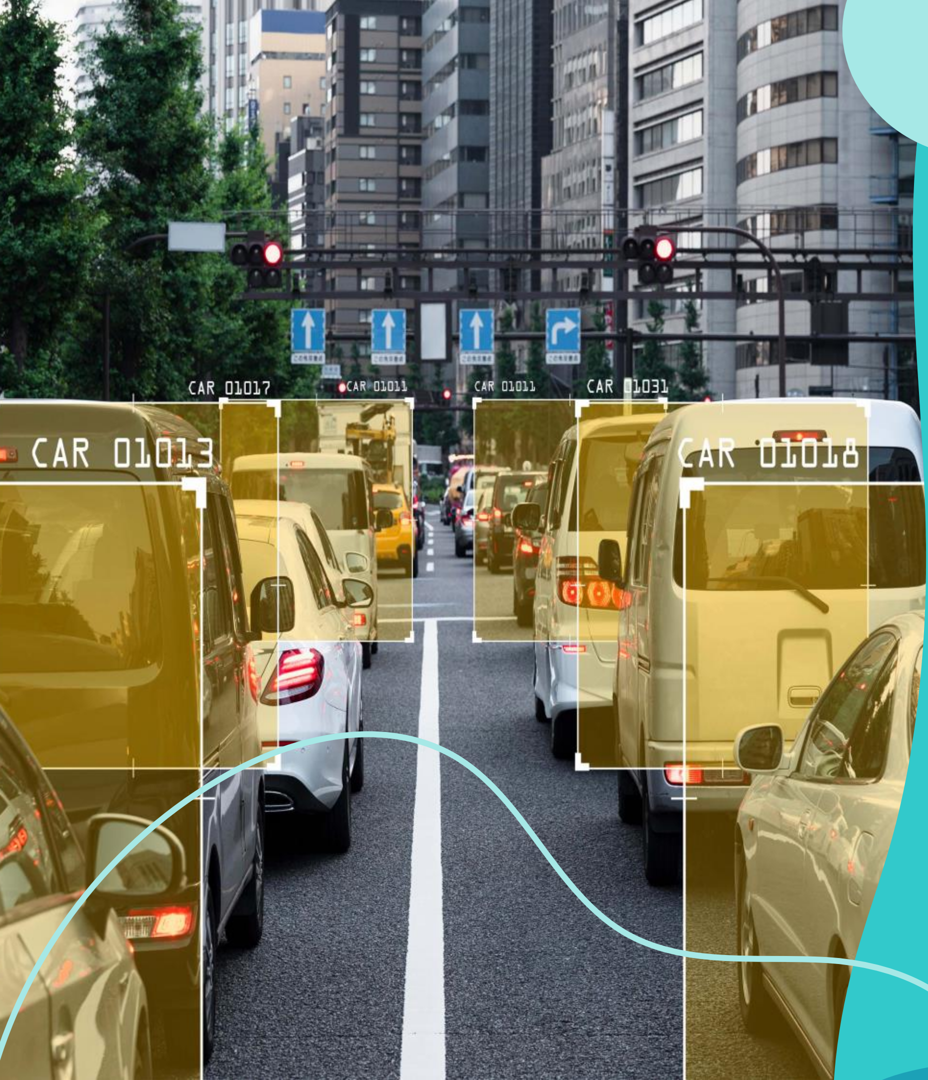
DESCRIPTION OF THE DATASET

DESCRIPTION OF THE DATASET

Variable	Description	Type
Make	Company of the vehicle	Qualitative
Model	Car Model	Qualitative
Vehicle Class	Class of vehicle depending on their utility, capacity, and weight	Qualitative
Engine Size (L)	Size of engine used in Liter	Quantitative
Cylinders	Number of cylinders	Quantitative
Transmission	Transmission type with number of gears	Qualitative
Fuel Type	Type of Fuel used	Qualitative
Fuel Consumption City (L/100 km)	Fuel consumption in city roads (L/100 km)	Quantitative
Fuel Consumption Hwy (L/100 km)	Fuel consumption in highways (L/100 km)	Quantitative
Fuel Consumption Comb (L/100 km)	The combined fuel consumption (55% city, 45% highway) is shown in L/100 km	Quantitative
Fuel Consumption Comb (mpg)	The combined fuel consumption in both city and highway is shown in mile per gallon(mpg)	Quantitative
CO2 Emissions(g/km)	**The tailpipe emissions of carbon dioxide (in grams per kilometer) for combined city and highway driving**	**Quantitative**

Kaggle obtained Dataset contains 7385 records with

- 5 qualitative variables
- 7 quantitative variables



03

DATA PRE-PROCESSING

Data Pre - Processing

Duplicated Entries

- 1103 duplicated entries were removed.
- "N" from Fuel Type & "France" from make_country were eliminated as there are only 2 and 1 records exist.

Now there are 6279 data excluding duplicates and those records

Car Make-Country Mapping

'Make' column contained 42 categories representing the vehicle's company with country

'make_country' column contained 8 categories representing vehicle's origin country only

Transmission Simplification

'Transmission' contained 27 categories representing transmission types with gear information

'transmission_general' classifies vehicles as either 'Manual' or 'Automatic,'

Vehicle Category Mapping

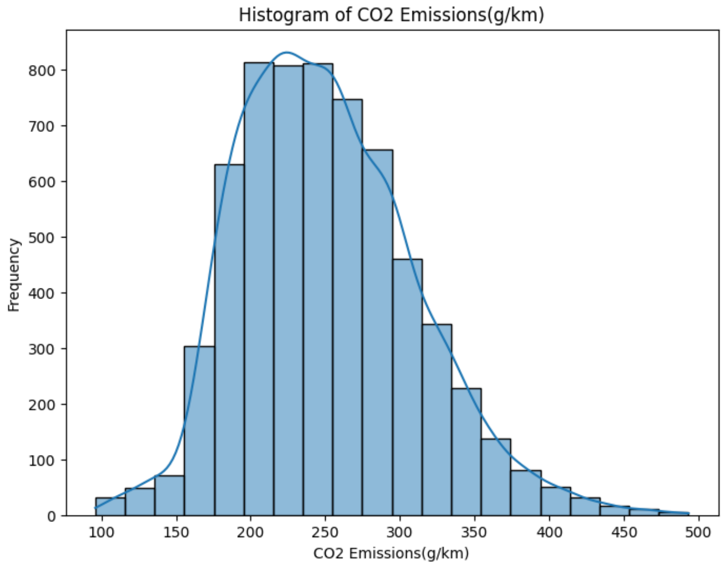
'Vehicle Class' contained 16 categories representing the vehicle's class with their utility, capacity, weight information

'vehicle' column contained 6 categories representing vehicle class only

04

DESCRIPTIVE ANALYSIS

Response Variable

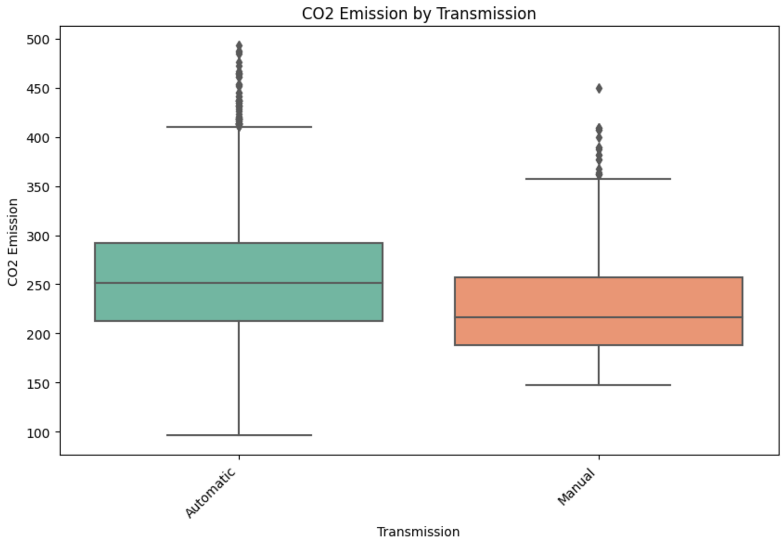


- Majority of vehicles tend to exhibit relatively lower CO2 emission values
- Right-skewness of the histogram suggests the presence of a tail towards higher emission values,

  → There are still a notable number of vehicles producing higher CO2 emissions

CO2 Emission by Transmission

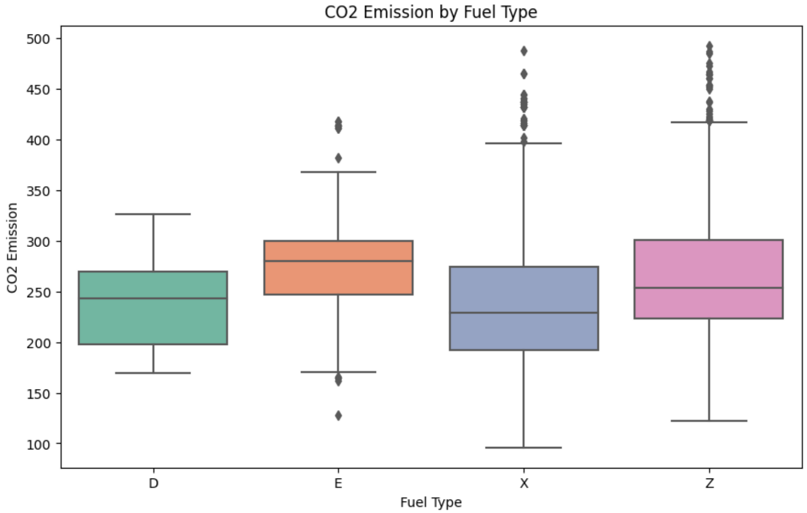
Vehicles with automatic transmissions exhibit a **wider spread** of CO2 emission range
and
a **higher median.**



The efficiency gains from automatic transmissions' can contribute to their relatively lower emissions in certain scenarios.

However,
contextual factors such as vehicle weight, type, engine tuning, and driver behavior also play significant roles in influencing emission levels.

CO2 Emission by Fuel type



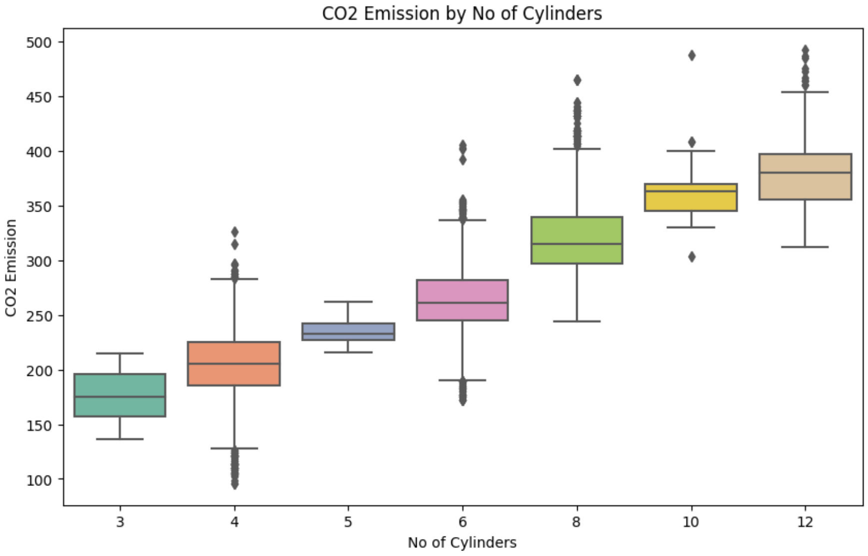
Type "E" (Ethanol) has the highest median CO2 emissions

This could be due to,
including ethanol's lower energy density,
distinct combustion characteristics,
and
potential tuning variations in vehicles designed for ethanol blends.

CO2 Emission by Number of Cylinders

As the number of cylinders rises, CO2 emissions seems to be increase.

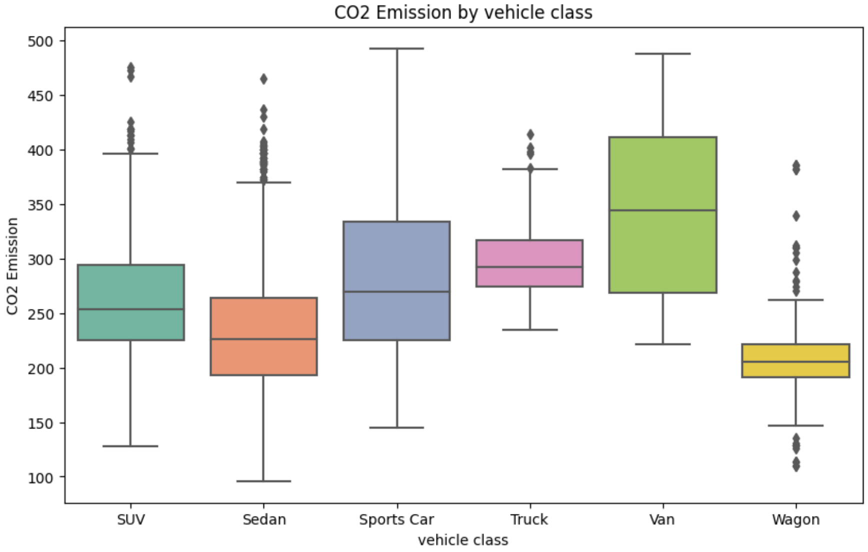
Larger engines with more cylinders inherently require more fuel to operate, leading to higher fuel consumption and subsequently elevated CO2 emissions.



CO2 Emission by Vehicle Class

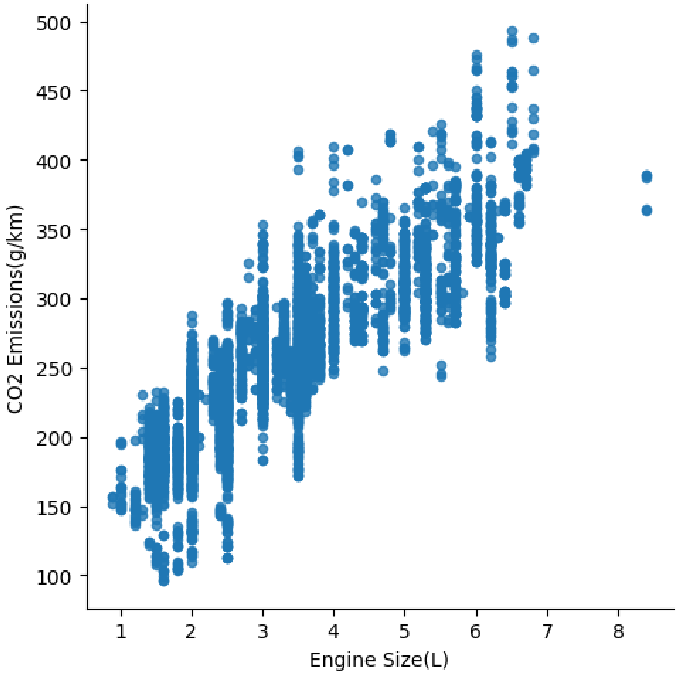
Vans and Trucks, exhibit the highest median CO2 emission levels.

- Larger size and typically higher weight
- Trucks are often owing to their heavy-duty nature and utilization for commercial purposes



Sports cars, known for their performance-oriented design and potentially larger engines, also present a notable median CO2 emission level.

CO2 Emission Vs Engine Size

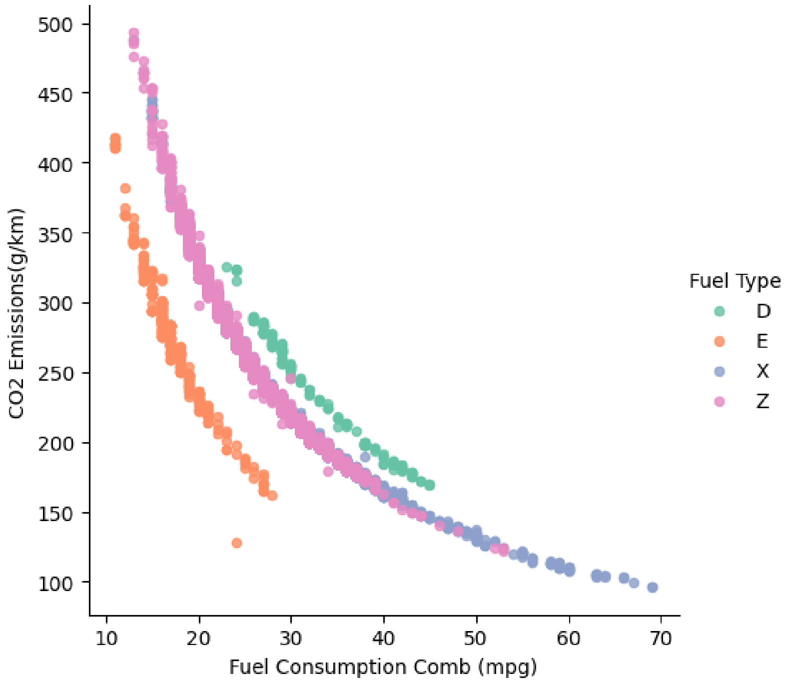


- Correlation between increased engine size and higher CO2 emissions can be attributed to the intricate relationship between engine capacity, fuel consumption, and power generation.
- Larger engine sizes lead to higher CO2 emissions due to increased fuel consumption required for greater power and often, heavier vehicle masses.
- This highlights the need for fuel-efficient technologies and alternative fuels to reduce the environmental impact."

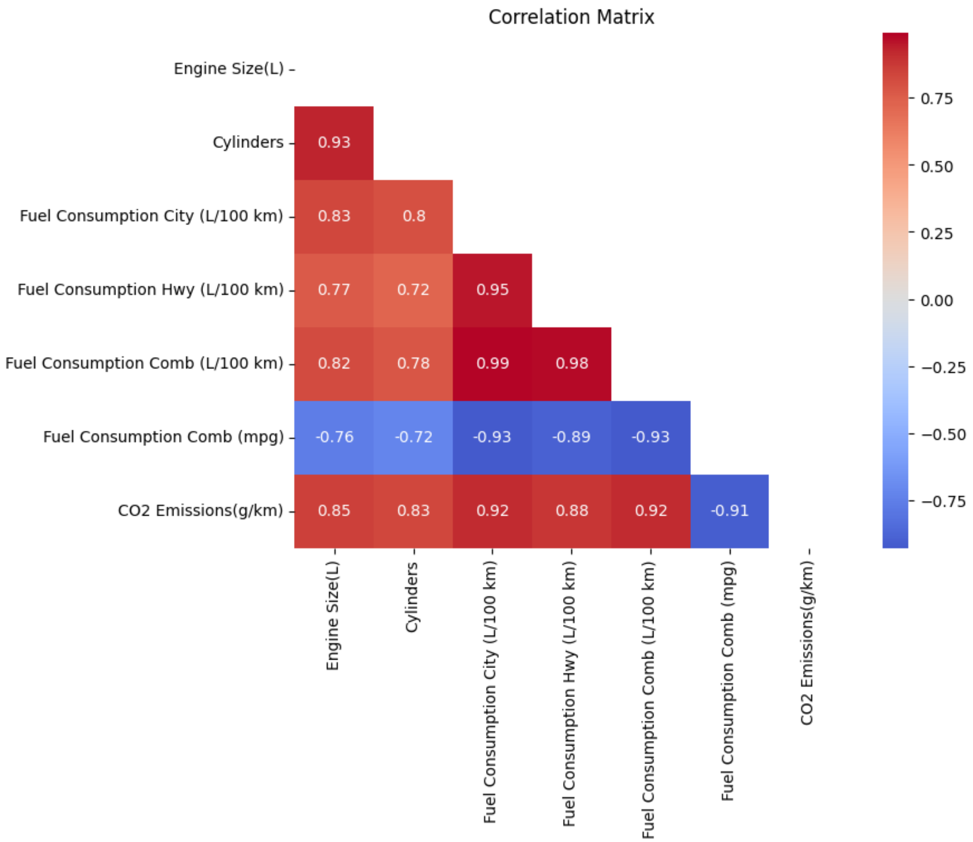
Fuel Efficiency and CO2 Emissions

- As fuel consumption decreases and mpg increases, vehicles show lower CO2 emissions, aligning with energy conservation principles.

- Higher mpg values signify better fuel efficiency, enabling vehicles to travel farther on less fuel, resulting in cost savings and reduced environmental impact in terms of CO2 emissions.



Association b/w Continuous Variables



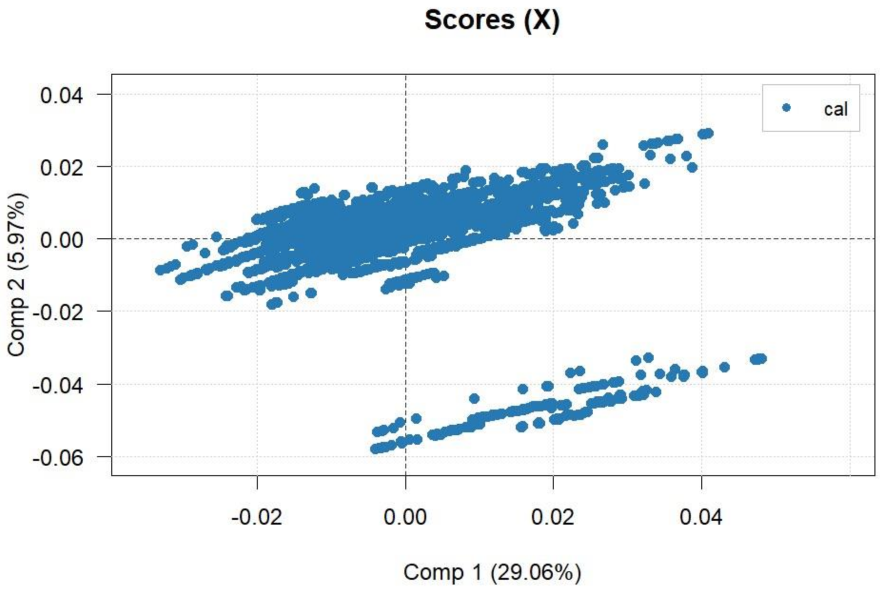
- **High Correlation: Fuel consumption variables (city, Hwy, combined) show a strong correlation.**

  For in-depth analysis, we'll concentrate on the Fuel Consumption Comb (mpg) variable.

- **High Correlation: Engine Size and Fuel Consumption variables show strong positive correlation.**

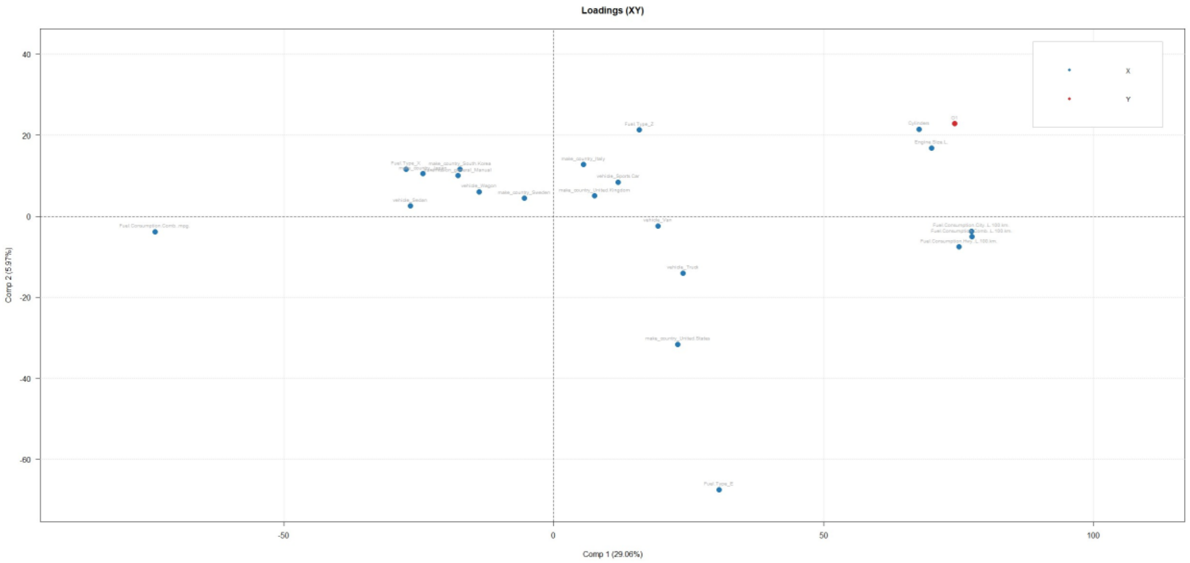
  Significant Multicollinearity issue exists.

Partial Least Squares – Scores Plot



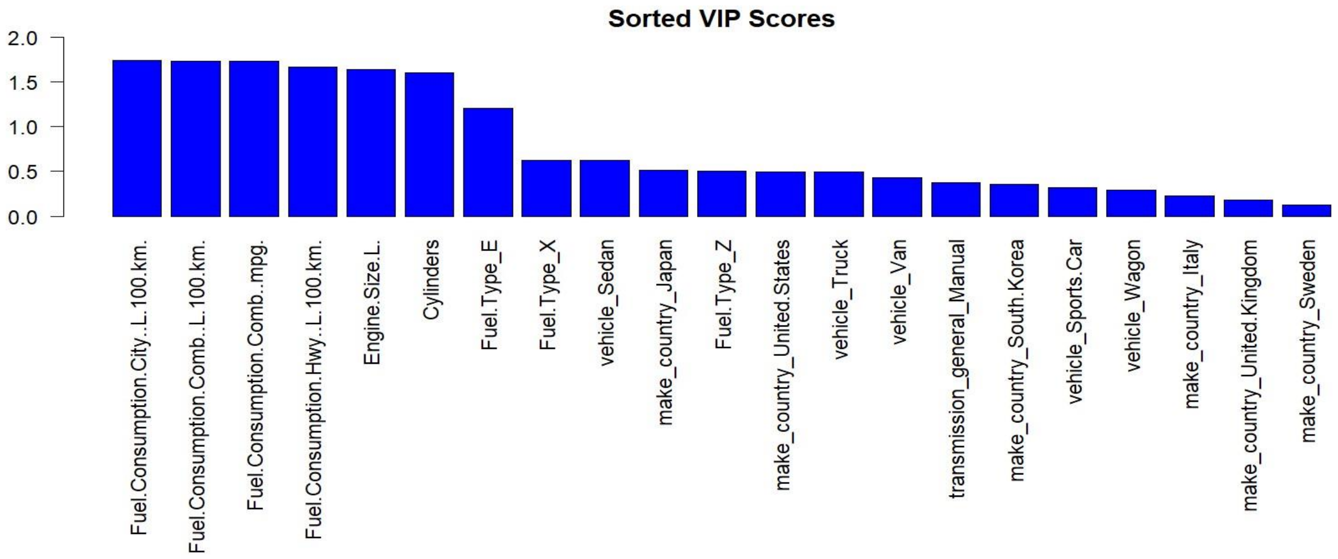
- **Purpose:** Employed PLS to uncover clusters within observations.
- **Observation Clusters:** Distinguished clusters observed.
- **Variation Explained:** Axes in the plot accounted for less than 70% (nearly 35%) of total variation in both predictor and response variables, suggesting some limitations in capturing data relationships.

Partial Least Squares – Loadings Plot



- **Purpose:** Identify strong correlations among predictors.
- **Observations:** Certain predictors show strong correlations with the response variable (Y).
- **Orthogonal Relationships:** Some predictors exhibit orthogonal relationships with the response, indicating no direct correlation.
- **Near Origin:** Certain predictors are located close to the origin, suggesting their lack of importance.
- **Variable Clusters:** Clusters of variables suggest the potential for deriving new variables or selecting important ones using feature importance.

Partial Least Squares – VIP Plot



- **Purpose:** Identifies input variables with the strongest relationships to the response variable, guiding variable selection for the model.
- **Model Simplification:** VIP plots also highlight redundant or unimportant variables that can be removed to simplify the model.
- **Valuable Tool:** A useful tool for optimizing the model's performance and reducing complexity.

05

ADVANCED ANALYSIS

EVALUATION METRICS

MSE

Mean Squared Error

- the average squared difference between predicted and actual values

R^2

R - Squared

- determines the proportion of variance in the dependent variable that can be explained by the model.
- shows how well the data fit the regression model (the goodness of fit)

RMSE

Root Mean Squared Error

- measures the average difference between values predicted by a model and the actual values.
- It provides an estimation of how well the model is able to predict the target value (accuracy)

MAPE

Mean Absolute Percentage Error

- the average percentage difference between predicted and actual values.

PERFORMANCE ON THE FULL MODEL

Model	Train			Test		
	RMSE	MAPE	R^2	RMSE	MAPE	R^2
Linear Regression	4.96	1.23%	0.9929	5.29	1.31%	0.9922
Ridge Regression	4.99	1.24%	0.9928	5.26	1.31%	0.9923
Lasso Regression	9.35	2.36%	0.9748	9.33	2.43%	0.9757
Random forest	1.68	0.40%	0.9992	3.68	0.40%	0.9962
XG Boosting	1.75	0.54%	0.9991	3.60	0.88%	0.9964

All models exhibited strong accuracy levels on both the training and testing datasets, indicating their consistent and reliable performance.

PERFORMANCE ON THE REDUCED MODEL

Model	Train			Test		
	RMSE	MAPE	R2	RMSE	MAPE	R2
Linear Regression	**15.92**	**4.30%**	**0.927**	**15.89**	**4.39%**	**0.9295**
Ridge Regression	**15.92**	**4.30%**	**0.927**	**15.89**	**4.39%**	**0.9297**
Lasso Regression	**17.20**	**4.59%**	**0.9148**	**17.31**	**4.76%**	**0.9163**
Random forest	**3.73**	**1.09%**	**0.996**	**4.31**	**1.24%**	**0.9948**
XG Boosting	**3.67**	**1.08%**	**0.9961**	**4.89**	**1.25%**	**0.9933**

To build a parsimonious model, certain variables with lower importance, as indicated by the Partial Least Squares (PLS) importance plot, have been intentionally removed.

Remaining predictors

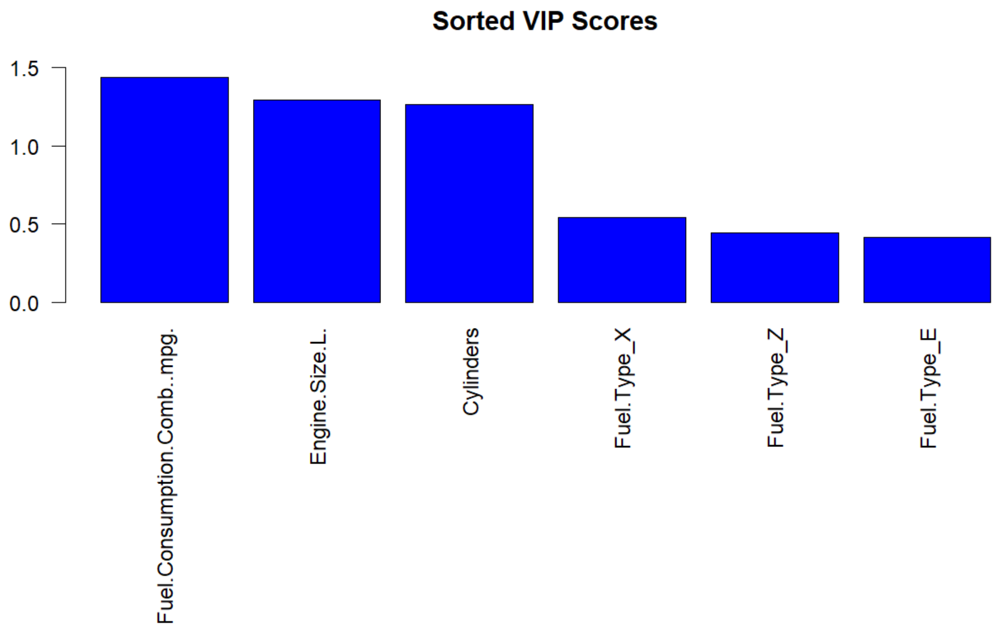
- Engine size
- Cylinders
- Fuel Consumption Comb (mpg)
- Fuel Type_E
- Fuel Type_X
- Fuel Type_Z

MOST IMPORTANT FACTORS

Predictors of the parsimonious model was taken from the Partial Least Squares VIP plot,

<u>Important Predictors</u>

1. Fuel Consumption Comb (mpg)
2. Engine size (in L)
3. Cylinders (no. of cylinders)
4. Fuel Type_X (Regular gasoline)
5. Fuel Type_Z (Premium gasoline)
6. Fuel Type_E (Ethanol (E85))



SUMMARY

- Random Forest was selected as the best model that a good fit for the data, Considering R^2, RMSE,MAPE and the risk of overfitting.

- Fuel Consumption Comb (mpg), Engine size (L), no. of Cylinders, Fuel Type_X (Regular gasoline), Fuel Type_Z (Premium gasoline), Fuel Type_E (Ethanol (E85)) play an important role in prediction of the CO2 emissions for combined city and highway driving.

06

DATA PRODUCT

A website, accurately predicts the CO2 emissions (in grams per kilometer) for combined city and highway driving

THANKS!

OUR TEAM
s15057 - Lakni Kodituwakku
s15067 - Gayani Pathirana
s15080 - Chamodi Siriwardhana
s15089 - Sanjani Wickramasinghe