**Component 2 – Fuel Consumption Rating**

**Abstract**

Fuel consumption and CO2 emissions are related because burning fuel releases carbon dioxide into the atmosphere. The more fuel a vehicle consumes, the more CO2 it will emit. As such, measures of fuel consumption, such as miles per gallon or litres per 100 kilometres, are often used as indicators of a vehicle's environmental impact. Reducing fuel consumption can help to reduce CO2 emissions and mitigate the effects of climate change.

**1.0. Introduction**

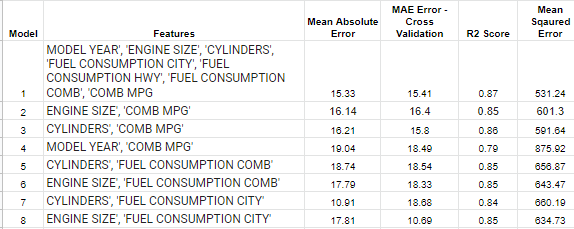
Linear regression is a statistical method that can be used to model the relationship between two variables. Linear regression models will be used to predict CO2 emissions based on the fuel consumption of a vehicle, as well as other factors using a dataset from 2010 to 2014.

**2.0. Methodology**

1. Prepare the data: This involves cleaning and pre-processing the data to make it suitable for training the model. This may include handling missing values, scaling the data, and encoding categorical variables.
2. The data is typically split into training, validation, and test sets. The training set is used to fit the model, the validation set is used to evaluate the model's performance and tune its hyperparameters, and the test set is used for the final evaluation of the model's performance.
3. A machine learning model is chosen based on the specific goals of the analysis and the characteristics of the data.
4. Train the model: The model is trained on the training set using a fitting algorithm. This involves adjusting the model's parameters to minimize the error in the training data.
5. The model's performance is evaluated on the validation set using a metric that measures how well the model's predictions match the true values.
6. If the model's performance on the validation set is not satisfactory, its hyperparameters can be adjusted to improve its performance.
7. Evaluate the final model's performance.

**3.0. Results**

**3.1. Exploratory Data Analysis**

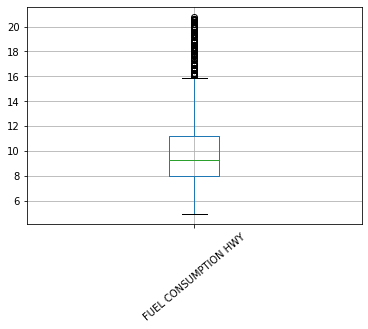
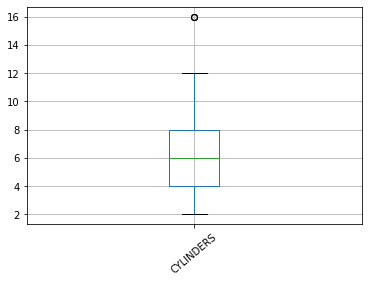
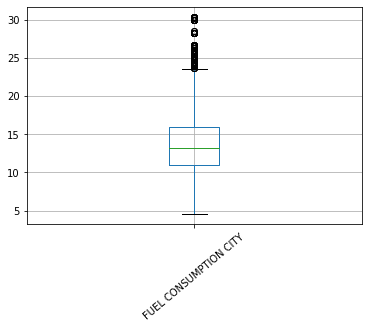


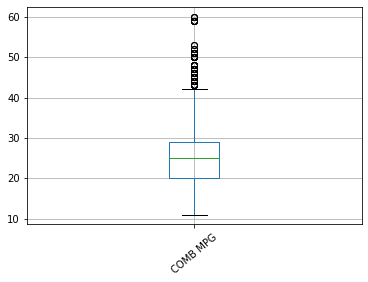
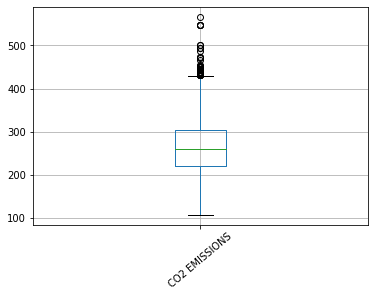
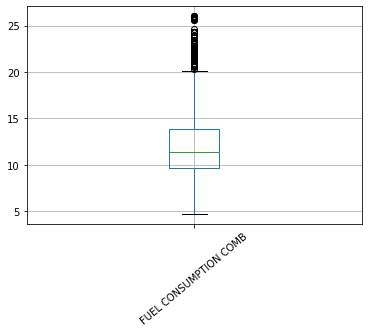
Based on the table above, it appears that the best model is model 1. It has the lowest mean absolute error and mean squared error and the highest R-squared score. This indicates that it is able to make the most accurate predictions on the target variable, while also being the best fit for the data

A strong correlation exists between CO2 emissions and COMB MPG and this closely followed by Fuel Consumption COMB and Fuel Consumption City

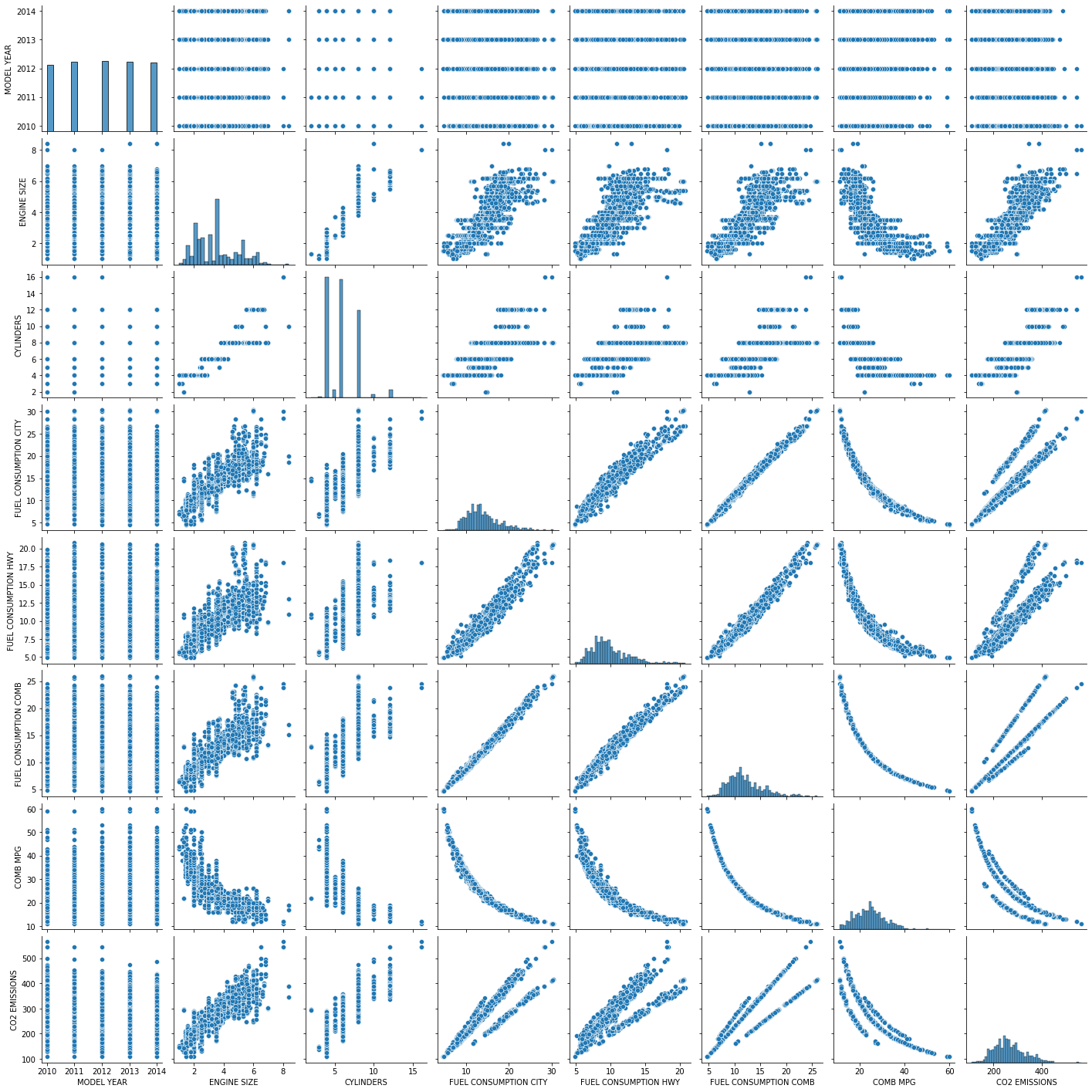
A picture containing chart

Description automatically generated

**Box Plot**

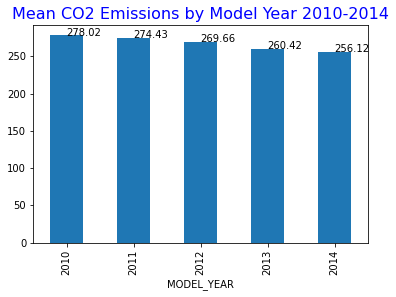


**PairPlot**



**3.3 CO2 emission from 2020-2014**

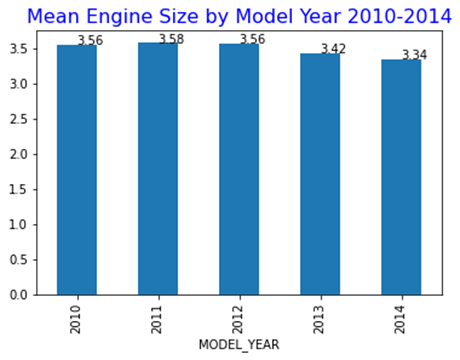
There appears to be only a slight improvement in C02 emissions between 2010 to 2014. The emission was gradually decreasing each year and the least emissions occurred within the year 2013 (mean = 260.42) and 2014 had the least emissions occurrence (mean = 256.12)



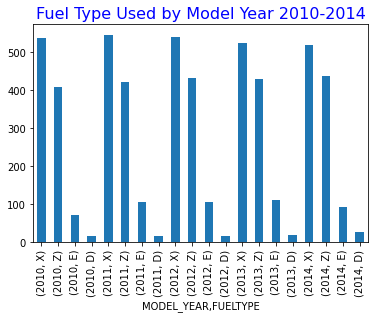
Text

Description automatically generated

Further analysis was done on the attributes that contribute to a car's emission of CO2 such as the engine type and the fuel type used in this period on what might have caused the slight improvements.

Table

Description automatically generated

The engine size used over the time period reduced in size for each year and the year 2014 which had the least CO2 emission had the smallest engine size used (engine size mean = 3.34).

There is a slight decrease in the use of fuel type X over the years, and the use of Z stayed the same.

**3.4. Best Performing categorical best in classifying the dataset**

The fuel type variable (X) performed best in classifying the dataset. The calculated proportions of each category in the data, and the output indicates that the X category is the most prevalent, with a proportion of 49.1%, while the D category is the least prevalent, with a proportion of 1.7%.

**3.5. Overfit**

The comparison of the Cross Validation score and the mean shows that the CV score is higher than the mean and this depicts that the model does not overfit the data. This comparison between the models shows that model 1, 2, and 3,4 overfits and model 5, 7, and 8 does not.

**3.6. Performance Measures**

Mean squared error (MSE) is a commonly used performance measure for regression models. It represents the average squared difference between the predicted values and the true values. A lower value of MSE indicates a better fit of the model to the data.

Mean absolute error (MAE) is another performance measure for regression models. It represents the average absolute difference between the predicted values and the true values. Like MSE, a lower value of MAE indicates a better fit of the model to the data.

R-squared (R²) is a performance measure for regression models that indicates the proportion of the variance in the dependent variable that can be explained by the model. A higher value of R² indicates a better fit of the model to the data.

Adjusted R-squared is a modified version of R-squared that adjusts for the number of predictors in the model. It takes into account the complexity of the model.

Root mean squared error (RMSE) checks the performance measure for regression models by calculating the square root of the average of the squared differences between the predicted values and the true values. Like MSE, a lower value of RMSE indicates a better fit of the model to the data.

**3.7. Models Deployment**

Model 8 also has the best performing model, its mean absolute error is the lowest, mean squared error and a very high R-squared value and does not overfit.

Table

Description automatically generated

The precision of the model is also high, with an average of 0.91 across all classes. In this case, the model has a high precision for all classes, which indicates that it is making accurate predictions.

**3.8.Internal and external evaluation metrics**

The DB index (0.39) and Sil index (0.34) indicate that the clusters are relatively distinct from each other, which suggests that the clustering algorithm has performed well.

**Conclusion**

A linear regression model was designed to make predictions about the CO2 emissions of a vehicle based on its fuel consumption and other characteristics. The accuracy and quality of these models were checked and the ability of the model to capture the underlying relationship between fuel consumption and CO2 emissions.