Jahangirnagar University

Course Code: WM-ASDS04
Course Title: Introduction to Data Science with Python

Final Project report on,

“Predicting CO_2 Emission from Vehicles Using Multiple Linear Regression Model”

Submitted by,

Nafisa Tabassum
Student ID: 20229035

&

A.F.S. Ahad Rahman Khan
Student ID: 20229005

Date of Submission: 21.04.2023

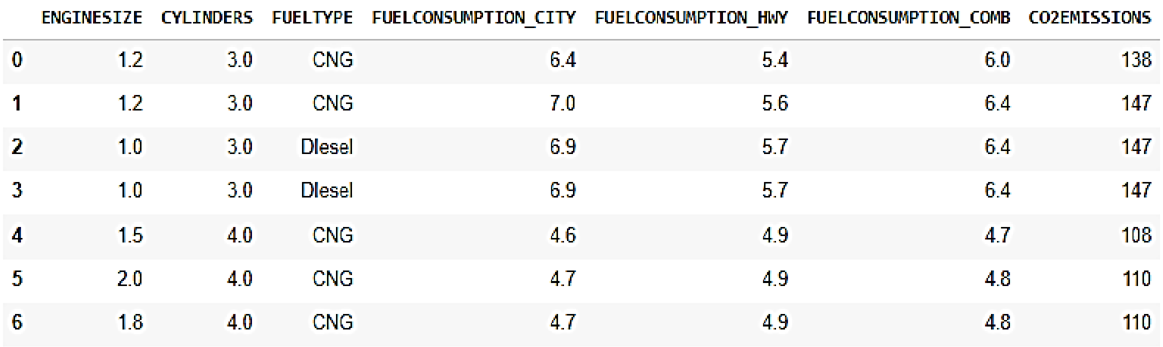
Predicting CO_2 Emission from Vehicles Using Multiple Linear Regression Model

1. Objective

The main goal of this report is to provide insights into the relationships between engine size, cylinder count, fuel consumption, and CO2 emissions, and to develop models that can accurately predict CO2 emissions in vehicles.

2. Description of Raw Dataset

Following table shows first few rows of the raw dataset that is used for the analysis:



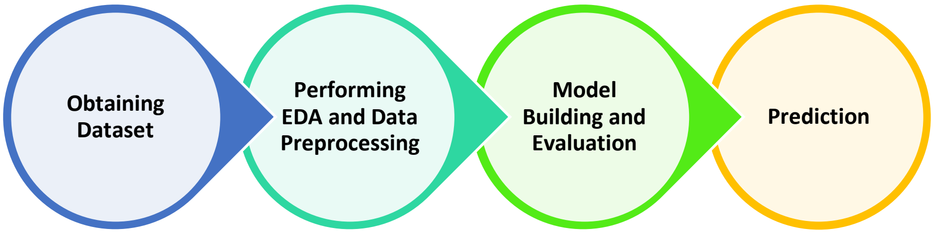
	ENGINESIZE	CYLINDERS	FUELTYPE	FUELCONSUMPTION_CITY	FUELCONSUMPTION_HWY	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	1.2	3.0	CNG	6.4	5.4	6.0	138
1	1.2	3.0	CNG	7.0	5.6	6.4	147
2	1.0	3.0	Diesel	6.9	5.7	6.4	147
3	1.0	3.0	Diesel	6.9	5.7	6.4	147
4	1.5	4.0	CNG	4.6	4.9	4.7	108
5	2.0	4.0	CNG	4.7	4.9	4.8	110
6	1.8	4.0	CNG	4.7	4.9	4.8	110

Here, The Data Frame contains 1067 entries with 7 columns, out of which 5 are of float64 type, 1 is of int64 type, and 1 is of object type. The "CYLINDERS", "FUELTYPE", "FUELCONSUMPTION_CITY", "FUELCONSUMPTION_COMB", and "CO2EMISSIONS" columns have non-null values, while "CYLINDERS", "FUELTYPE", "FUELCONSUMPTION_CITY", and "FUELCONSUMPTION_COMB" columns have missing/null values.

The fuel consumption ratings for city and highway driving are indicated in units of litres per 100 kilometres (L/100 km), while the combined rating, which reflects a 55% city and 45% highway driving mix, is also expressed in L/100 km. Additionally, the amount of carbon dioxide emitted by a vehicle's tailpipe, based on both city and highway driving, is measured in grams per kilometre.

3. Methodology:

Following flow-chart can be developed for the methodology of the task:


Obtaining Dataset
Performing EDA and Data Preprocessing
Model Building and Evaluation
Prediction

3.1 Obtaining Dataset: Data was collected from a publicly available dataset of used cars from Kaggle with some modifications according to the domain knowledge.

3.2 Performing EDA and Data Pre-processing:

The following steps were performed for Exploratory Data Analysis (EDA) and Data Preprocessing:

1. Imported the necessary libraries including Pandas, Seaborn, Matplotlib and Numpy.
2. Read the CSV file using Pandas and stored the data in a dataframe called "df".
3. Inspected the dataset using the following methods:
   - **df.head(7)** to display the first 7 rows of the dataframe
   - **df.tail(7)** to display the last 7 rows of the dataframe
   - **df.describe()** to display descriptive statistics of the dataframe
   - **df.info()** to display information about the dataframe including the column names, data types, and number of non-null values.
4. Identified the presence of missing values in the dataframe. The columns "CYLINDERS", "FUELTYPE", and "FUELCONSUMPTION_COMB" had null values.
5. Decided to drop rows with missing values using the **dropna()** function and stored the resulting dataframe in a new variable called "emission".
6. Renamed the columns of the "emission" dataframe using the **rename()** function to make the names more readable.
7. Conducted Data Visualization using Seaborn and Matplotlib libraries.
   - Created **scatter plots** for all numeric features against the target variable CO2_EMISSIONS using **sns.pairplot()**.
8. Detected Outliers using the following methodology:
   - Defined the numerical columns for which we wanted to detect outliers.
   - Calculated the z-score and IQR for each numerical column using the **stats.zscore()** and **stats.iqr()** functions from the SciPy library.
   - Detected the outliers using the z-score and IQR methods and stored the outliers in new variables called **zscore_outliers** and **iqr_outliers**.
   - Printed the number and values of the outliers detected by each method using **print()**.
9. Decided to treat the outliers detected using z-score method.
10. Graphically displayed the outliers detected using z-score method using Seaborn and Matplotlib libraries.

- Created a new column to store the z-score of each data point for each numerical column using **np.abs()**.
- Created scatter plots for all numeric features against the target variable CO2_EMISSIONS using **sns.pairplot()**.

11. Find rows with z-score outliers for each numerical column and drop them from the dataset.
12. Create a new dataframe named emission_t with the outliers excluded values only.
13. Create a correlation **heatmap** to identify how strongly the variables are correlated.
14. Select numerical features to remove multicollinearity through **p-test**.
15. Encode categorical variables using **one-hot encoding** technique.
16. Perform **ANOVA** test on the encoded categorical variables.

Overall, the above methodology provided a comprehensive understanding of the data and allowed for effective data preprocessing to prepare the dataset for further analysis.

3.3 Model Building and Evaluation

1. A linear regression model is built using the preprocessed data to predict CO_2 emissions from vehicle features.
2. The model is trained on a subset of the data and tested on another subset using the **train_test_split** function from **sklearn**.
3. The performance of the model is evaluated using the mean squared error and R-squared value, which indicate how well the model predicts CO2 emissions.
4. The model appears to perform well, with a high R-squared value and low RMSE value.
5. An alternative model is built using a different train-test split to check the stability of the model.
6. Another alternative model is built without any data preprocessing to see the effect of multicollinearity on the model's performance.
7. The performance of these models is also evaluated using the mean squared error and R-squared value.

3.4 Prediction

1. The code begins by taking user inputs for engine size, fuel consumption, and fuel type.
2. A new dataframe is created with the user inputs, where the fuel type is initially set to 0 for all columns.
3. The value of the fuel type is set to 1 for the appropriate column based on the user input using an if-else statement.
4. A function called "preprocess_new_data" is defined to scale the numerical variables in the new data. The function divides the engine size and fuel consumption by their respective maximum values.
5. The new data is preprocessed using the "preprocess_new_data" function.
6. The final model is used to make a prediction on the preprocessed new data.
7. The predicted CO2 emissions are printed to the console.

4. Findings:

Findings of the full project are listed briefly:

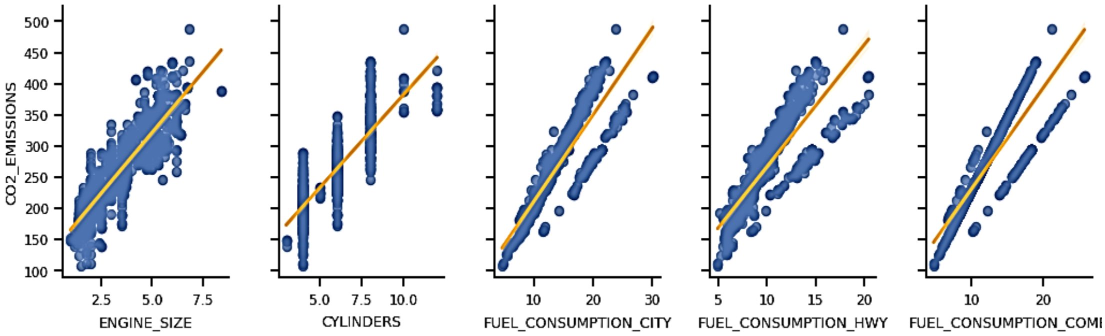
1. **EDA and Data Pre-processing:**
   The initial dataset had 1067 rows and 7 columns, with missing values in CYLINDERS, FUELTYPE, FUELCONSUMPTION_CITY, and FUELCONSUMPTION_COMB.

   The missing values were dropped from the dataset as there were enough data available for modelling. The column headings were also changed to make the dataset more readable.

2. **Data Visualization:**
Scatter plots were created for all the numeric features (ENGINE_SIZE, CYLINDERS, FUEL_CONSUMPTION_CITY, FUEL_CONSUMPTION_HWY, FUEL_CONSUMPTION_COMB) against the target variable CO2_EMISSIONS.

The scatter plots revealed that there is a positive correlation between CO2_EMISSIONS and all the numeric features, especially ENGINE_SIZE and FUEL_CONSUMPTION_COMB.

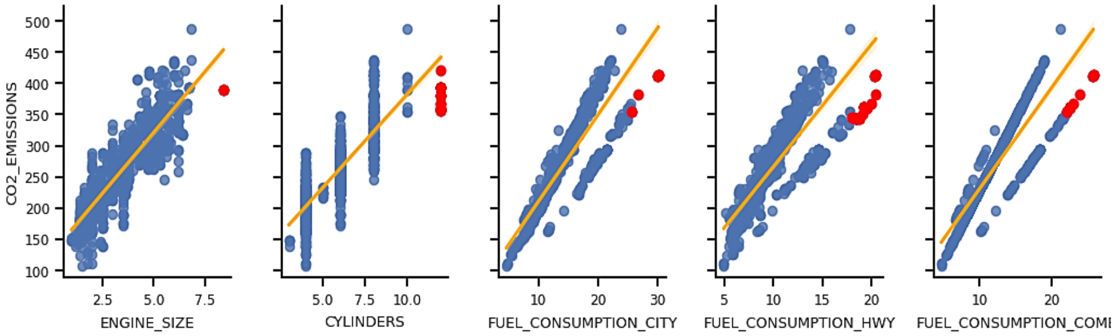
3. **Linear Relationship:**



It can be observed from the scatter plots that there is a strong positive linear relationship between CO2 emissions and the engine size, cylinders, and fuel consumption variables (city, highway, and combined).

4. **Outlier Detection:**

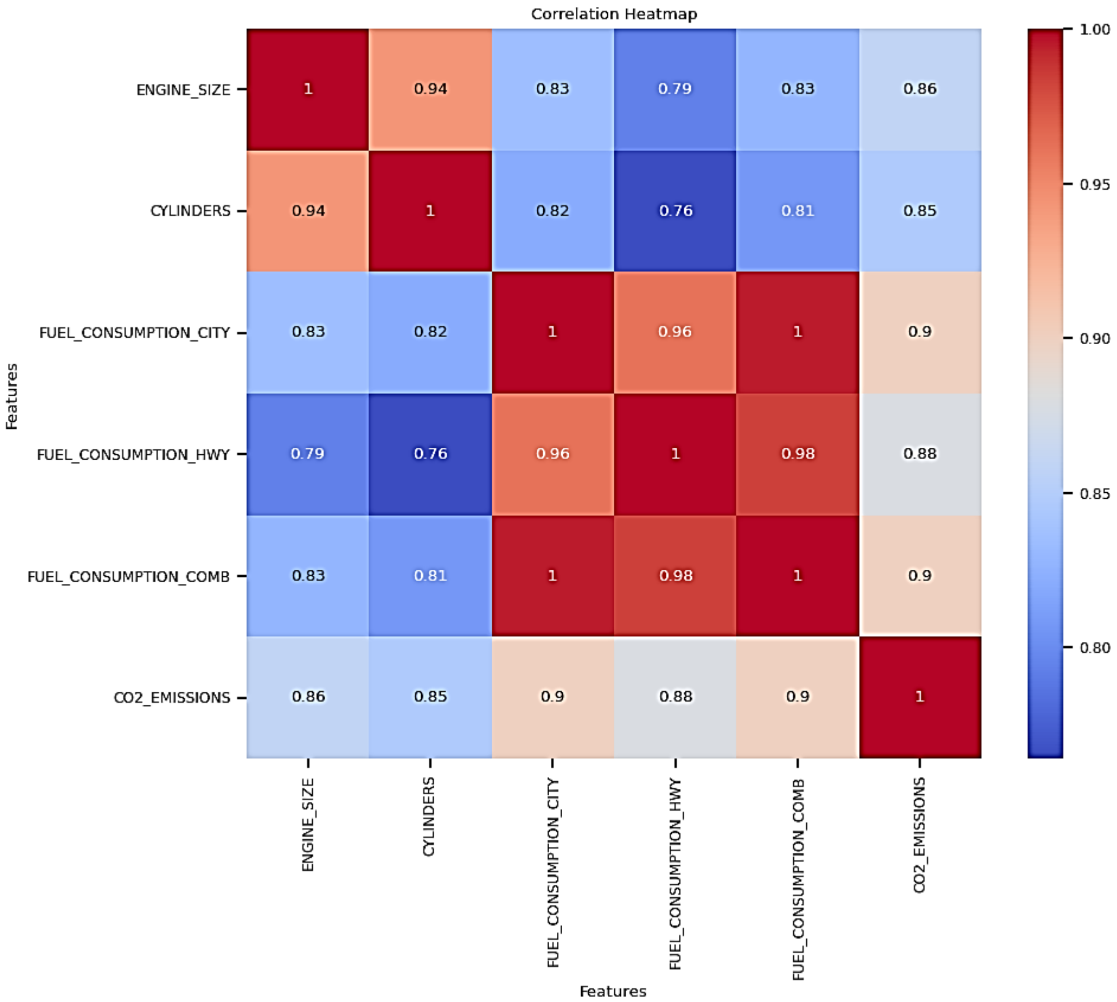
- The z-score and IQR methods were used to detect outliers in the dataset. The z-score method detected 2 outliers for the ENGINE_SIZE column and 20 outliers for the FUEL_CONSUMPTION_HWY column, while the IQR method detected 2 outliers for the ENGINE_SIZE column and 35 outliers for the FUEL_CONSUMPTION_HWY column. No outliers were detected for the CYLINDERS, FUEL_CONSUMPTION_CITY, and FUEL_CONSUMPTION_COMB columns using either method. Outputs are shown below:



Here, the red colored dots are identified as outliers.

Based on domain knowledge, it was decided to treat the outliers determined using the z-score method.

5. **Correlation Analysis:**
A correlation matrix and heatmap are created to identify the strength of correlation between variables.



The results show that CO2 emissions have a high positive correlation with engine size, cylinders, and fuel consumption variables.

6. **Multicollinearity:**
   Multicollinearity is identified in the fuel consumption variables, as they all measure fuel consumption in different driving conditions, and also in the engine size and cylinders variables, which are related to the size and power of the engine. To remove multicollinearity, a p-test is conducted and fuel consumption combined and engine size variables are selected for predicting the CO2 emission.

7. **Significance test for categorical variables:**
   A significance test (ANOVA) is conducted to test if there is a difference in CO_2 emissions between different fuel types. The results show that there is a statistically significant difference in CO_2 emissions between at least one pair of fuel types.

8. **Scaling of variables:**
   The descriptive statistics of each variable are obtained to identify if scaling is necessary. Based on the information provided, it appears that scaling may not be necessary for the linear regression model. This is because the range of values for each variable is relatively small, with most variables having a range of less than 10, indicating that the variables are already in a similar range and do not require scaling to bring them into a similar range. Furthermore, the

units for each variable are consistent (i.e., all in L/100 km or g/km), eliminating the need to scale them to ensure that they are in the same units.

9. **Model Performance:**
   The performance of the model is evaluated on a test set, and the root mean squared error (RMSE) and R-squared values are calculated. The high R-squared value of 0.99 indicates that 99% of the variance in the dependent variable (CO2 emissions) can be explained by the independent variables. The low RMSE value of 5.41 indicates that there is an average difference of 5.41 grams per kilometer between the predicted CO2 emissions and the actual CO2 emissions.

10. **Model's Stability:**
    To test the stability of the model, a different train-test split is used, and the performance of the model is evaluated again. The variation is found to be small, indicating that the model is stable.

11. **Effect of Multicollinearity:**
    A model without any preprocessing, (except for one-hot encoding for categorical data), is built and compared with the preprocessed model. The preprocessed model is then compared with this new model and the results are evaluated.

    Though it is generally expected that the preprocessed model will have a better performance with a lower RMSE value and higher R-squared value, in our case, the model without any pre-processing showed similar type of accuracy (RMSE value of the model: 5.415069530277368 and R-squared value of the model: 0.9917250270689348). Possible reasons behind this scenario can be stated as:
    1. The multicollinearity in our dataset may not be severe enough to significantly affect the performance of your model. In some cases, multicollinearity may not have a strong impact on the model's performance, especially if all the predictor variables have almost similar type of significance.
    2. The model built without preprocessing may have learned to ignore the highly correlated predictor variables and focus on the other relevant features in the dataset. This is possible when the dataset has enough samples and the highly correlated features do not have a strong impact on the target variable.

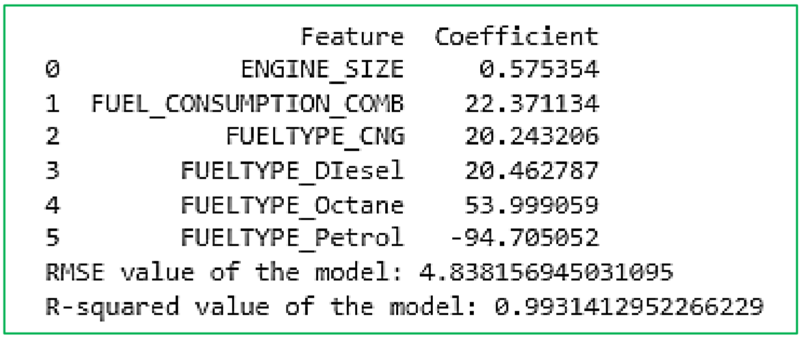
5. Final Multiple Linear Regression Model

Generally, it is recommended to remove highly correlated predictors to avoid multicollinearity issues and to simplify the model. This can help in interpreting the coefficients of the remaining predictors and make the model more robust to overfitting.

In the context of modeling, robustness refers to the ability of a model to maintain its performance and accuracy under different conditions, such as changes in the dataset or the model's parameters. A robust model is less affected by outliers, noise, and other sources of variability in the data, and it can generalize well to new or unseen data. A robust model is important because it can provide reliable predictions and insights, even in challenging or uncertain situations.

In this case, the preprocessed model was chosen despite having almost similar accuracy to the unprocessed model. This decision was made to address the issue of multicollinearity and to improve the interpretability and robustness of the model.

Our selected final model has the following properties:



```
                  Feature  Coefficient
0             ENGINE_SIZE     0.575354
1   FUEL_CONSUMPTION_COMB    22.371134
2            FUELTYPE_CNG    20.243206
3         FUELTYPE_DIesel    20.462787
4         FUELTYPE_Octane    53.999059
5         FUELTYPE_Petrol   -94.705052
RMSE value of the model: 4.838156945031095
R-squared value of the model: 0.9931412952266229
```

The final model has a high R-squared value of 0.993, which means that it explains 99.3% of the variance in CO2 emissions. This indicates that the model is a good fit for the data and has strong predictive power.

The model coefficients show that engine size and fuel consumption have a positive effect on CO2 emissions, while the type of fuel has a varying effect. In particular, vehicles that run on Octane tend to emit more CO2, while those that run on Petrol tend to emit less CO2.

The RMSE value of the model is relatively low at 4.84, which indicates that the model's predictions are accurate and have low error.

Overall, the findings suggest that the final model can be used to predict CO2 emissions for vehicles based on their engine size, fuel consumption, and fuel type with a high degree of accuracy. The insights gained from the model can also help policymakers and manufacturers make more informed decisions regarding vehicle emissions and fuel efficiency.

6. Prediction

Using the final model, a user can give input of the variables and find out the expected CO_2 emission for that specific type.

7. Conclusion

In conclusion, this report aimed to provide insights into the relationships between engine size, cylinder count, fuel consumption, and CO2 emissions, and to develop models that can accurately predict CO_2 emissions in vehicles. The linear regression model performed well, with a high R-squared value and low RMSE value. Overall, the report provides useful insights into the factors that affect CO_2 emissions in vehicles and demonstrates the effectiveness of a multiple linear regression model in predicting CO_2 emissions.