**CO2 EMISSIONS (Predicting CO2 Emissions of Cars)**

Instructions -

1. Please read the question carefully before attempting them.

2. Solve all the questions in a SINGLE jupyter notebook file.

3. In case name of the variable to be used is mentioned in the question, use the same name while

Coding (marks are associated with it)

4. In your answers include your descriptions as and when mentioned. Think yourself as a Data

Analysts, wxxho needs to suggest and explain solutions to the client based on Data.

**Climate change due to carbon dioxide (CO2) emissions is one of the most complex challenges threatening our planet. This issue considered as a great and international concern that primary attributed from different fossil fuels.**

**To reduce the effect of Global Warming caused mainly due to CO2 cars manufacturing companies are pre-estimating the CO2 emissions for a newly manufactured car that is about to launch in near future based on the features of car.**

**This project deals with estimating the CO2 Emissions for a newly manufactured car by making a Regression models which can accurately predetermine the car CO2 emissions before it is being launched.**

**Note** - For working on this Project, we recommend the use of Python 3.x version with libraries such as Numpy, Pandas, Matplotlib, Seaborn, SkLearn, Scipy, statsmodels), Jupyter Notebook, MySQL. Also, the knowledge of Supervised Learning: Regression is required.

The Data given here is for Year 2014 manufactured Cars.

YEAR – Year of manufacturing of car.

MAKE – Manufacturing company name.

VEHICLECLASS – Type of vehicle like SUV or medium sized etc.

ENGINESIZE – Size of the car’s engine (expressed in cc or cubic centimetre).

CYLINDERS – Number of Cylinders in the engine.

TRANSMISSION – Automatic or manual transmission with number of gears.

FUELTYPE – It indicates the type of fuel car use i.e. Diesel, Petrol, Z (Unleaded Petrol) etc.

FUELCONSUMPTION\_CITY – Fuel consumption or Fuel economy of car while running in city expressed in miles per gallon.

FUELCONSUMPTION\_HWY - Fuel Consumption or Fuel economy of car on highway expressed in miles per gallon.

FUELCONSUMPTION\_COMB – Net or combination of Fuel Economy expressed in miles per gallon.

FUELCONSUMPTION\_COMB\_MPG – Total fuel economy expressed in miles per gallon.

CO2EMISSIONS – The CO2 emitted by the car expressed in grams.

**Your task is to create a Predictive Model using Linear Regression Technique by splitting the original data into test and train datasets.**

**Data file** – CO2 emission.csv

**Task 1 – Loading the Data**

a) Install the Pandas library. Import the library into the code file

(Hint: install pandas in jupyter notebook using

**!pip install pandas**

)

b) Import the required libraries. Store the data into Pandas dataframe and call it emission.

(Import all libraries like:-

**import pandas as pd**

**import seaborn as sns**

**import matplotlib.pyplot as plt**

**import numpy as np**

**from sklearn import linear\_model, metrics**

)

**Task 2 - Analysing and cleaning data**

a) Analyse the shape of the Data using emission dataframe variable. Check the no. of rows and Columns available.

b) Print the top 5 rows and check the values.

c) Pick out four columns from the above dataframe which contributes towards CO2 emissions for a car and rename it as emission\_new.

**Task 3 - Perform explanatory analysis**

a) Check the basic statistics of the given new dataset columns. Describe the Dataset in your words including the columns, their type, ranges and other stats. Statistics should include the mean, median, standard deviation, different quartiles, minimum and maximum values of each column.

b) Check the units of all the columns.

**Task 4 - Visualizing the Data**

a) Create the Box-plot of all the numeric columns and explain the result in your words for emission\_new dataset using pandas hist () function.

(Hint: In this you need to plot a histogram for the complete new dataframe built using Pandas inbuilt function hist ()).

b) Next you need to find out the dependent variable from all the selected 4 columns from initial dataframe using the corr () function available in pandas dataframe.

c) Plot a heatmap for correlation between all columns of emission\_new dataframe by passing the corr () function to heatmap which will automatically plot the correlation heatmap for the dataframe.

d) Create a figure and plot 3 scatterplot graphs in same row.

* The first plot should be between ENGINESIZE and CO2EMISSIONS.
* The second plot should be between CYLINDERS and CO2EMISSIONS.
* The third plot should be between FUELCONSUMPTION\_COMB and CO2EMISSIONS.

Analyse the graphs and conclude which columns are best fit for the independent variables for carrying out Linear Regression and which column can act for Dependent Variable.

e) Now as we have concluded that CO2EMISSIONS is a dependent variable and ENGINESIZE, FUELCONSUMPTION\_COMB are the two independent variables so plot an lmplot between

* ENGINESIZE and CO2EMISSIONS
* FUELCONSUMPTION\_COMB and CO2EMISSIONS

**Task 5 – Linear Regression using Train and Test data**

In this step we will proceed with splitting the original data into train data and test data in a random order.

a) Use emission\_new Pandas dataframe to split data into training dataset and test dataset and perform Linear Regression keeping all the parameters as decided above.

b) Plot two plots into a same graph

* The first plot should be scatter plot between train data ENGINESIZE column and train data CO2EMISSION column.
* The second plot should be a regression line.

[Hint: - Plot a simple line plot between independent variable of train data and calculated value of dependent variables (CO2EMISSIONS) using above built model i.e. by using coef\_ and intercept\_ parameter of model].

c) Plot a scatter plot between actual value of CO2EMISSIONS and predicted value of CO2EMISSIONS.

d) Find out the mean squared error using sklearn metrics module.

**Task 6 – Multivariate Prediction Model using Train and Test data**

a) Use emission\_new Pandas dataframe to split data into training dataset and test dataset and perform multivariate Regression keeping all the parameters as decided above.

b) Plot a scatter plot between actual value of CO2EMISSIONS and predicted value of CO2EMISSIONS.

c) Find out the mean squared error using sklearn metrics module.

d) Explain which model is suitable i.e. **Linear Regression using Train and Test data** or **Multivariate Prediction Model using Train and Test data** using mean square value and R-value of regression.