

**CSE523 : Machine Learning**

Winter 2021 - 2022

**Weekly Report - 1**

Dt - 09-02-2022

**Group Name : Discover Decipher**

**Group Members**

Nimisha Patel - AU1940146 - Btech Computer Science and Engineering.

Sakshi Shah - AU1940213 - Btech Computer Science and Engineering.

Astha Patel - AU1940312 - Btech Computer Science and Engineering.

Kareena Matwani - AU1940314 - Btech Computer Science and Engineering.

**Prediction of CO2 emission based on the car/vehicle using Machine Learning Algorithms**

**Abstract**

The current global CO2 emissions are bizarre and transportation/vehicles contribute to the CO2 emissions largely. Machine learning algorithms are widely used to solve complex environmental problems. The aim is to predict the CO2 emissions using relevant supervised learning techniques such as regression models, single variable linear regression, multi valued linear regression, polynomial regression and non-linear regression or SVM algorithm (used for linear and non linear classification, regression and detection) by entering a few details of the car/vehicle like model, engine capacity, fuel consumption, etc. The model will also be evaluated with the help of mean squared error and R2 scores.

**Task Performed this week**

1. **Finding relevant information**

* **Dataset**

The dataset included 7385 light-duty vehicles observed from 2017-2021 with the predictive view of the fuel consumption and carbon dioxide emissions of different vehicles models and brands. Based on the data, we wish to make suggestion pointers regarding features that can help reduce environmental impacts due to vehicles.

* **Features**
* **Vehicle company :** Company of the vehicle
* **Model :** Car Model
* **Vehicle class :** Class of vehicle depending on their utility, capacity and weight.
* **Engine Size :** Size of engine in terms of litre.
* **Cylinders :** Number of cylinders for the vehicle engine. Cylinder is the power unit of the engine and is the place where fuel is burned and converted to mechanical energy used for the vehicle power.
* **Transissions :** Composed of a different gear set that allows the driver to control the power to the vehicle without changing engine speed.
* **Fuel type :** Type of fuel used for vehicles.
* **Fuel consumption City (L/100 km) :** Fuel consumption on city roads
* **Fuel consumption Hwy (L/100 km) :** Fuel consumption on highways
* **Fuel consumption Comb (L/100 km) :** The combined fuel consumption (55% city, 45% highway) is shown in L/100 km
* **Fuel consumption Comb (mpg) :** measure of how far a vehicle can travel per gallon of fuel, where MPG is measure of vehicle's fuel efficiency.
* **Label - Prediction column :**

**CO2 emission in (g/km) :** Co2 emission due to fuel burning in g per kilometer.

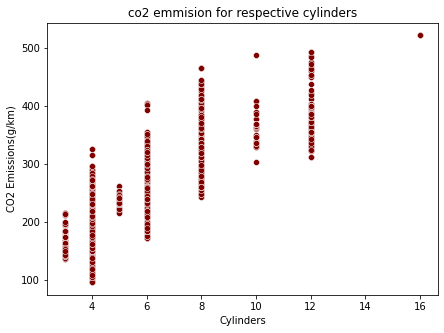
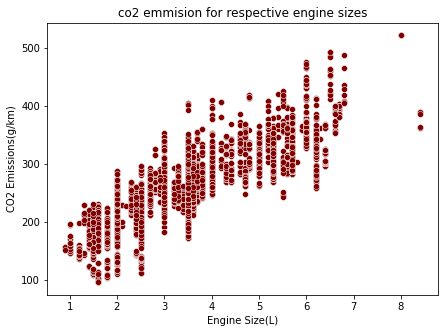
* **Dataset**
* <https://www.kaggle.com/gangliu/oc2emission/tasks>
* <https://www.kaggle.com/debajyotipodder/co2-emission-by-vehicles>
* <https://data.worldbank.org/indicator/EN.ATM.CO2E.LF.KT> (secondary if required)
* **Other Important Links**
* <https://towardsdatascience.com/how-machine-learns-from-data-a-simple-method-for-co2-pollution-prediction-af18430ce12b>
* *CO emissions from cars: The facts - transport & environment*. (n.d.). Retrieved February 7, 2022, from <https://www.transportenvironment.org/wp-content/uploads/2021/07/2018_04_CO2_emissions_cars_The_facts_report_final_0_0.pdf.>
* **Machine Learning algorithms that can be used for solving the problem and its basic information**
* Regression for continuous numerical output
* Linear regression
* SVM - Simple Vector Machine : Algorithm to perform classification as well as regression.
* Random Forest : Ensemble of decision trees - searches for the feature value in the random subset of features when growing trees instead of searching for best feature when splitting a node.
* **Measures to evaluate the algorithms**
* Mean Square Error
* Mean Absolute Error
* R squared coefficient of determination

1. **Data Exploration and Understanding**

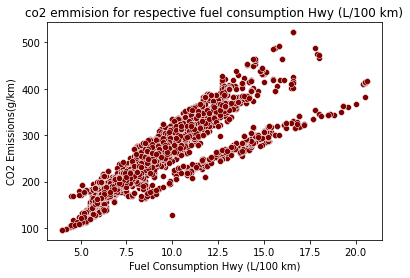
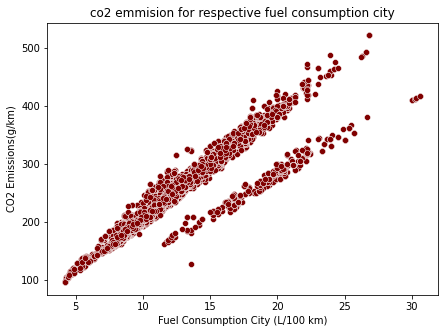
After checking the data, we tried to understand the data by plotting different features with CO2 emission which helps depict the correlation and measure dependency of CO2 emission of any particular feature to a certain extent.

**Scatter plot for different features vs. CO2 Emissions**

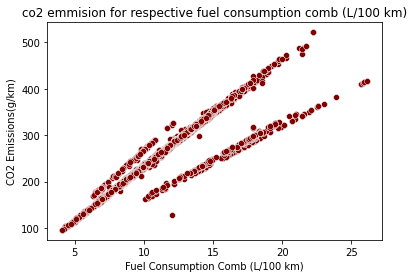
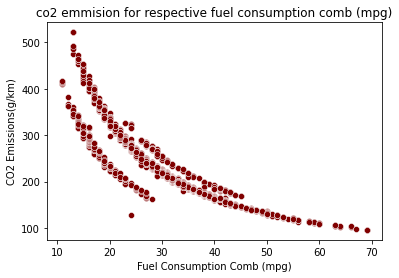
From the scatter plots (with help of python matplotlib) for different features vs CO2 emission (that is to be predicted), it can be inferred that there is some relationship between the fuel consumption and the co2 emissions. While from engine size and cylinders one can visualize the range of CO2 emission.

****

**Engine Size Vs Co2 emission Cylinders vs Co2 emission**

****

**Fuel consumption City (L/100 km) vs Co2 emission Fuel consumption Hwy (L/100 km) vs Co2 emission**

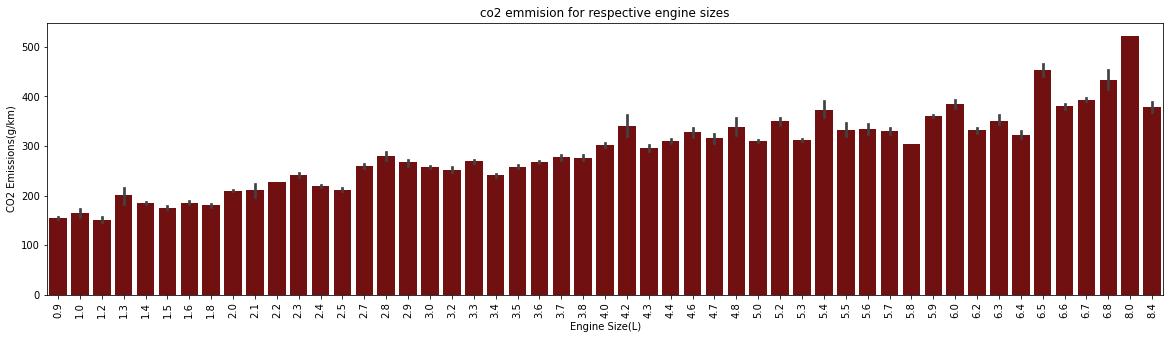
****

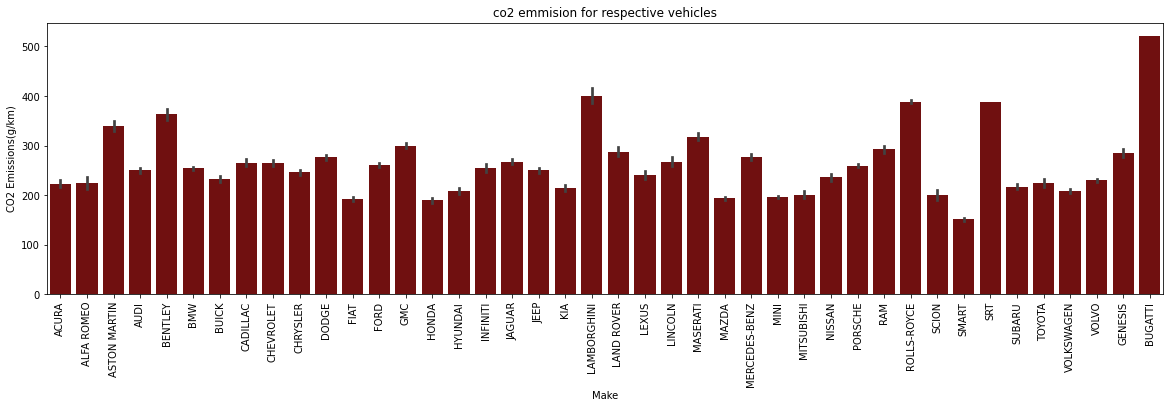
**Fuel consumption Comb (L/100 km) vs Co2 emission Fuel consumption Comb (mpg) Vs Co2 emission**

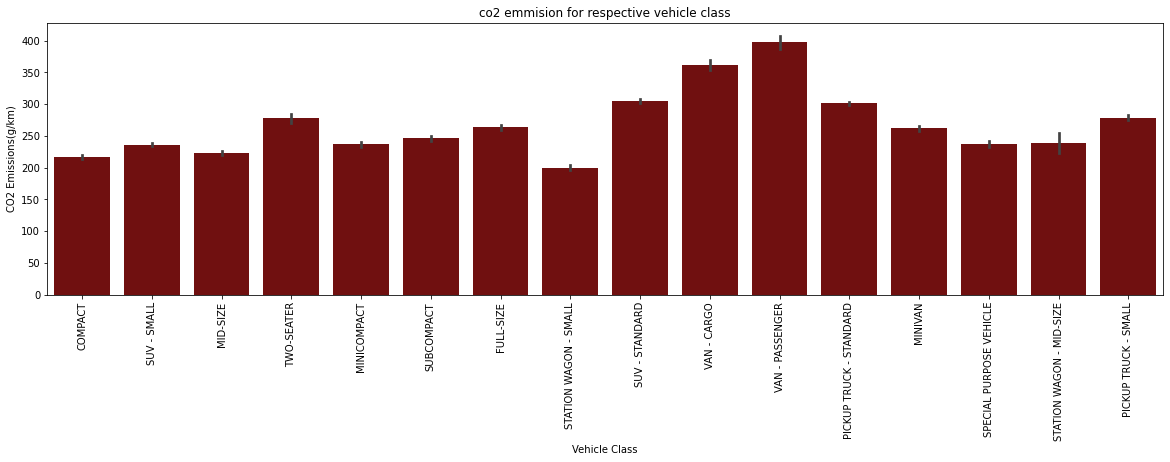
**Bar graph for different features vs. Co2 Emissions**

Bar plots(with help of python matplotlib) are to visualize, compare and understand the co2 emission company wise, enginewise and by the type of fuel consumption.

**Engine Size(L) :**

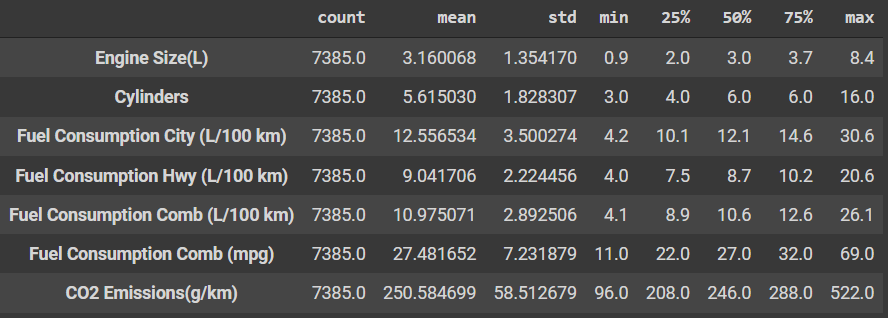
****

**Vehicle company : **

**Vehicle class : **

**Fuel type : **

**Basic statistics measures (obtained using python describe() function) for the variables in the dataset**



**Task to be performed next week**

1. Data preparation and division into train, test, validation.
2. To see if this problem can be solved as a classifier problem?

* Classes
* severe emission (out of permissible limit)
* high emission
* medium emission
* Less emission

1. Data analysis and model preparation.