# 

**CSE523 : Machine Learning**

Winter 2021 - 2022

**Weekly Report - 2**

Dt - 16-02-2022

**Group Name : Discover Decipher**

**Group Members**

Nimisha Patel - AU1940146 - Btech Computer Science and Engineering.

Sakshi Shah - AU1940213 - Btech Computer Science and Engineering.

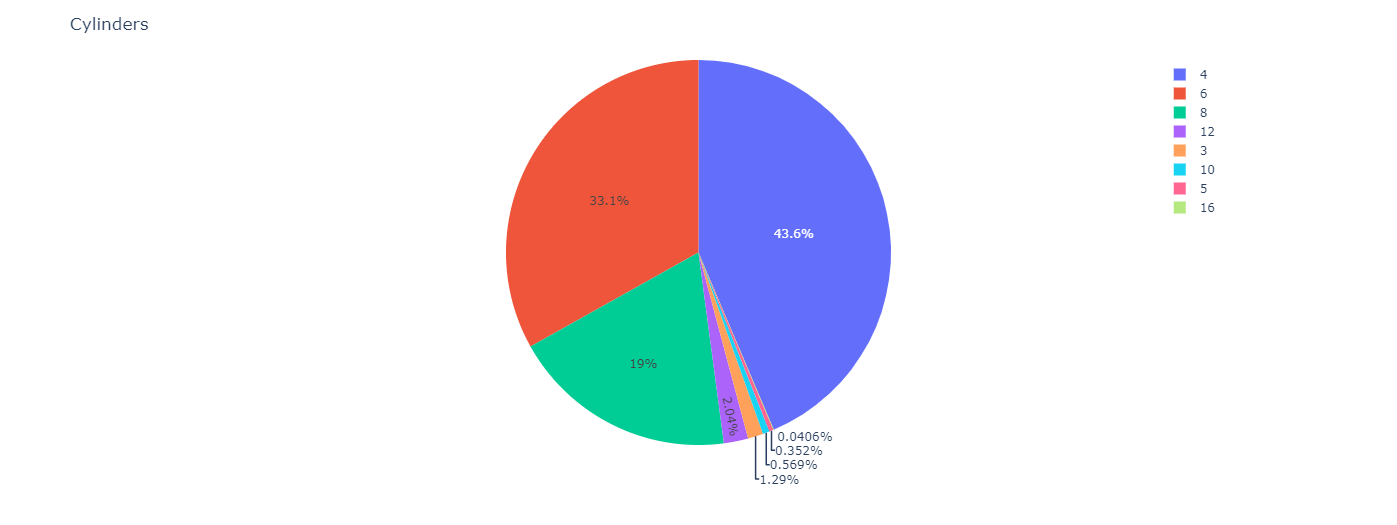
Astha Patel - AU1940312 - Btech Computer Science and Engineering.

Kareena Matwani - AU1940314 - Btech Computer Science and Engineering.

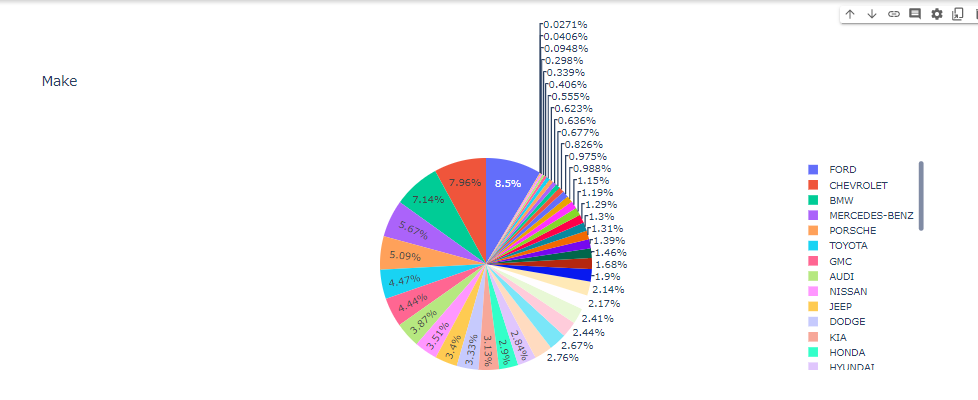
**Task Performed this week**

1. **Interpretation of the obtained results on understanding data**

* The below graphs interpret that the number of cylinders contributing to the Co2 emissions which is for 7385 light-duty vehicles and to what extent a small or larger number affect the emissions as shown below respectively.

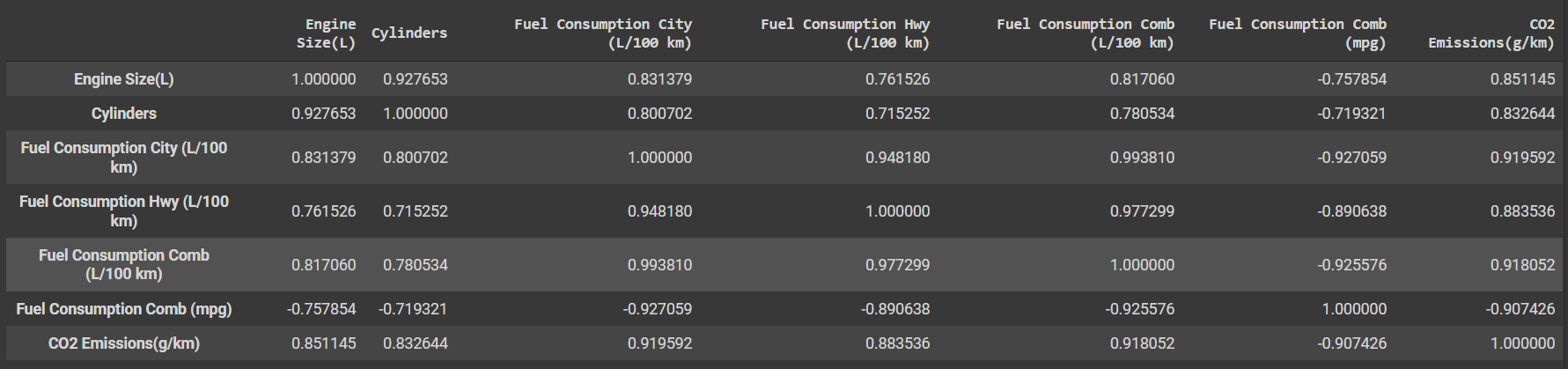


* The type of CO2 emissions collectively has taken the percentage by what type of cars is interpreted as shown below which may help us in further interpretation of data.

****

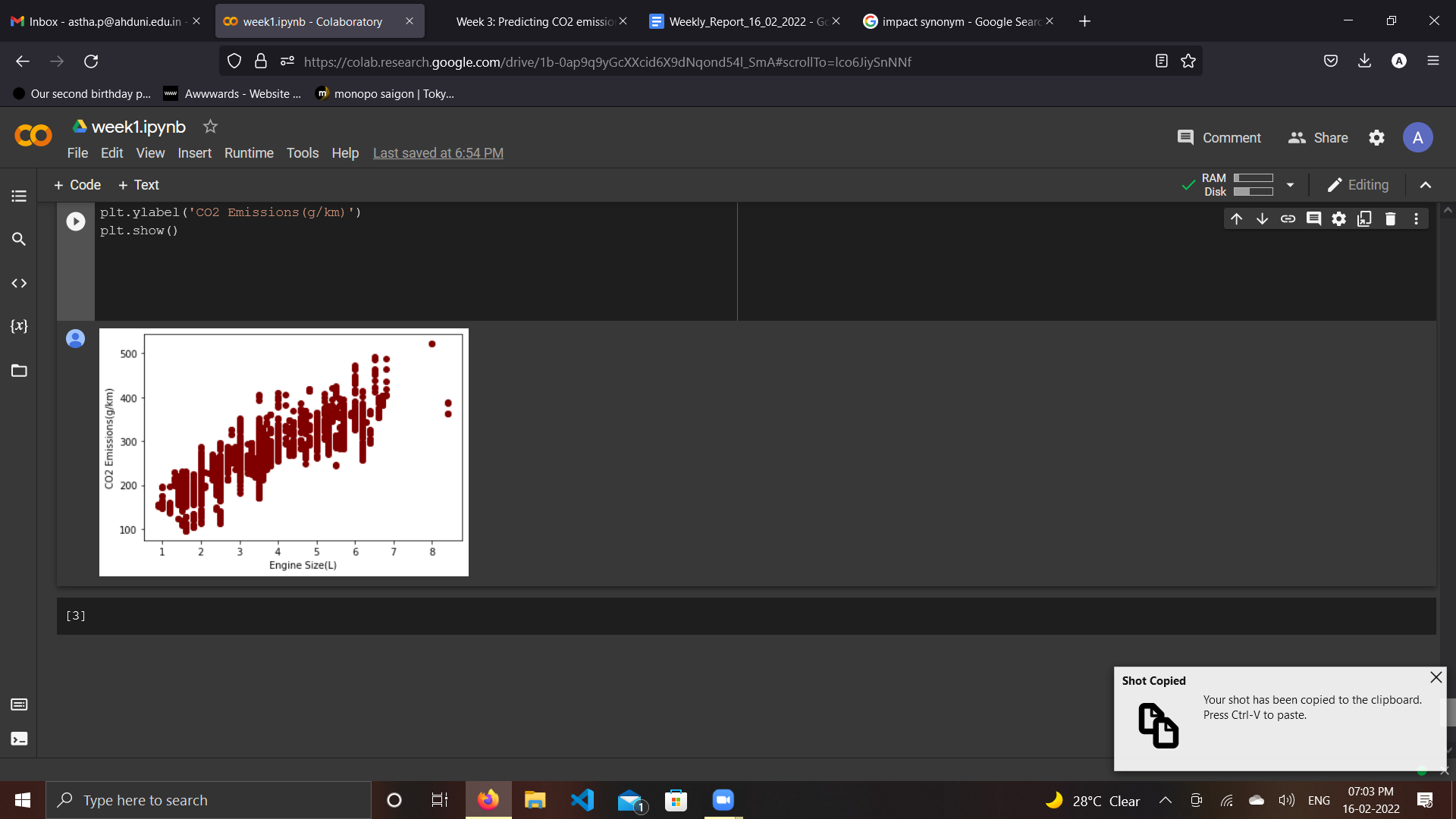
1. **Data preparation and division into train, test, validation.**

* Correlation matrix table shows correlation coefficients between variables. Each cell value shows the correlation between two random variables. This matrix used to summarize data as an input into more advanced analysis.
* The diagonal of 1.0000s shows that each variable always perfectly correlates with itself. This matrix is symmetric. The same correlation is shown above the main diagonal being a mirror image of those below the diagonal.
* To make a summary of a huge amount of data with the purpose of identifying patterns. The observable pattern in our example is that all of the variables are significantly correlated.



**Correlation Matrix**

* The engine size (power of engine) of a car affects the CO2 emissions.
  + Proof: J Mater Environ Sci.com. 2022. [online] Available at: <https://www.jmaterenvironsci.com/Document/vol11/vol11_N2/JMES-2020-11-18-Al-Arkawazi.pdf> [Accessed 16 February 2022].
* The data is splitted into two parts.
* The feature vector is defined (given the variable name) as engine which contains Engine Size(L).
* CO2 emissions(g/km) is defined as co2.
* With the help of train\_test\_split data is splitted into train dataset and test dataset.
* Here is the scatter plot of the correlation between features of train dataset



**Studying the correlation between Engine size and CO2 emissions**

1. **Important Links**

* Medium. 2022. *Machine Learning Concept behind Linear Regression*. [online] Available at: <https://towardsdatascience.com/how-machine-learns-from-data-a-simple-method-for-co2-pollution-prediction-af18430ce12b>

[Accessed 16 February 2022].

* Medium. 2022. *Machine Learning Concept behind Linear Regression*. [online] Available at: <https://towardsdatascience.com/how-machine-learns-from-data-a-simple-method-for-co2-pollution-prediction-af18430ce12b> [Accessed 16 February 2022].

**Tasks to be Performed Next Week**

* Classification model based on the existing columns in data
* How to incorporate Linear Regression with the data set