# Introduction:

Global warming is the most important and sensitive topic of 21st Century. Global worming endangers our health as well as the health of earth. CO2 is a greenhouse gas that is mostly responsible for the global warming. One of the most carbon dioxide emitting tool is car.   
In this project we are going to explore what variables responsible for carbon emission.  
For this purpose we have a data set form EU-28. It contains 26 variables and approx.419000 observations. Where from my research point of view Carbon emission rate (e gm/km) is the dependent variable, as it can be predicted using other independent variable.

# Overview of the project:

The data contained many missing and many unwanted values. Thus in the pre-processing step, first the data was converted into columns as it was in text format and then using R the rest of the pre-processing was done like imputing wrongly inputted data, same data in different ways, space problems, o written as 0 etc.  
The next step was Variable selection as many of the variables were conveying similar type of information and some were not needed as per our research purpose. Thus selection initially by observation and then by statistical screening was done and finally factor scores were used instead of the continuous variables as dropping one variable may cause loss of information.   
The next step was building a model that can predict carbon emission.

# Missing Values:

The given CO2 data have 15 categorical variable and 10 numeric variable.

The distribution of the continuous variable are given in the graph.

%missing

id 0.000

MS 0.000

MP 0.000

Mh 0.000

MAN 0.000

MMS 0.000

TAN 0.000

T 0.000

Va 0.011

Ve 0.000

Mk 0.000

Cn 0.000

Ct 0.000

r 5.864

e..g.km. 6.102

m..kg. 5.961

w..mm. 11.311

at1..mm. 11.781

at2..mm. 14.976

Ft 0.000

Fm 0.000

ec..cm3. 6.539

ep..KW. 26.291

z..Wh.km. 99.720

IT 0.000

Er..g.km. 99.990

Some variables in the dataset are insufficient and some are not useful thus we drop some variable and name the new dataset “Work”.

Thus find the missing values in the dataset as it contains the variables of our concern.

The result is given below:

%missing

MS 0.000

MP 0.000

TAN 0.000

Mk 8.035

Ct 0.000

r 5.864

e 6.102

m 0.000

w 0.000

at1 0.000

at2 14.976

Ft 5.896

Fm 0.000

ec 0.000

ep 0.000

So, from the “mis\_detect” function created to find the missing values, we get that there are 5 variables that have missing value.

To work better we converted the entire dataset into lower i.e. it will be easier to deal with the other impurities like misplaced character, un-wanted character etc.

Some variables here like Mk, r, e, at1, Ft have missing values.

Amongst which Mk, r, Ft are categorical variables.

And e, at1 are continuous variables.