SYSTEM ANALYSIS

For the system analysis we will analyze the data set for the further usage of it in our project.

- **Cleaning of Data:-**

Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a Data set. Hence we find out the missing and 'NaN' values in the data set. There were many 'NaN' values specially in the last rows of data set.

```
#droping end rows with NaN values
df1.dropna(how='all',inplace=True)

#Counting Missing Values from each Column
print('Missing values:\n\n',df1.shape[0]-df1.count())
clean=(df1.shape[0]-df1.count()).sum()
print("\n")
if(clean==0):
    print("No Missing Values")
```

- **Formatting of Data:-**

Data formatting depends upon the purpose of your data, elements in the data and much more. Data formatting enhances the visual appearance of our worksheet. Hence we find out that Time and Date column are in Object Data type which can create a problem afterwards. So, I converted TIME Column (HH:MM:SS) format into a new column i.e. 'HOUR' column and DATE Column (YYYY-MM-DD) format into a new column i.e. 'MONTH' column.

```
#Splitting Column TIME(HH:MM:SS) into new column(of int64 type)
df1['HOUR']=df1['TIME'].apply(lambda x: int(x.split(':')[0]))

#Defining Month from DATE column
df1['DATE']=pd.to_datetime(df1.DATE, format='%d-%m-%Y')
df1['MONTH']= df1['DATE'].dt.month
```

- **Understanding Correlation between variables:-**

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. It can be of 3 types:-

1. Positive Correlation – when the value of one variable increases with respect to another.

2. Negative Correlation – when the value of one variable decreases with respect to another.
3. No Correlation – when there is no linear dependence between the two variables.

   Hence a correlation can be from -1.0 to +1.0. So, here we have used 'heatmap' (by importing seaborn and matplotlib libraries) for this which is a very simple and efficient method to understand the correlation among variables.

```
sns.heatmap(df1.corr(),annot=True, linewidths=.4)
plt.title('Heatmap of co-relation between variables',fontsize=16)
plt.show()
```

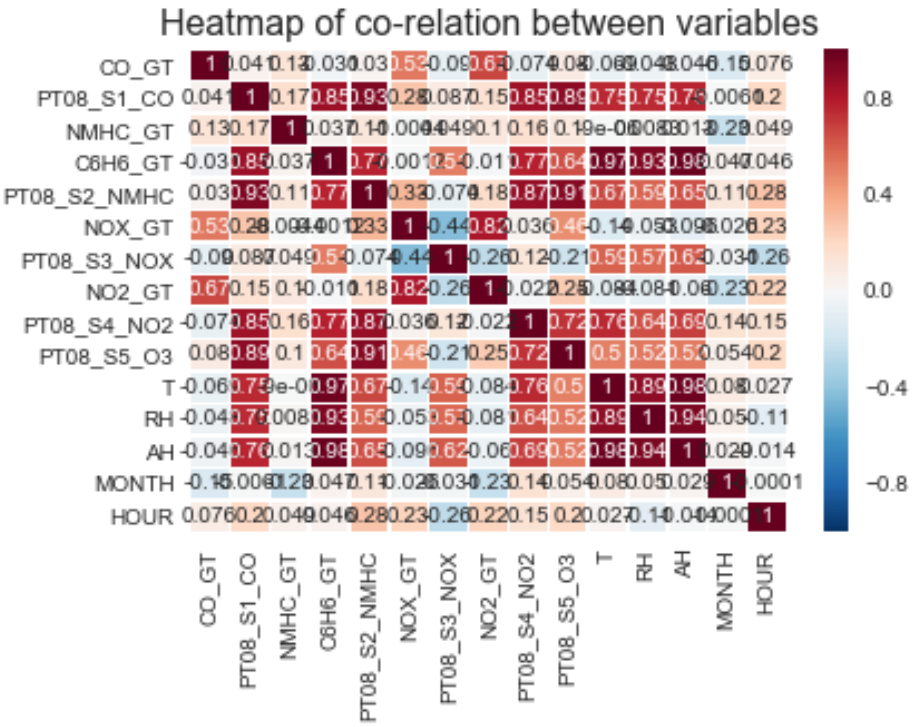

Figure 1: Understanding Co-relation between variables.

- **<u>Understanding Linearity between Relative Humidity and other variables:-</u>**

  Linearity is the property of a mathematical relationship or function which means that it can be graphically represented as a straight line. Here we have used 'lmplot' to plot out and analyze the linearity between Relative Humidity (RH) and other variables.

```
col1=df1.columns.tolist()[2:]
for i in df1.columns.tolist()[2:]:
    sns.lmplot(x=i,y='RH',data=df1,markers='.')
```

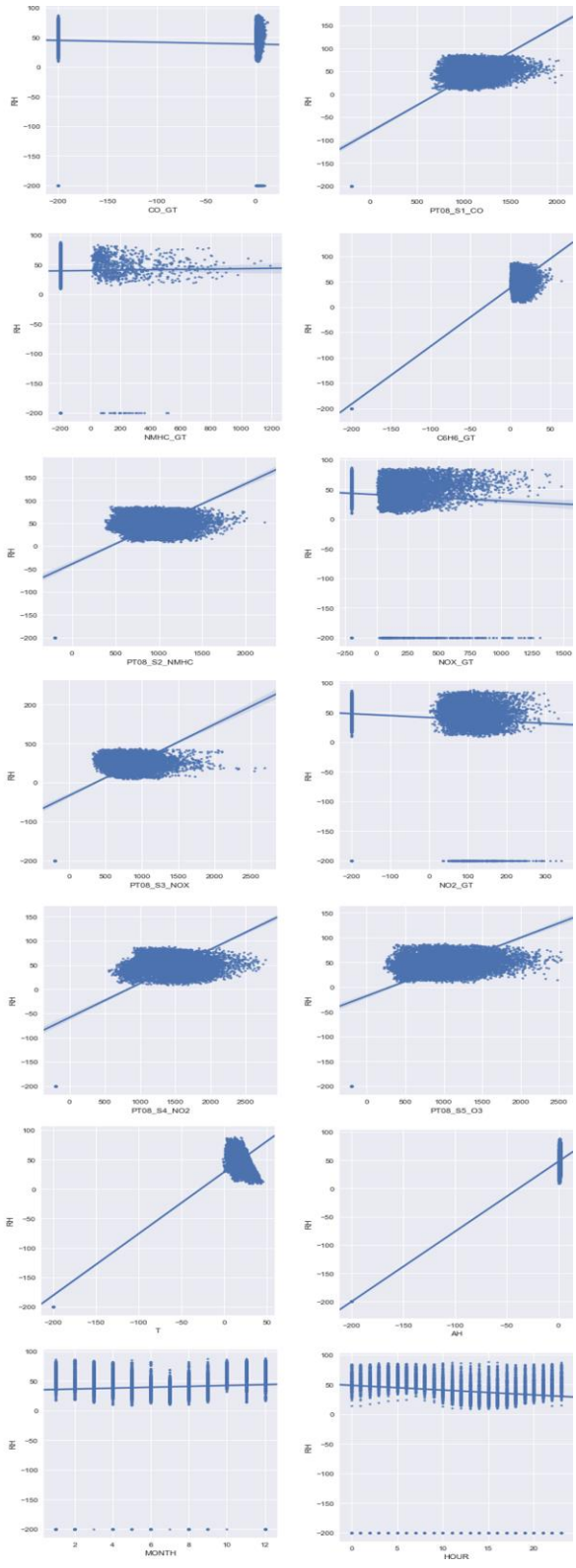

Figure 2: Understanding Linearity between Relative Humidity and other variables.