**Statistics for Data Science Assignment Report**

**Subhadeep Dash,**

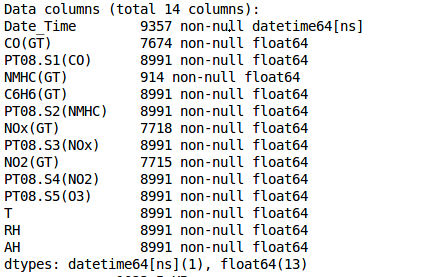
**S20160010021**

**Abstract:**

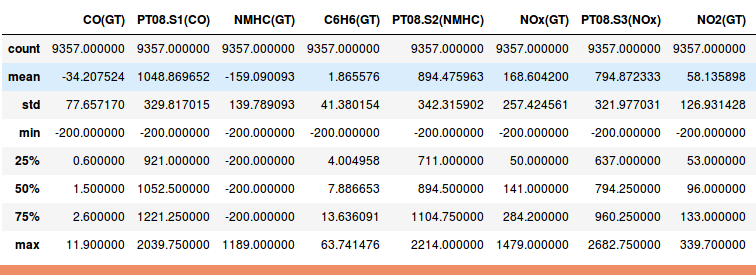
This paper is a report on the data analysis, regression performed and time series analysis performed on the Air Quality Dataset. In this paper, various ways of representing the statistical data and further processing on the dataset will be discussed.

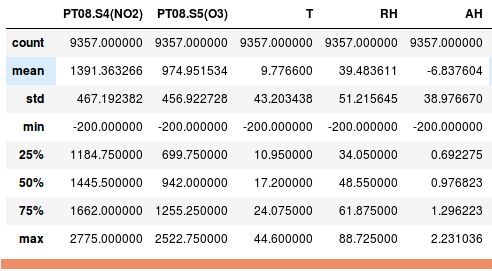
**Problem Statement:**  
Perform Regression, Data Analysis and Time Series Analysis on the Air Quality Dataset provided.

**Initial Data Pre-processing:**  
The dataset was scanned and various details were checked upon. According to the dataset report, the null values in the data-set were replaced by -200 which could be seen in the overall summary of the data. The statistics after-200 was replaced with null from the dataset is as follows:



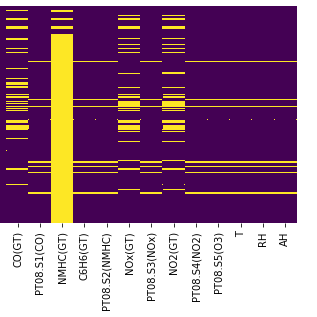
We can observe that there are 9357 records are present out of which no field is completely filled for all records. The detailed description of the data is available below. The count includes the null attributes and the minimum is displayed as -200 for most of the columns because the -200 values have not been removed.





**Tables displaying the data before missing values replacement.**

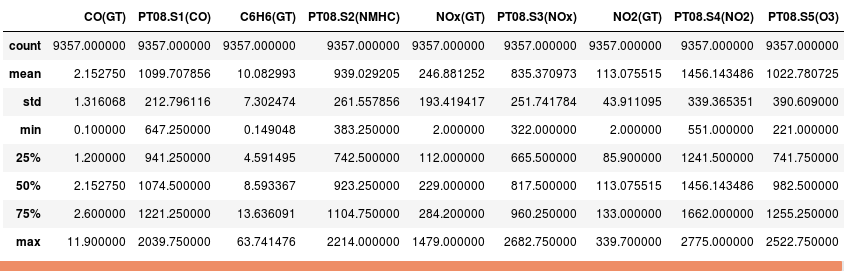
The heat map for the missing values in each field for the data-set is as follows:

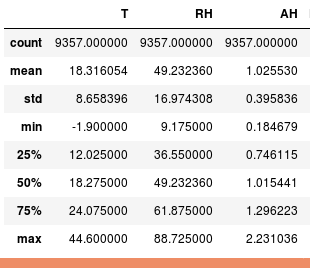


**Heat Map Displaying Null values at various positions of the dataset**

It can be seen from the head map that the NMHC(GT) column consists of a lot of null values, so it is better to discard it. For other columns, we replace the missing values in them with the means of their respective fields.

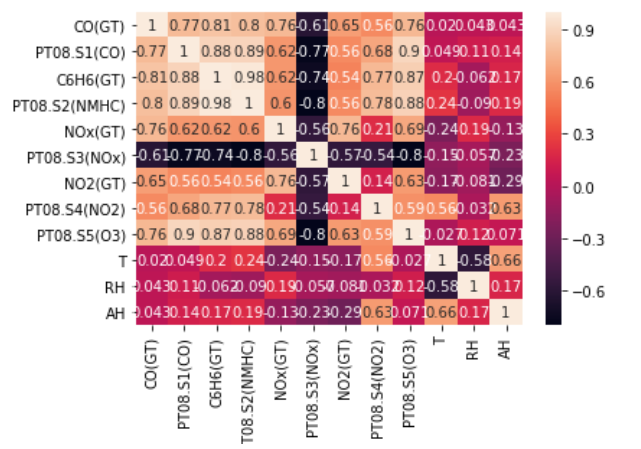
**Modifying missing information:**

After the removal of null values and discarding unnecessary fields, we obtain the following modified information from the data:  




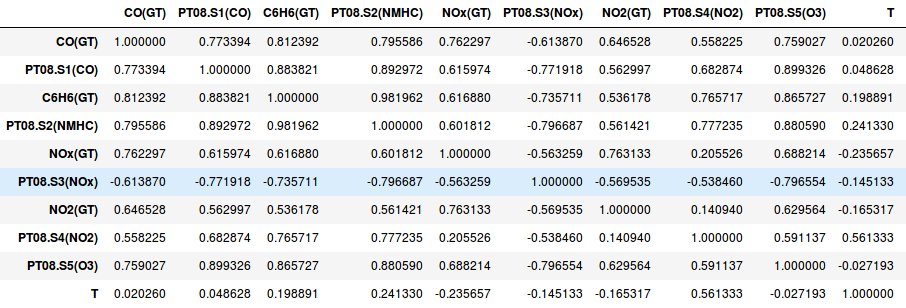
**Tables representing modified values after removal of NULL from the dataset**

Let’s have a look on the correlation matrix obtained with the help of a heat map.



After removal of a field and filling missing values with mean, we obtain the above correlation matrix from which we can observe that it is symmetric and the diagonals represent the correlation with the element itself. Therefore, there are only ones across the diagonal which are represented in white colour. Some variables can be seen as negatively correlated which are towards the darker side of the heat map while others are positive correlated which have lighter colours in the heat map.

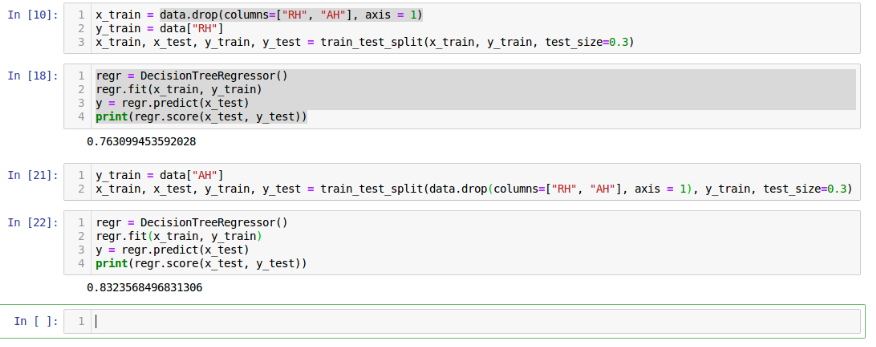
Our data attributes which we will be predicted through regression are AH and RH i.e. Absolute Humidity and Relative Humidity. Detailed and evident correlation with more precision between the features can be looked below:



**Correlation Matrix of training features**

**Regression:**

Regression is a process of predicting a value of a specific attribute by implementing a well-trained model. Regression may seem similar to classification but classification involves assigning each data record ‘d’ to a well-defined class ‘c’ by predicting based on the values of the attributes whereas Regression predicts the value of a specific feature ‘f’ by working on the data of several attributes i.e. Regression is used for continuous labelling whereas Classification is used for discrete labelling.  
 In this assignment, the values of RH and AH were predicted by using a Decision Tree Regressor. The implementation was using Python and Sci-kit learn library was used for training the construction of the regressor. SVMs weren’t chosen, as they were taking a lot of time to train and Decision Trees were built quickly were compared to them. A Regression Score of 76.31 was obtained when the regressor was used to predict the Relative Humidity value. When the Absolute Humidity was predicted using the same regressor, a score of 83.24 was obtained. The code snippet can be seen in the figure below:



**Code Snippet of training and testing in Python**

The same data had been divided into two parts out of which one was used for training and the other was used for testing. Decision Trees are considered one of the best structures involved in classification and regression. They are easy to interpret and just like the ones how human brain involves decision making(step-by-step).