1. **Introduction:**
   1. **Background Information:**

Air pollution is a very serious and growing concern as the quality of air keeps deteriorating all over the globe and this downfall in the quality will result in the increase of health-related risks. In the U.S. approximately 134 million are facing a risk of premature deaths or serious health issues due to air pollution, as per the American Lung Association [1]. There are many articles and research on air pollution daily trying to show the severity of the situation and spread the awareness among the public about the problem and its consequences. For example, an article in New York times about a research published in JAMA Psychiatry which concluded that high level of air pollution can cause mental health issues in teenagers [2].

There are many factors that cause the air pollution such as the pollution from various industrial factories, vehicle pollution, and other natural or accidental factor. According to an article in the Houston Chronicle, 48% of smog-forming nitrogen oxides came from automobiles as per the EPA figures [3]. Therefore, as a result of these growing concerns it becomes highly important that we are able to predict the air quality in advance to better prepare for deteriorating air quality. These predictions can be found more useful for the people will health conditions and also can be used to better suggestions for the hospital locations. Mostly, in big cities with increasing number of automobile and factories around these cities are main cause of high amount of air pollution in cities like New Delhi. Recently, New Delhi tested an odd and even automobile formula to control the growing air pollution, according to which only odd number cars will run certain days and even number cars will run other days. The pollutants that are the main cause of air pollution that include four type of pollutants which are NO2, O3, CO and SO2. These pollutants are mainly released by high number of automobiles and industries that surround the cities.

* 1. **Problem Statement:**

As a result of such growing issues, the demand for air quality prediction increases to predict the AQI at least one or two days in advance. Data Mining algorithms that are used for prediction can be applied to the data of these pollutants and we can predict the future AQI to get an estimate of increasing or decreasing air quality. In this paper we use machine learning algorithms (such as Decision Tree Regression and Decision Tree Classifier) to predict the near future value if the air quality index.

1. **Literature Review:**

Air Pollution is major problem and therefore, there are many research literatures focused on this topic like the Brunelli et al., 2007 [4] and Mesin et al., 2010 [5]. However, the feature selection is very common which include the temperature, solar radiation, humidity, wind speed etc. A large feature selection can cause the algorithm to become specific to a particular data set. Therefore, most relevant attributes are required to be considered for prediction and increasing the accuracy, so new methods/algorithms are required to be implemented [6].

Air pollution is a major issue in the USA and worldwide; therefore, there are many research literatures focused on this topic. The different methods of data mining used by the researchers for prediction of air pollution in the USA. Siwek & Osowski (2016) stated that past literatures and researcher clearly indicated the importance of data mining model to predict the air pollution in the future. The research used the genetic algorithm, linear method and decision trees of data mining to predict the air pollutions in the form of PM10 (particulate matters of diameters up to 10 μm), SO2, NO2 and O3. The finding of this research disclosed the preselection of the most important features to use data mining approaches allowed enhance the forecasting accuracy or predict of atmospheric pollution in the country. Moreover, the proposed approaches of the data mining (genetic algorithm, linear method and decision trees) help to better accuracy of prediction of air pollution in the country (Siwek & Osowski, 2016).

Moreover, Kurt & Oktay (2010) presented similarly reviewed and explained the different data mining methods for prediction of sulfur dioxide (SO2), carbon monoxide (CO) and particulate matter (PM10) to analyse air pollution. In this study, the authors applied the neural network architecture and Airpol system to identify the most important environmental factors that will enhance the air pollution in the future. The results of this research presented and discussed the 3 main air pollutants (CO2, SO2, and PM10) will be predicted the high air pollution in the future. These data mining models finding represents the forecasting high air pollution in urban areas in the future. This study concluded the air pollution in the future created more health issues among urban areas people compared to the rural areas people (Kurt & Oktay, 2010).

Pooja & Vanishree (2018) stated that the industrialization has improved human life, but it also negatively affected the environment by carbon emissions and enhance pollution. The research mentioned that carbon-dioxide, nitrogen oxide, excess levels of ozone, etc., are reasons of air pollution that a major threat to human life in the future. The downfall trends of air quality standards bad effects on environment and health of human being. Moreover, the various factors like PM 10, PM 2.5, NO2, SO2, are the most dangerous pollutant factors of air pollution. Random Forest Regression Algorithm, linear method of stepwise fit, and decision trees method of data mining applied in this research to predict air pollution in the coming years (Pooja & Vanishree, 2018).

In addition, Bellinger, Jabbar, Zaïane & Osornio-Vargas (2017) predicted that the past literatures using different data mining techniques and machine learning methods to measure and predict the airborne pollutants in the air that cause air pollution. The authors mentioned that the recent research articles using various data mining techniques and machine learning methods to forecast/predict of air pollution/quality, such as the APRIORI algorithm, genetic algorithms, k-means clustering, support vector machines, artificial neural network (ANN), natural language processing (NLP), bayesian hierarchical models, and decision trees. This research concluded that data mining techniques are becoming a common tool in identifying the current trends, challenges and new directions to explore of air pollution in the US. Moreover, this research shows that data mining technique forecast/predict the air pollution continues to grow in the US in the future (Bellinger, Jabbar, Zaïane & Osornio-Vargas, 2017).

Ma, Richards, Ghanem, Guo & Hassard (2008) clarified the transport has a significant impact on climate change, noise pollution, watercourse pollution and local air quality that impact upon the environment. The MoDisNet system, Grid computing technology, and P2P e-Science Grid architecture are data mining techniques that integrated approach used by the researcher to find the relationships between urban transport and the air pollution. Moreover, the research mainly concluded that integrated all three data mining techniques identify the urban air pollution and capturing the environmental change. This research identified the significant relation in traffic or transport and air pollution in the east London. Furthermore, integrated data mining techniques predicted that trend of enhance in traffic in urban areas, like east London cause of high pollution in the future (Ma, Richards, Ghanem, Guo & Hassard, 2008).

1. **Data Description:**
   1. **Data Source:**

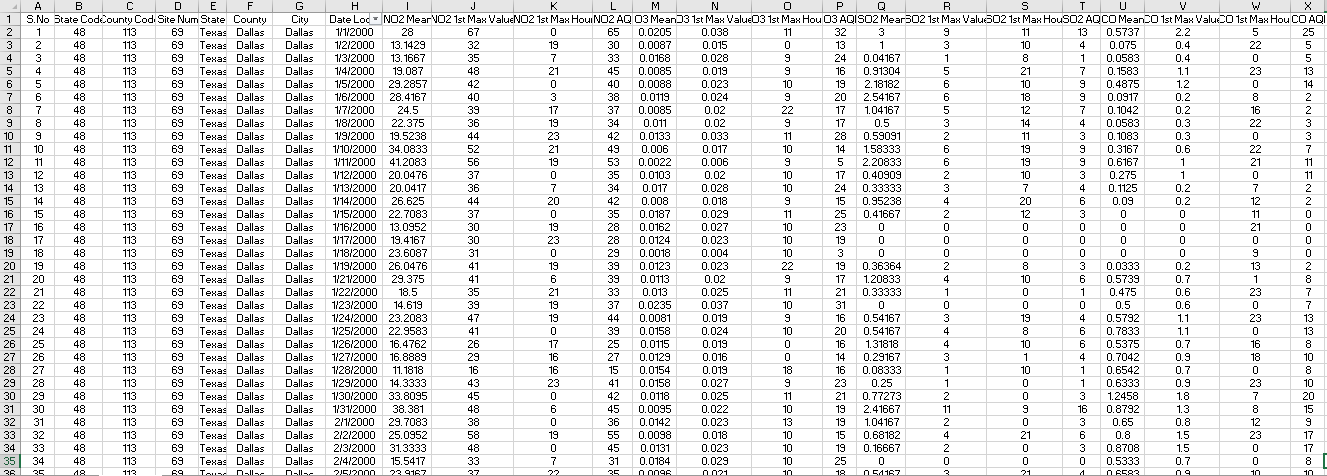
The data was collected from the Kaggle website (<https://www.kaggle.com/sogun3/uspollution>) [7]. This data was available for public which was collected from the United States Environmental Protection Agency (U.S. EPA) and include the air quality index data about 4 pollutants (Nitrogen Dioxide, Ozone, Sulphur Dioxide and Carbon Monoxide) that are primary constitute of the air pollution.

* 1. **Data Size:**

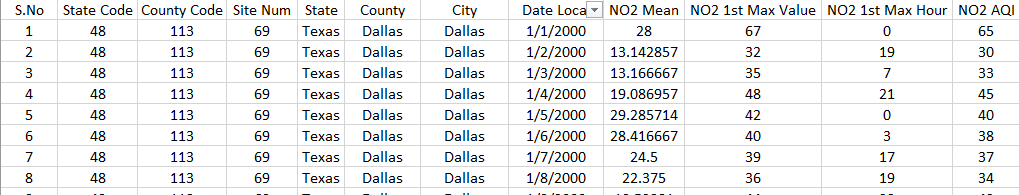
The original data was approximately 1.4 million as it included the data from all over the United States. As, this project will be focused on predicting the air quality in Houston area, therefore, the data sets for this project will be approximately 4,000 values from year 2000 – 2009. The data that was used for this project has been truncated from the original data just to extract the air pollution data of Houston.

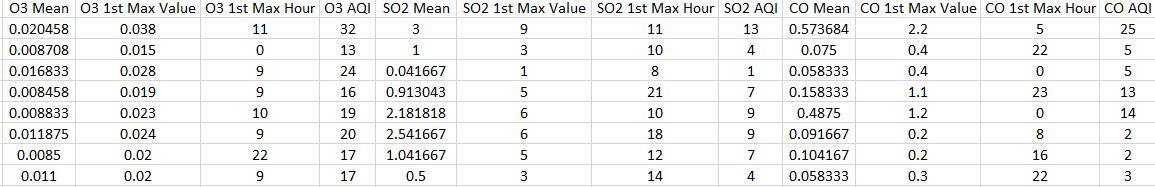
* 1. **Data Visualization:**

In this Houston data set, there are 4,000 rows and 24 columns. These columns are S.NO, State Code, County Code, Site Num, State, County, City, Date Local, NO2 Mean, NO2 1st max value, NO2 1st max hour, NO2 AQI, O3 Mean, O3 1st max value, O3 1st max hour, O3 AQI, SO2 Mean, SO2 1st max value, SO2 1st max hour, SO2 AQI, CO Mean, CO 1st max value, CO 1st max hour, CO AQI. In the excel file, the structured data looks like the following figure.



For better visualization of the attributes, the Zoomed in screenshots have been shown here:

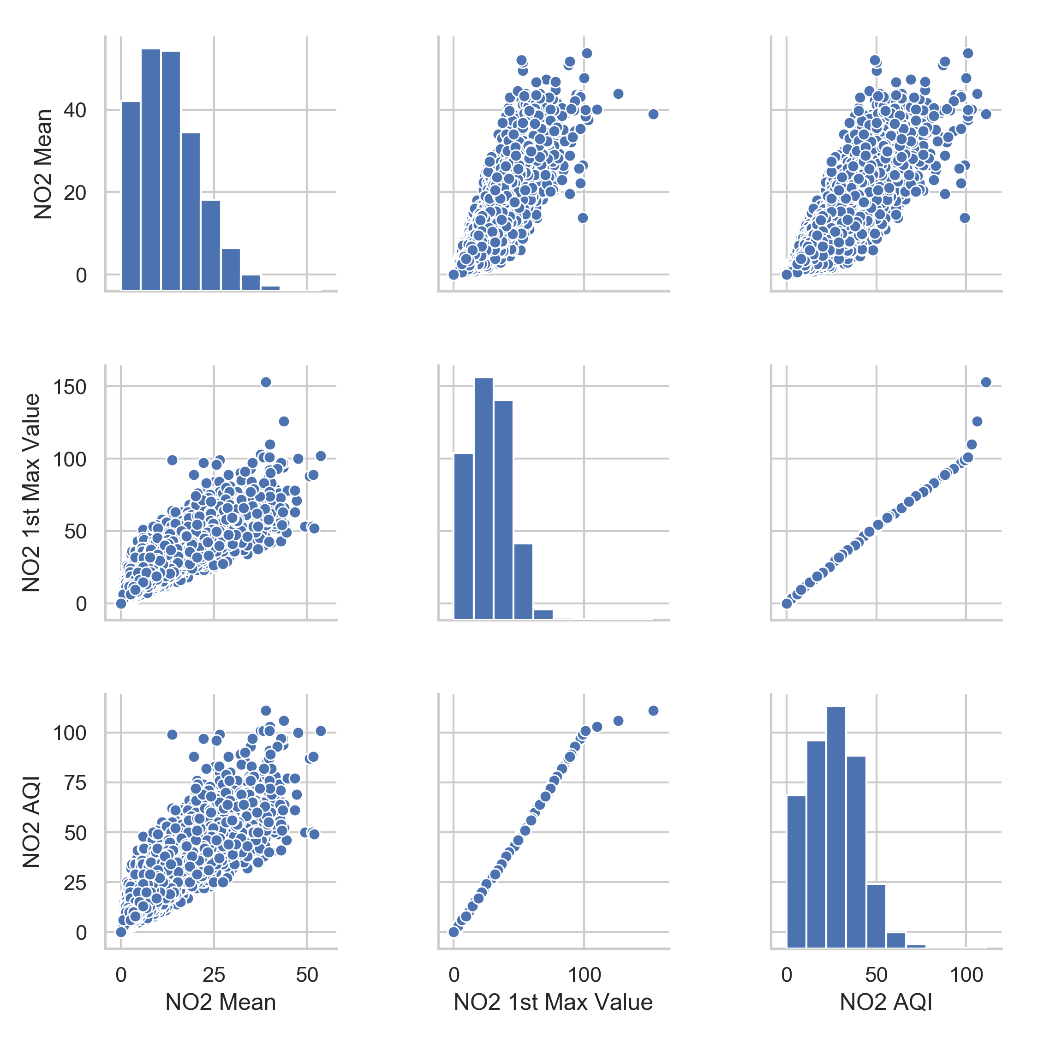




Before training a machine learning model, it is very important to some graphical Exploratory Data Analysis using Matplotlib. It helps us to detect the outliers in the data, the correlation between the features, and the distribution of it.   
Here, we have created scatterplot matrix for NO2, O3, SO2 and CO separately to see how the respective features (Mean, 1st Max Value and AQI) are pairwise correlated. The pairplot function from the Seaborn library have been used. The following figures show the correlation between the features for individual gas.

In the scatterplots below we will see that there is high correlation between the Mean - 1st Max value and Mean – AQI. But there is a perfect correlation between 1st Max value and the AQI. To quantify the correlation we have plotted a correlation matrix for each of the set of features.

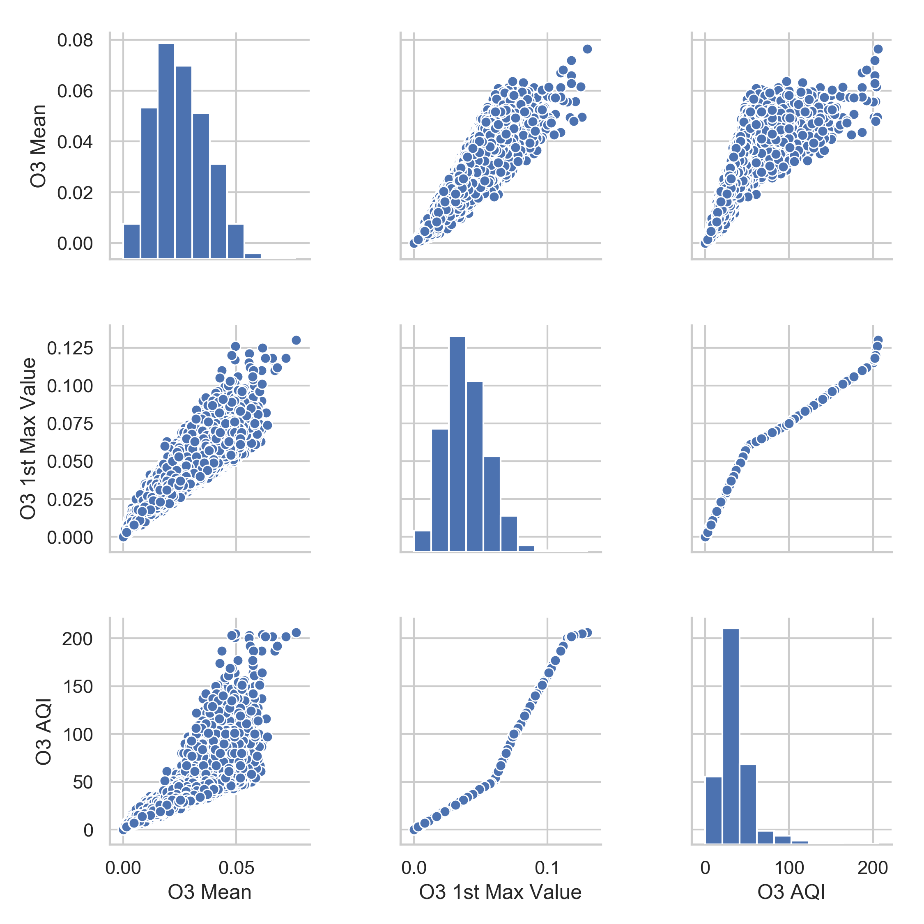
For NO2:



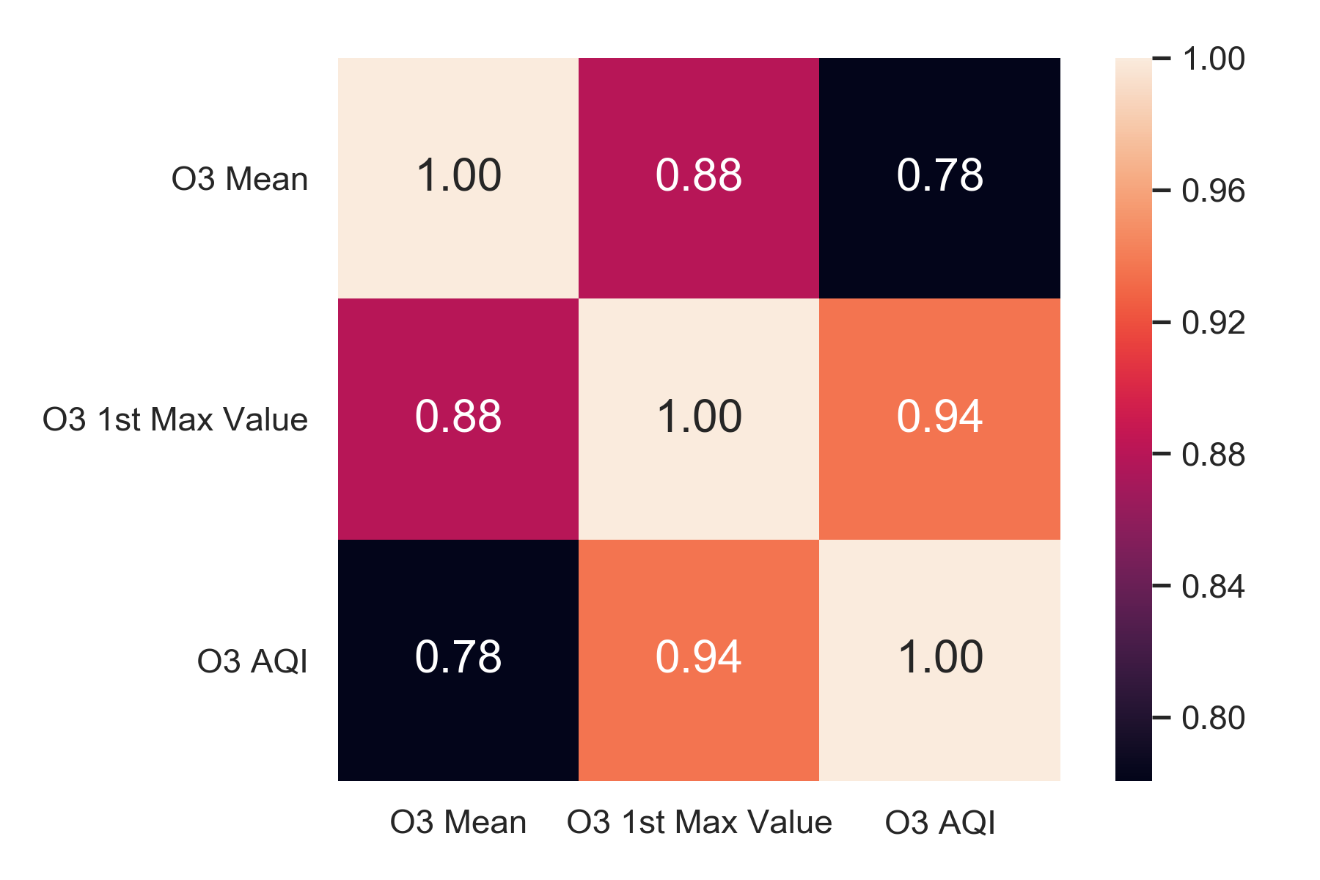
The correlation matrix has been shown with a colormap here:



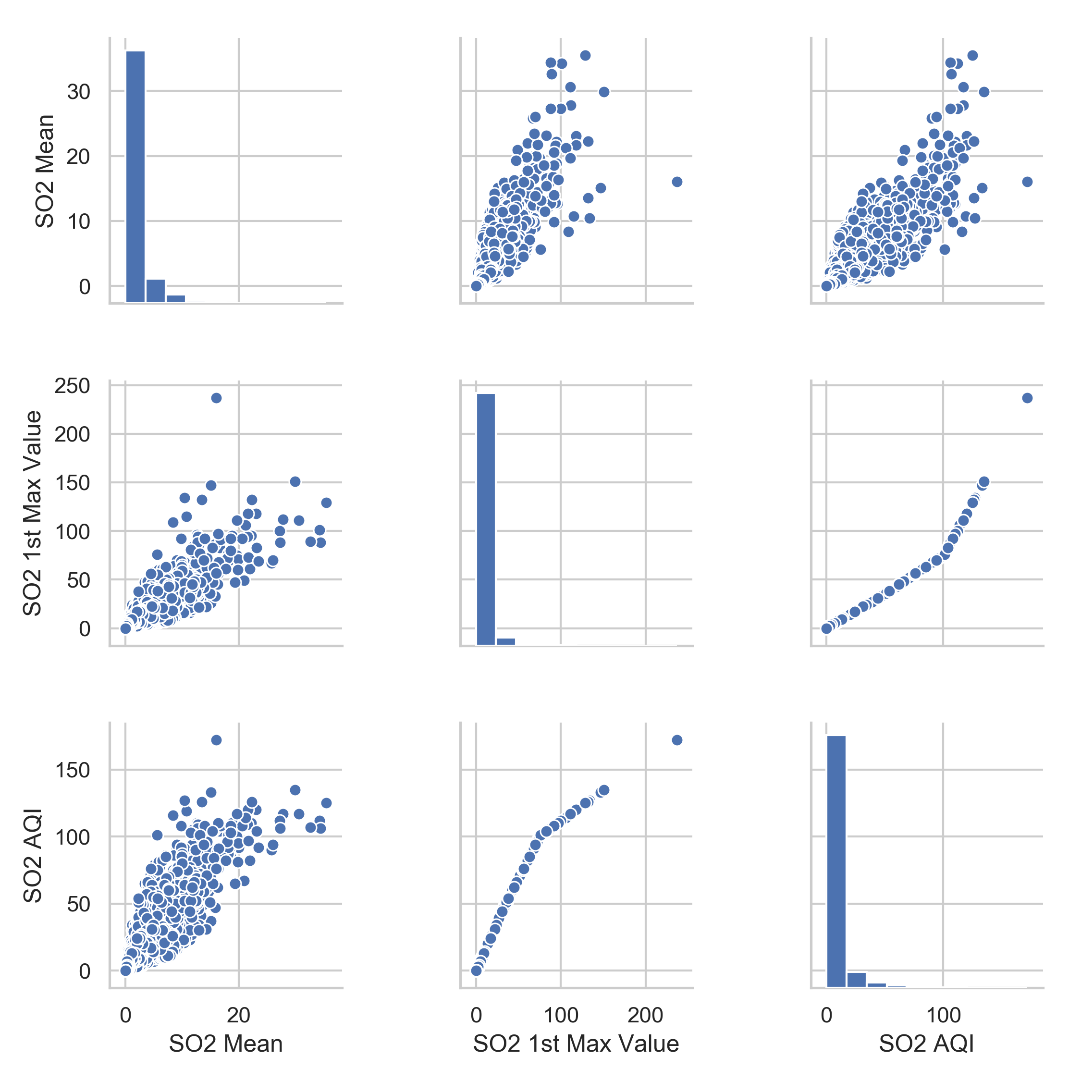
For O3:



The correlation matrix has been shown with a colormap here:



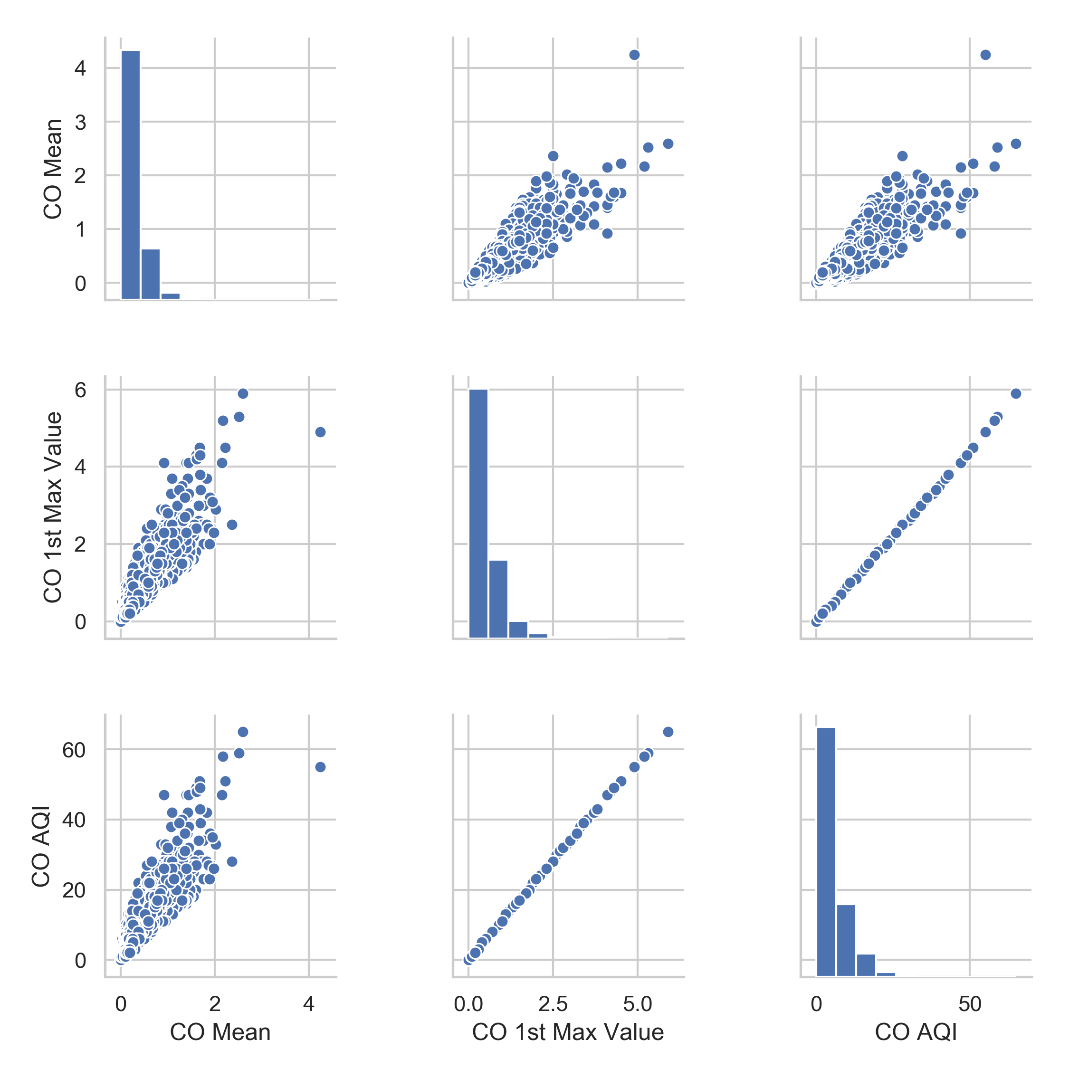
For SO2:



The correlation matrix has been shown with a colormap here:



For CO:



The correlation matrix has been shown with a colormap here:

