US Air Pollution - Environmental Dataset

Mogollon Monster Data Analysis

Team Members: Nibedita Bal, Zach Paquin, Upendra Rajendran, David Terracino

The following performance metrics are used to evaluate the performance of the models discussed in this project data set.

5 Number Summary:

The 5 Number summary below represents 4 environmental air pollutants Nitrogen Dioxide NO2, Ozone O3, Sulfur dioxide SO2, and carbon monoxide CO. The performance metrics is used for descriptive analyses or the preliminary investigation gives the minimum, 25 percentile (Q1) , median (Q2), 75 percentile (Q3) and maximum values on each pollutants in the dataset and helps to provide a quick check on the distribution observations and AQI's of the data. The summary function also adds the mean of the data, which can be compared to the median to identify any skewness in the data. With this 5 numbers summary, all air pollutants AQI's holds the min of 0 to max of 200 range and gives an insight of the data is right skewed. This helps to understand and categorically find the air quality health index.

This five-number summary can be further represented in a Box Plot which is explained in the following parts

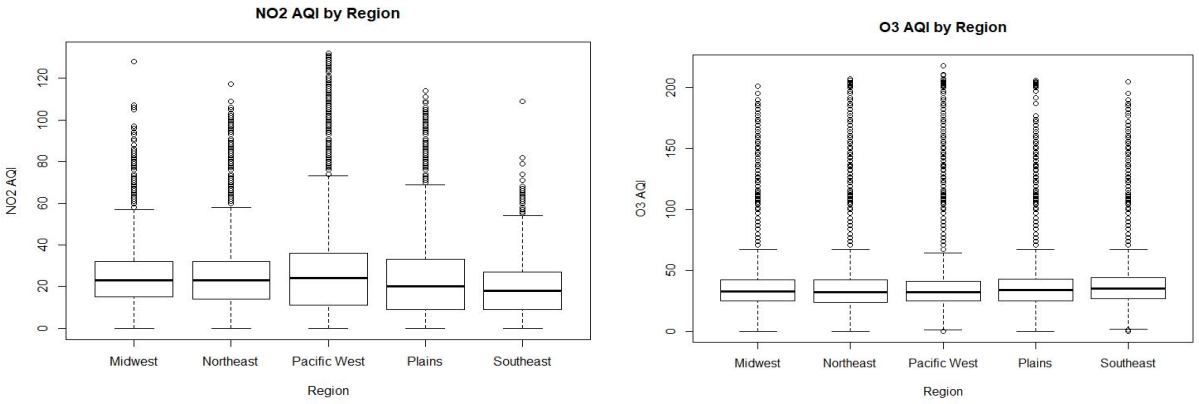
Variable	Min	Q1	Median	Q3	Max
NO2.Mean	-2.00	5.75	10.74	17.71	139.54
NO2.1st.Max.Value.ppb	-2.00	13.00	24.00	35.70	267.00
NO2.AQI	**0**	**12**	**23**	**33**	**132**
O3.Mean	0.00000	0.01787	0.02587	0.03392	0.09508
O3.1st.Max.Value.ppm	0.0000	0.0290	0.0380	0.0480	0.1410
O3.AQI	**0**	**25**	**33**	**42**	**218**
SO2.Mean	-2.0000	0.2652	1.0000	2.3333	321.6250
SO2.1st.Max.Value.ppb	-2.000	1.000	2.000	6.000	351.000
SO2.AQI	**0**	**1**	**3**	**9**	**200**
CO.Mean	-0.4375	0.1917	0.2958	0.4708	7.5083
CO.1st.Max.Value.ppm	-0.4000	0.2000	0.4000	0.7000	15.5000
CO.AQI	**0**	**2**	**5**	**8**	**201**

US Air Pollution - Environmental Dataset

Mogollon Monster Data Analysis

Box Plots:

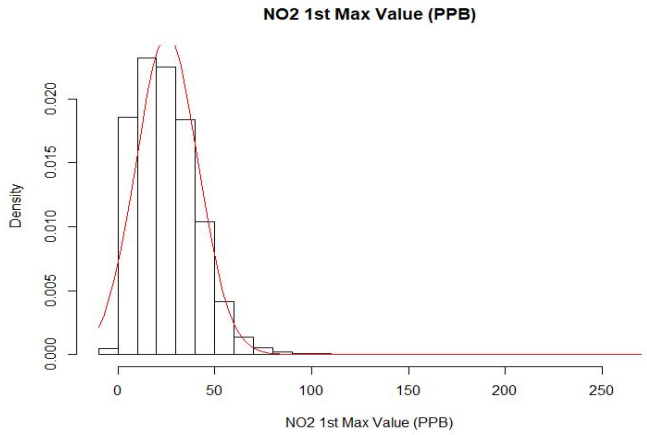
Below are two box plots of the pollutants NO2 and O3 Air Quality Index (AQI) assessed by 5 regions; Midwest, Northeast, Pacific West, Plains and Southeast respectively. Immediately, our first observations are the significant amount of NO2 outliers above the Q3 interquartile range for every region. Similarly, we observe the same evidence of O3 outliers above the Q3 interquartile range by region in the second box plot. Further analysis will be conducted to determine the viability and reasoning of the outliers. Potentially more data or additional insights in the collection methodology may be required when assessing outliers. Next, the mean for NO2 AQI by region remains relatively close to + or - approximately 5 Parts per Billion (PPB) from 20 PPBs. Subsequently, the O3 mean by region is even closer, around 40 parts per million (PPM).

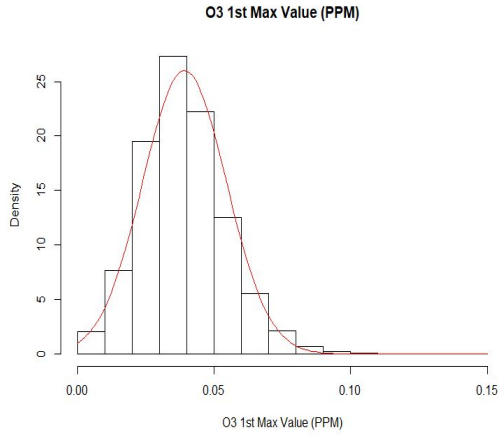


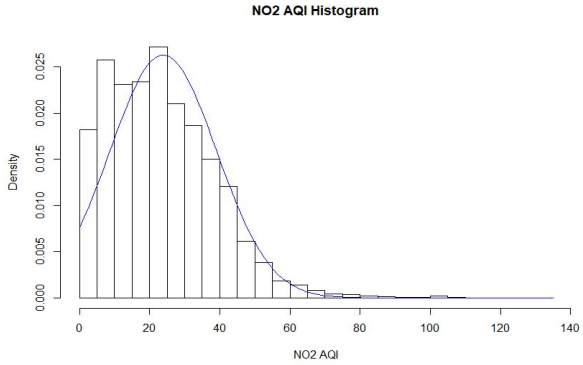
Histograms:

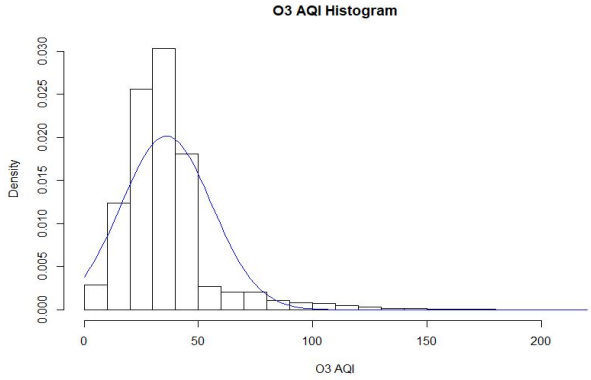
The metrics shown below demonstrate the distribution of two pollutants NO2 and O3 respectively. The first two visualizations are NO2 (parts per billion) and O3 (parts per million) "first max value" readings taken across multiple locations throughout the continental United States, and Hawaii. The initial observations demonstrate a heavily skewed right distribution of variables indicating outliers above the Q3 range. In the next two histograms were observed the NO2 and O3 Air Quality Index (AQI). Similarly, the observations reveal heavily skewed right values further demonstrate potential outliers above the Q3 interquartile range as well. Next, when comparing NO2 (AQI) and NO2 (first max values) based on distribution, we can determine a potential direct relationship among the values. In addition, the same assessment can be made between the O3 AQI and O3 first max value as well. Further evaluation of scatter plots

substantiates our observations. Lastly, normalization of values such as logarithmic scaling may be required to handle outliers in data.





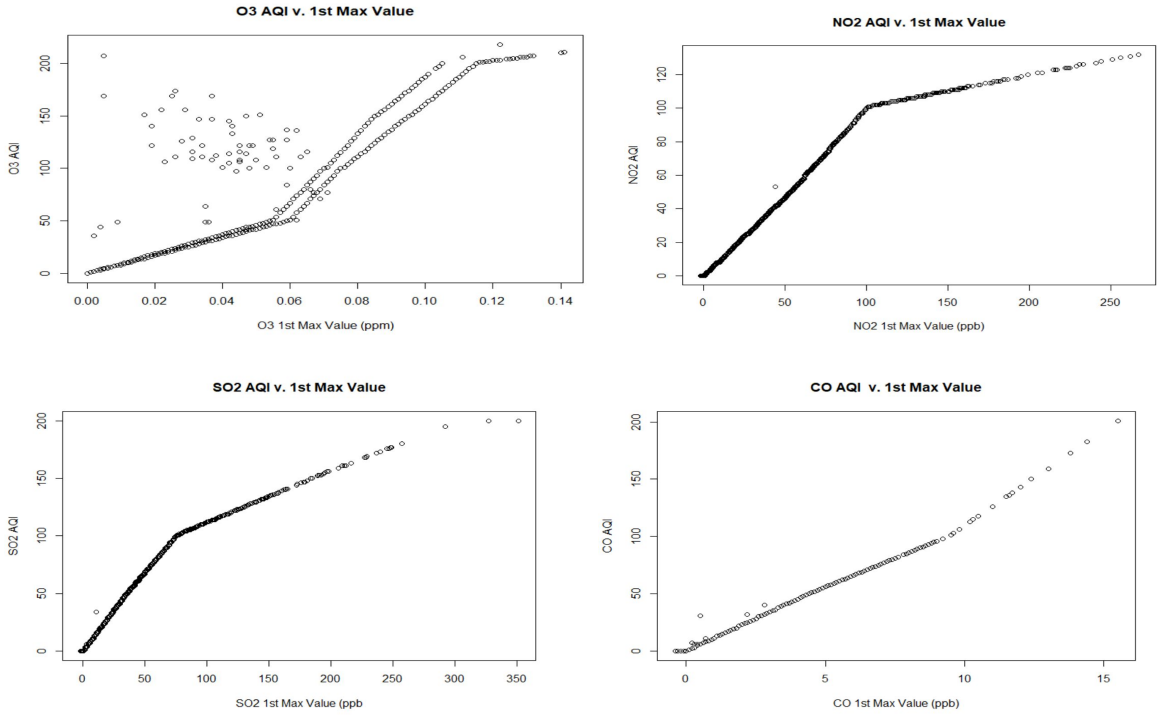




Scatterplot:

Below scatter plot for all 4 air pollutants shows an overall positive, strong and linear correlation. The scatterplot between O3 AQI in Y-axis and 1st max value in X-axis shows a linear , positive, and strong association with a slight spread in the middle and shows a steady relationship at the end and also, the line has three linear patterns. The regression line has one linearity in the range of 0-50 in the X-axis, then from 0.05-0.06, and so on. Between NO2 AQI and 1st maximum value shows a linear association. The direction is positive and has good air quality strength. The

scatterplot between SO2, CO AQI and 1st maximum value shows a linear association. The direction is positive and has good air quality strength.



Conclusion:

Based on the initial analysis, this data set helps to identify several models which can predict the AQI levels and classify them based on 4 different pollutant bands (NO2, SO2, CO and O3) and their individual contribution to the AQI response variables can be experimented.

In the future scope, the exploratory data analysis will be implemented for the prediction models (multiple regression on first order and second order models) to get the interesting correlations on air pollutant data in the USA. This could obtain several notable outcomes from the predictive models that are worth being discussed...