* **Statistical/ Hypothetical Questions**: which of the above listed air pollutants is released the most to the atmosphere? Carbon monoxide (CO), Nitrogen dioxide (NO2), Sulphur dioxide
* **Outcome of your EDA.** Since the dataset was large with over 16 years of data, the EDA process was followed to go through each variable and understand their significance better. For example, NO2 and SO2 were measured in billions and O3 and CO were in millions. I therefore grouped the pollutants with the same units together
* **What do you feel was missed during the analysis?** I focused on getting codes to work first before any analysis. There are several things I would have loved to discuss but got stuck because codes were not working. So, I concentrated only on those items where the code gave an output. For example, I wanted to do a regression analysis to come up with a line of best where which could help in an extrapolation of the data analysis, where we could calculate the AQI knowing the pollutants. Pollutant values given in billions and millions was not making things any easy
* **Were there any variables you felt could have helped in the analysis?** My dataset was large, it started on 1/1/2000 and went to 12/31/2016 and would have trouble with Excel opening it up and there were also outliers. Interstate pollution would have been another good element to look at. Where the pollutants originate from neighboring states.
* **Were there any assumptions made you felt were incorrect?** In the beginning I assumed all 4 pollutants had the same measurements and started comparing them equally, then I noticed the difference in measurements (billion and million) and I had to change my methods.
* **What challenges did you face, what did you not fully understand?** Some of the challenges I faced were when it came to the analysis when I created the graph I wanted and then I needed to understand the significance of what it meant. I needed to go back and review the terms to make sure I was using the proper graph. I would have also loved to do a regression analysis and not just the scatter plot. I had an idea on how the line of best fit would look like. But it was going to be challenging coming up with the equation Y= mx + b