Capstone 1 - Data Story

Author: Varsha Gopalakrishnan

1) Figure 1 is a plot showing all the monitoring data points on a map for Black Carbon (BC). This map shows that there are very few points (close to highways) where the concentration of black carbon is as high as 8 ug/m3. The average concentration of blackcarbon in the area is approximately 0.72 ug/m3. Similarly, Figure 2 is a plot of monitoring data points for Nitrogen dioxide (NO2). The average concentration of NO2 is ~17 ppb, with concentration as high as 60 ppb in some points close to highways.

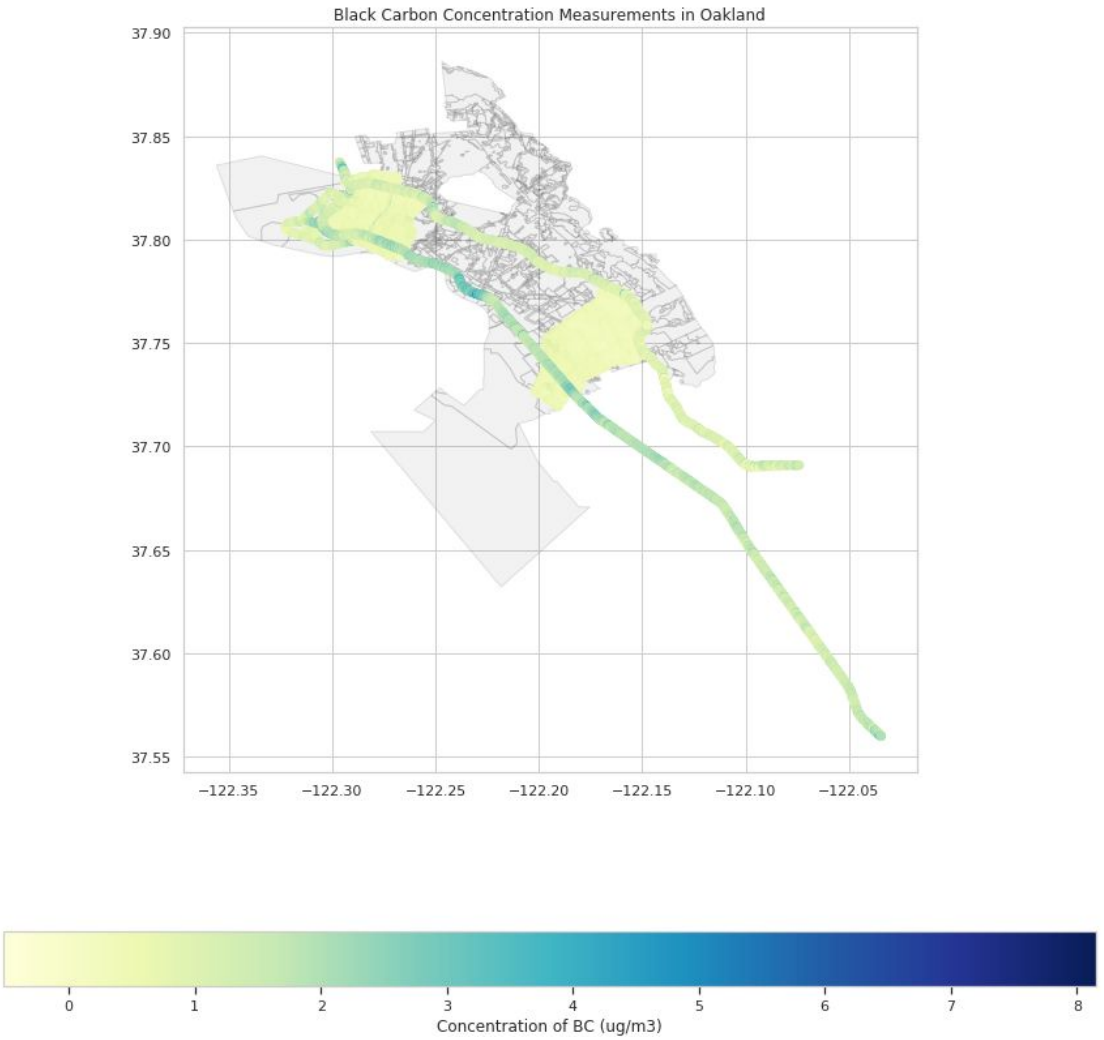

Figure 1: Concentration of Black Carbon Measured at Different locations

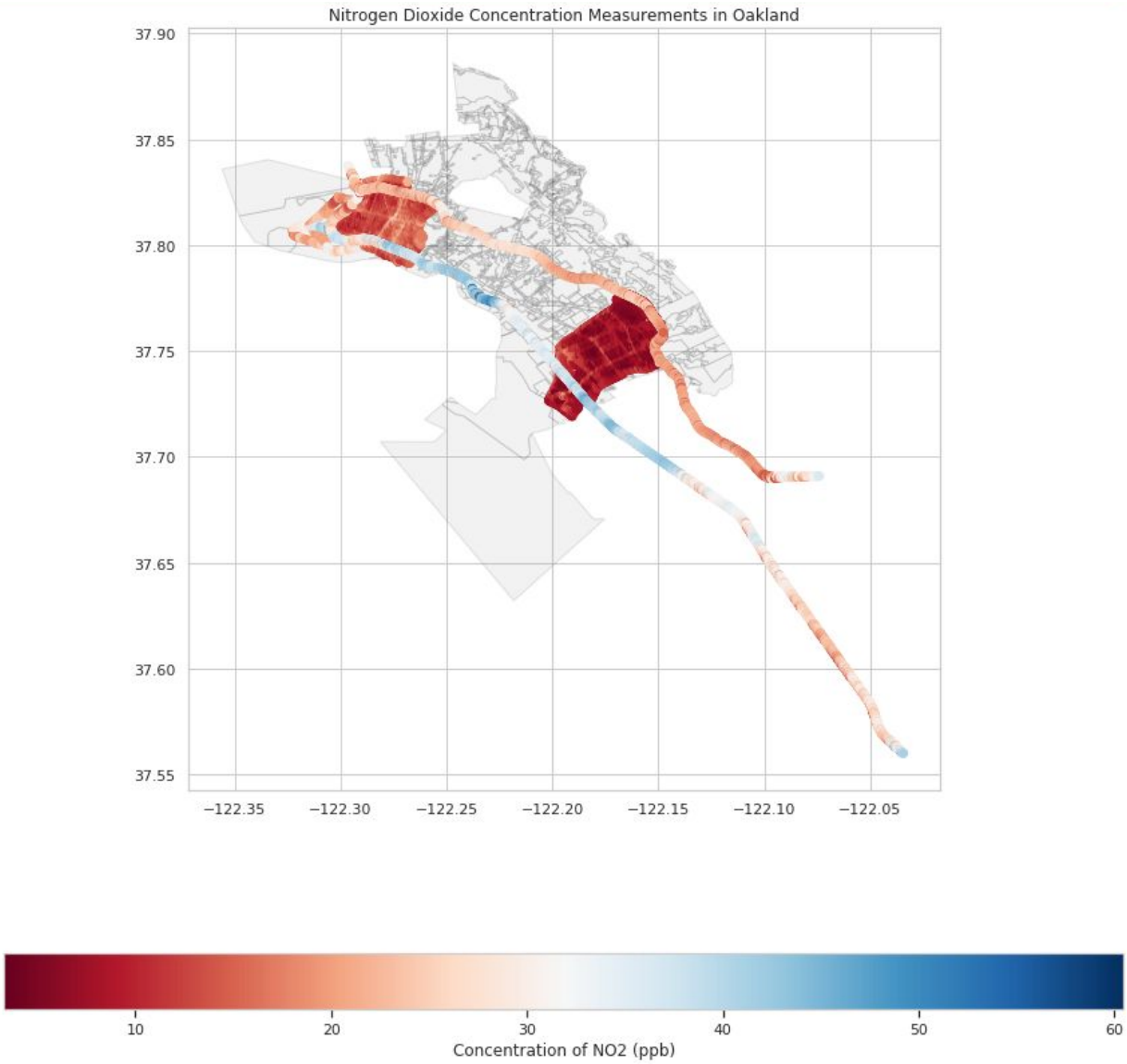

Figure 2: Concentration of Black Carbon Measured at Different locations

2) The hypothesis that I'm exploring here is air pollution concentration in a given location is correlated with the sources that emit air pollution, traffic (background concentration) and meteorological parameters. The hypothesis tree that I'm exploring is shown in *Figure 3* below.

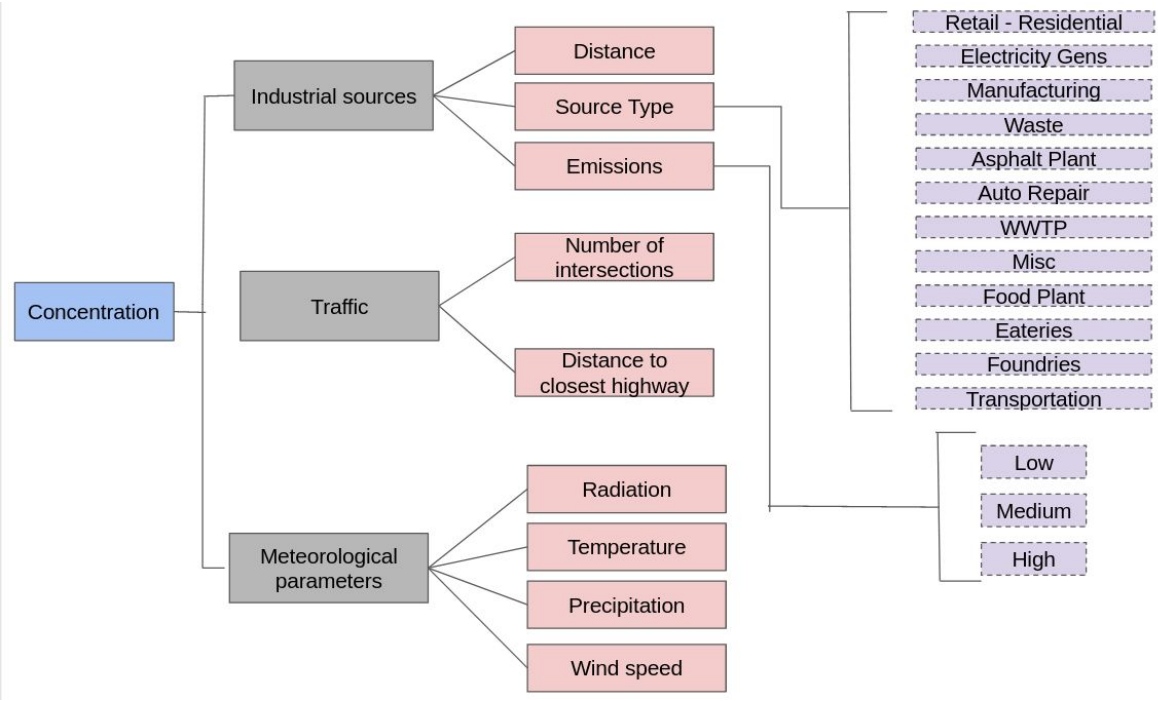

Figure 3: Hypothesis Tree

3) To explore whether ***air pollution is correlated with sources that emit air pollution***, it was necessary to plot histograms to understand the distribution of sources. *Figure 4 and Figure 5* show the histogram and transformed data for BC and NO2, respectively. The distribution shows that there are several small sources of emissions, that make up more than 80% of the dataset. Creating a boxplot by source type confirms the same. The boxplot shown on *Figures 6* and 7 for BC and NO2, respectively, indicate that all the small sources of emissions (with emissions < 5 tons) can be combined into larger source groups in order to minimize multicollinearity. Combining the sources into larger groups resulted in a boxplot as shown in *Figure 8 and 9*. By observing the boxplot, one can see that there are some very large sources of emissions, and some relatively smaller sources. We can further assign a categorical variable to each source as ‘low’,’medium’ and ‘high’ depending on their quantiles. Low indicates that emissions are lower than first quantile, medium indicates emissions is between first and third quartile, and high indicates emission is above third quartile.

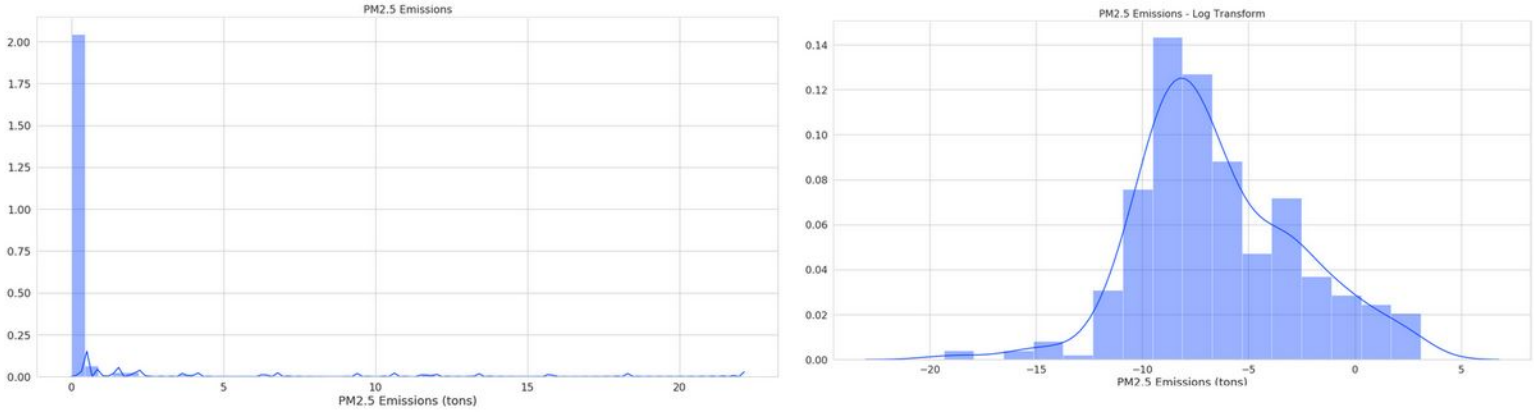

Figure 4: Histogram of PM2.5 Emissions

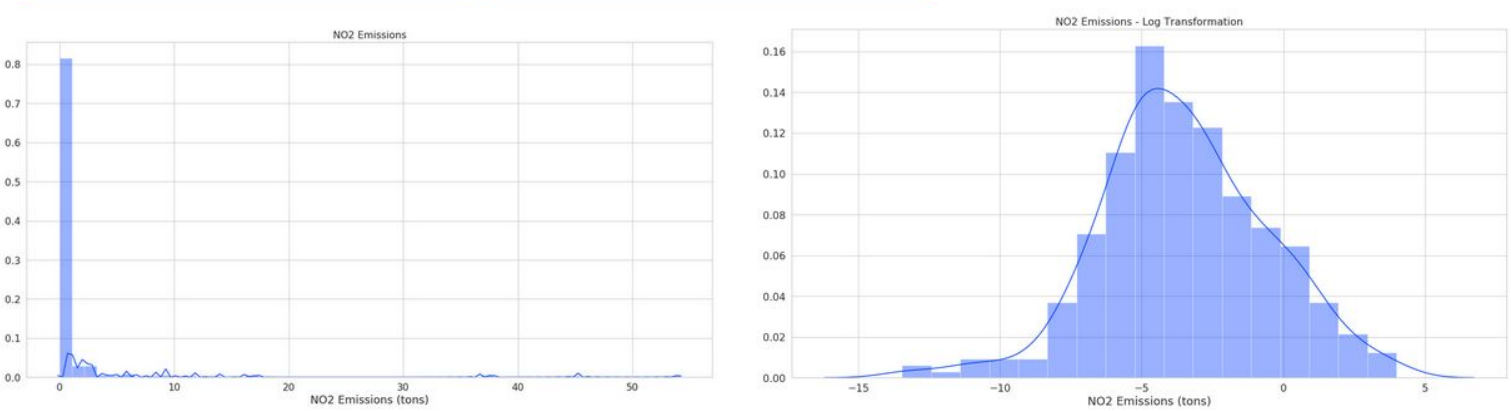

Figure 5: Histogram of NO2 Emissions

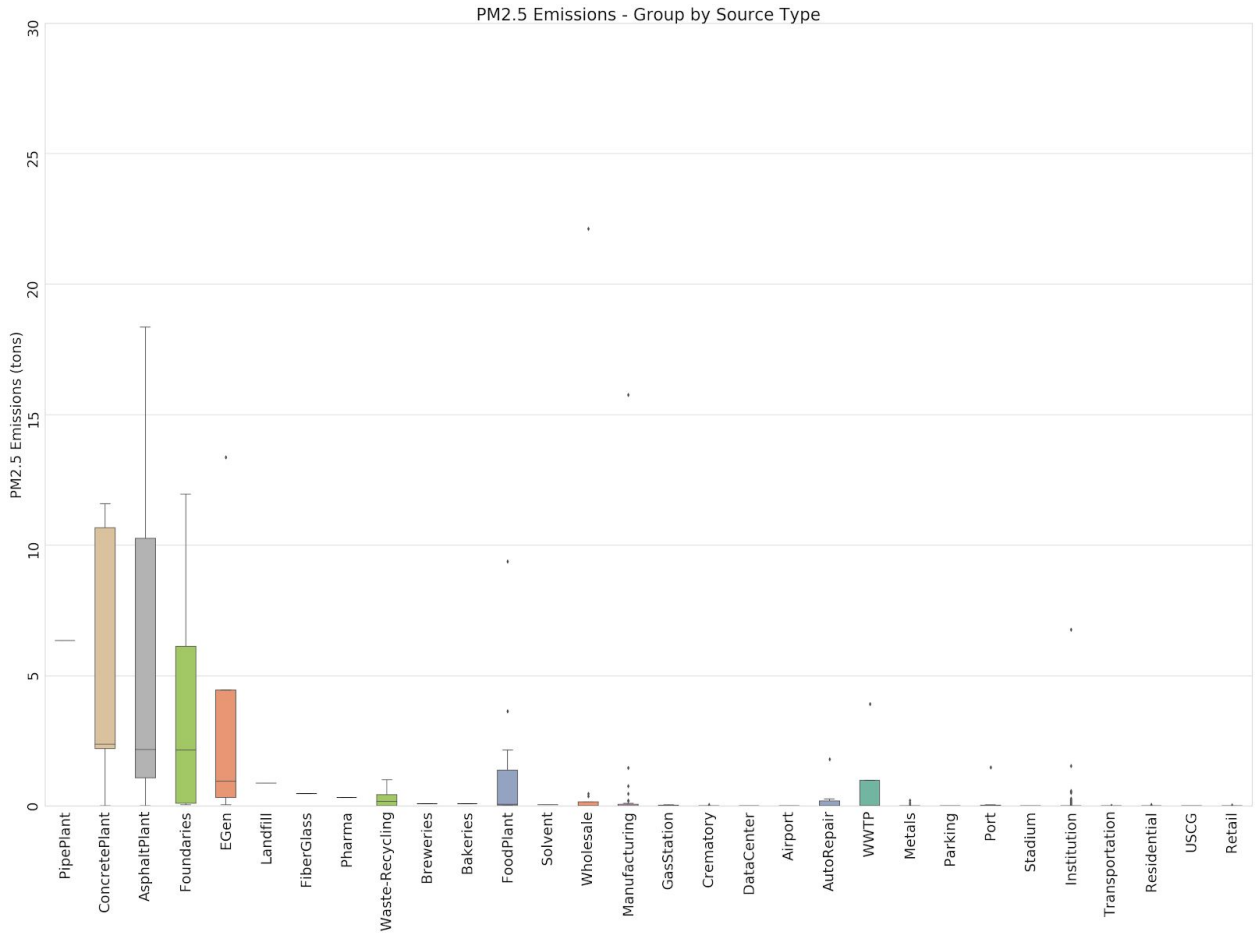

Figure 5: Boxplot of PM2.5 emissions by different source types

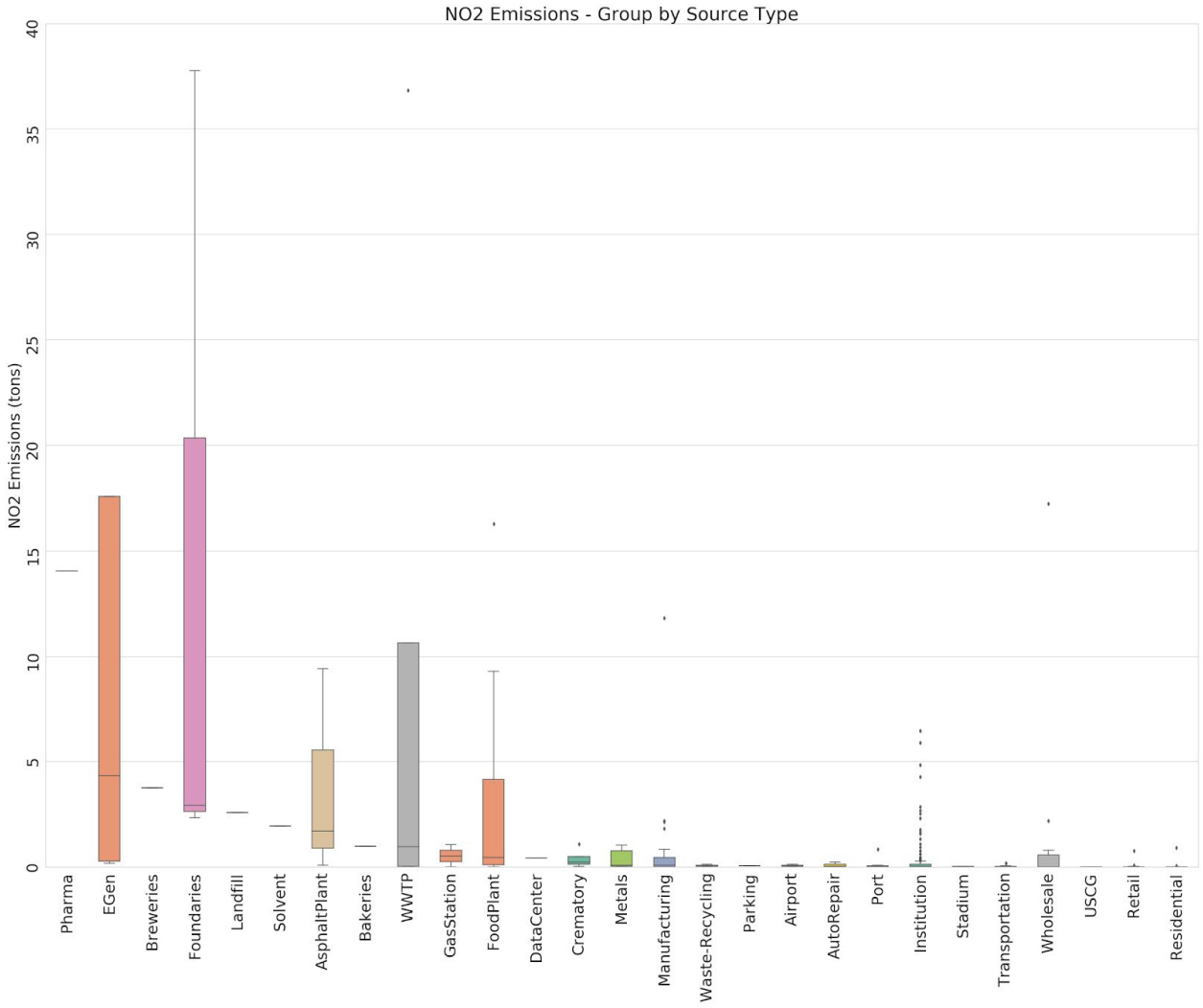

Figure 7: Boxplot of NO2 emissions by different source types

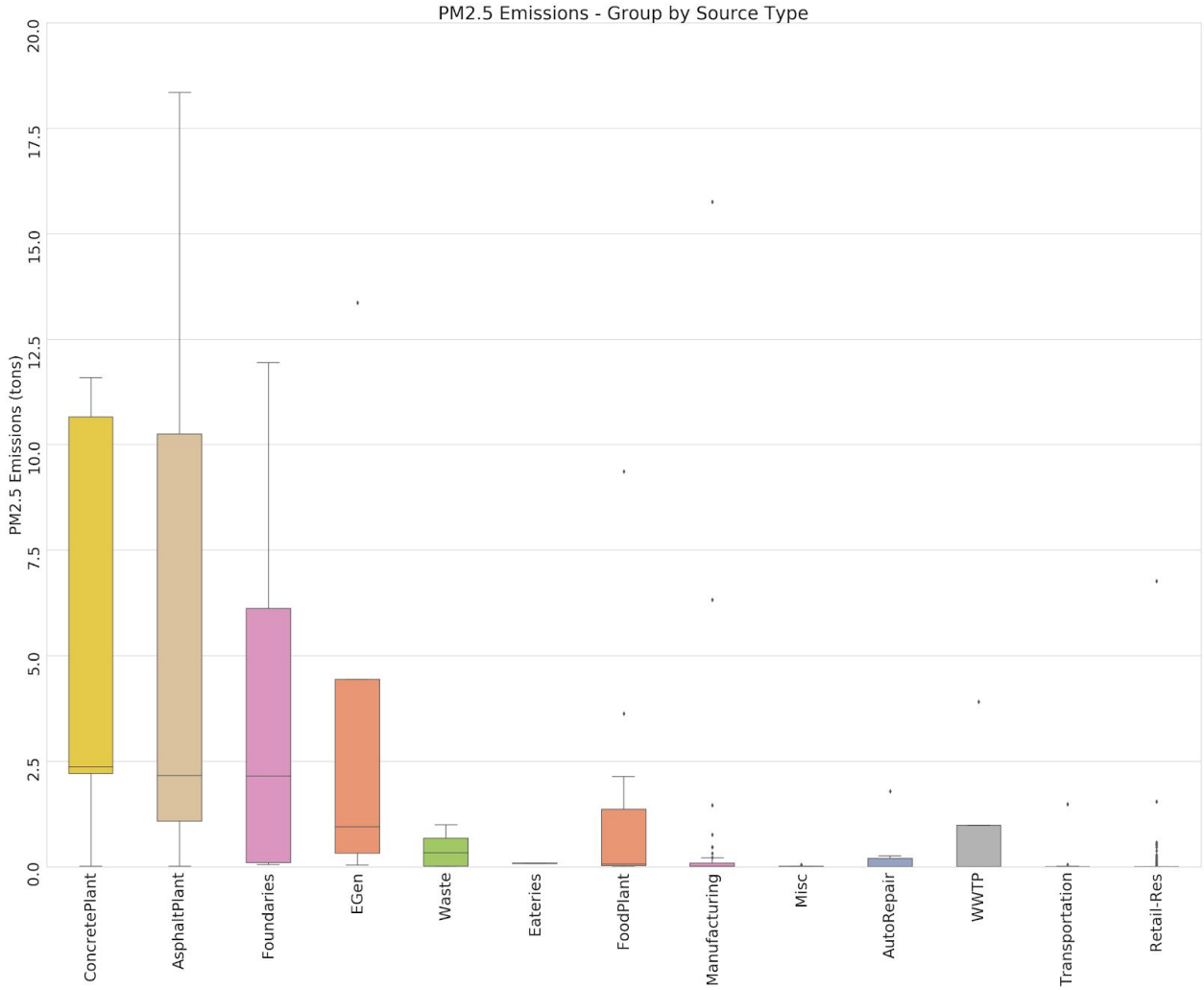

Figure 8: Boxplot of PM2.5 emissions by different source types - grouped

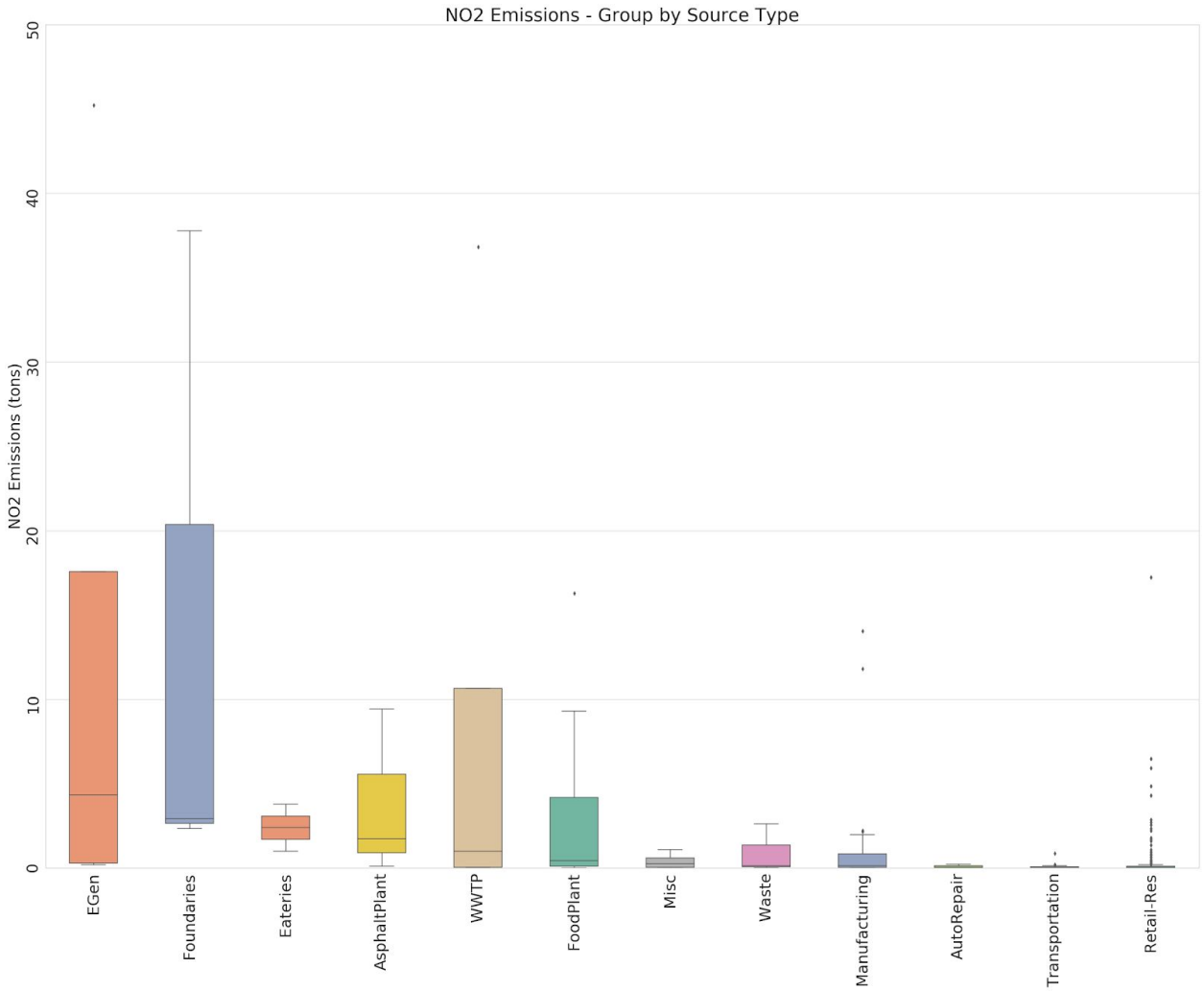

Figure 9: Boxplot of NO2 emissions by different source types - grouped

4) The next question that I'm exploring is ***how does air pollution concentration vary with distance to these sources of emissions? Is concentration directly correlation with distance, or to (distance)^2 or to emissions/distance?***

   A plot of BC concentration vs distance to the largest source as shown in *Figure 10* indicates that there could be a weak correlation with concentration and distance. A similar trend was observed for the plot of NO2 concentration vs distance as shown in *Figure 11*, but the Pearson r value for this plot is very low. A plot of concentration vs distance^2 for NO2 showed slightly higher correlation, as shown in *Figure 12*. A plot of concentration vs. emissions/distance as shown in *Figures 13 and 14* indicate there is almost no correlation between the two parameters.

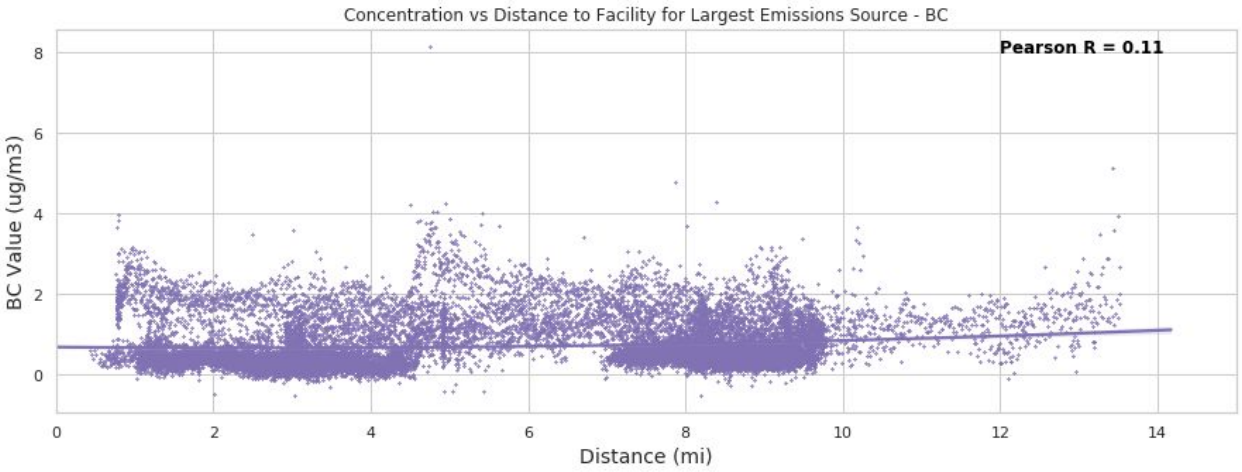

Figure 10: Scatter plot of concentration vs distance to facility for largest source - BC

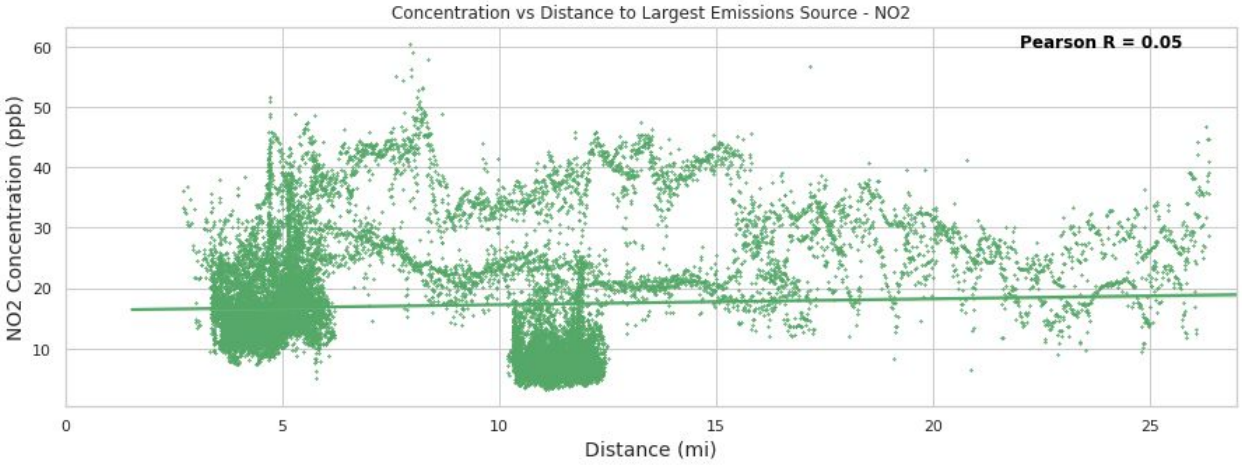

Figure 11: Scatter plot of concentration vs distance to facility for largest source - NO2

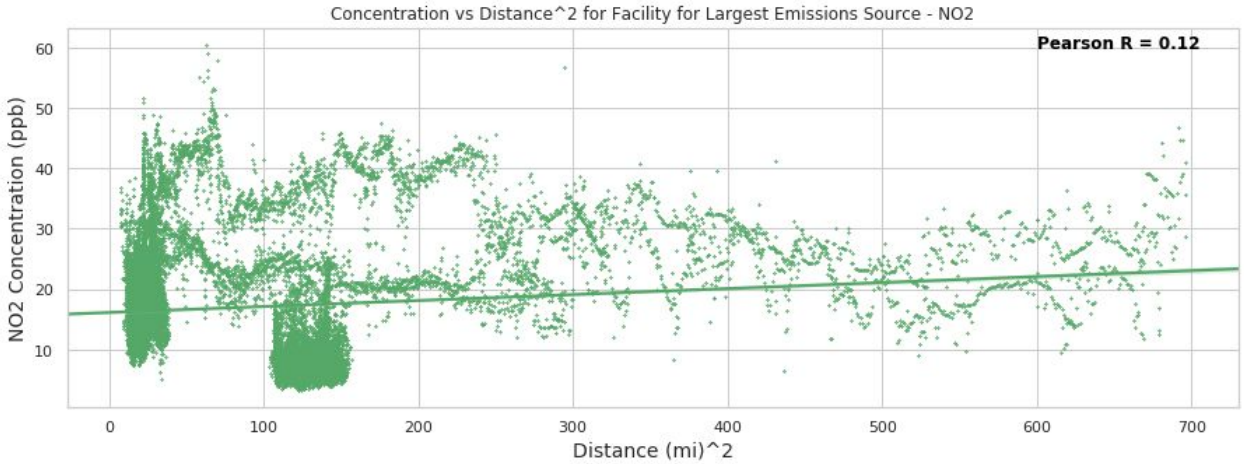

Figure 12: Scatter plot of concentration vs distance^2 to facility for largest source - NO2

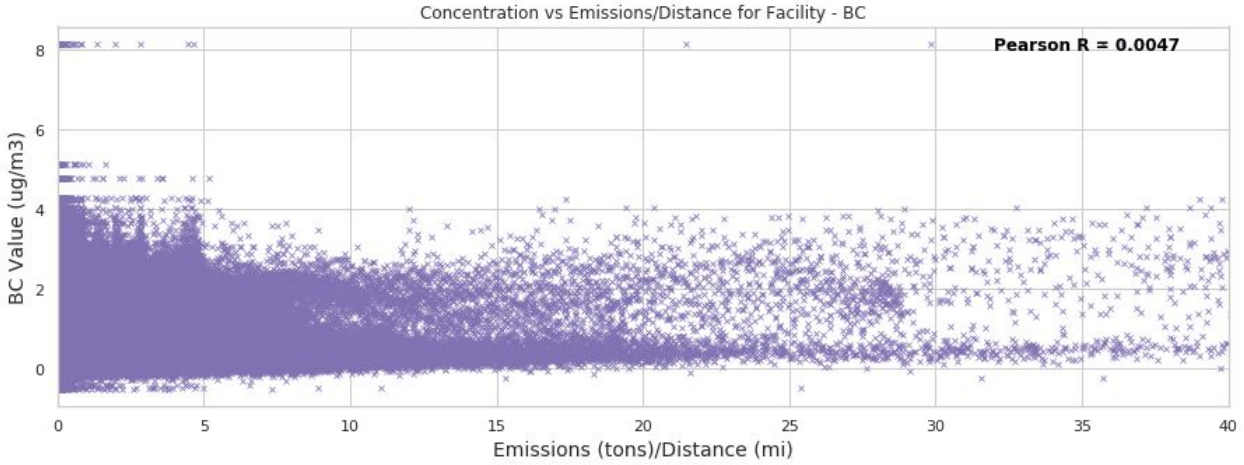

Figure 13: Scatter plot of concentration vs emissions/distance for each facility - BC

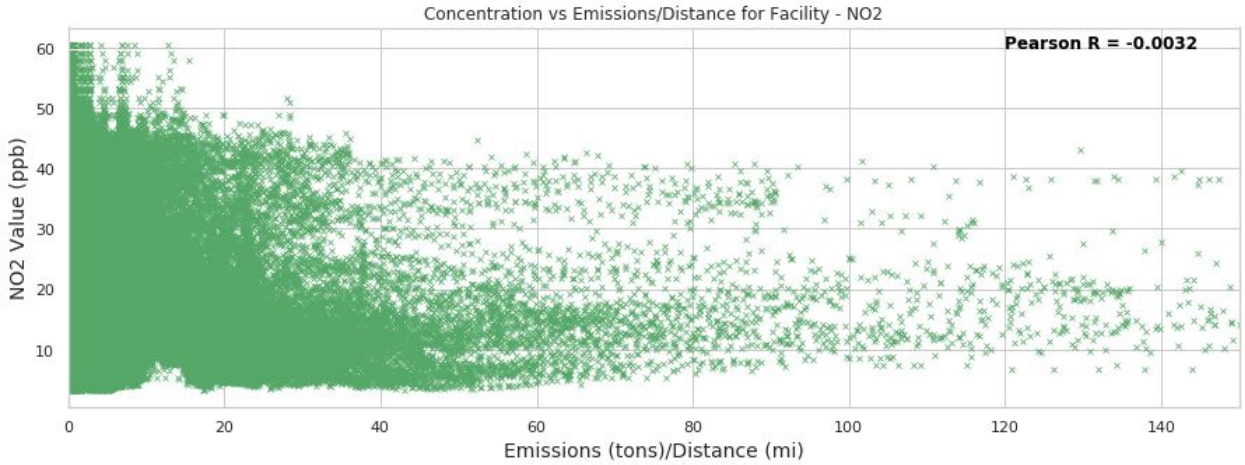

Figure 14: Scatter plot of concentration vs emissions/distance for each facility - NO2

5) Next, we explore correlations with traffic data. A traffic score for each monitoring location is calculated based on the total number of traffic intersections within a 1,000 ft. buffer. The traffic score varied from 2 to 35, indicating that there are a minimum of two traffic intersections, up to a maximum of 35 intersections. *Figures 15 and 16* show the **correlation between concentration and number of traffic intersections**. The plot indicates that the high concentration does not necessarily mean a larger number of traffic intersections indicating there could be other factors like number of vehicles that pass through an intersection, and other background sources that are contributing to high concentrations. However, we don't have enough data to confirm how other parameters are contributing to concentration. A correlation between distance to closest highway and concentration is also explored in *Figures 17 and 18*. The scatter plot indicates that a closer distance is correlated with higher concentration, at least for NO2.

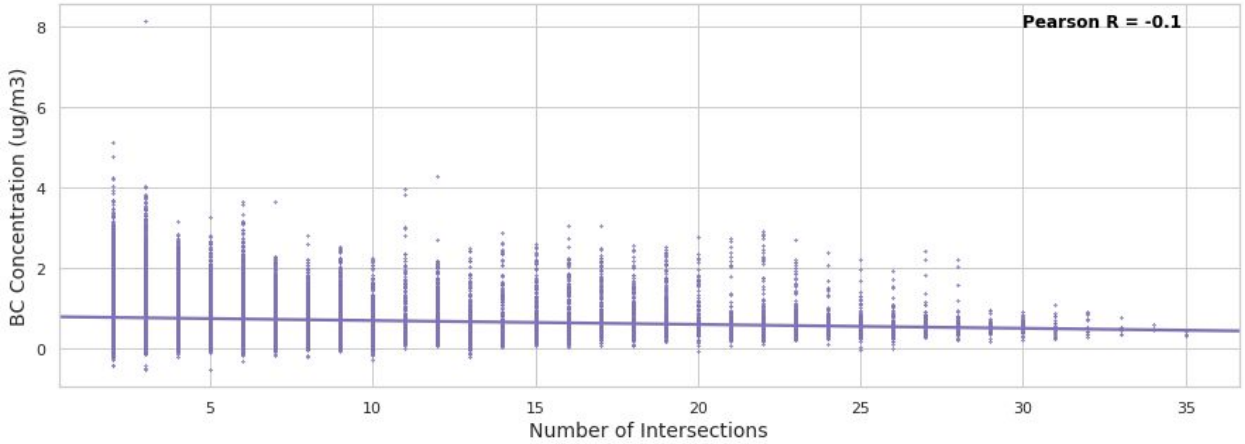

Figure 15: Concentration vs number of traffic intersections - BC

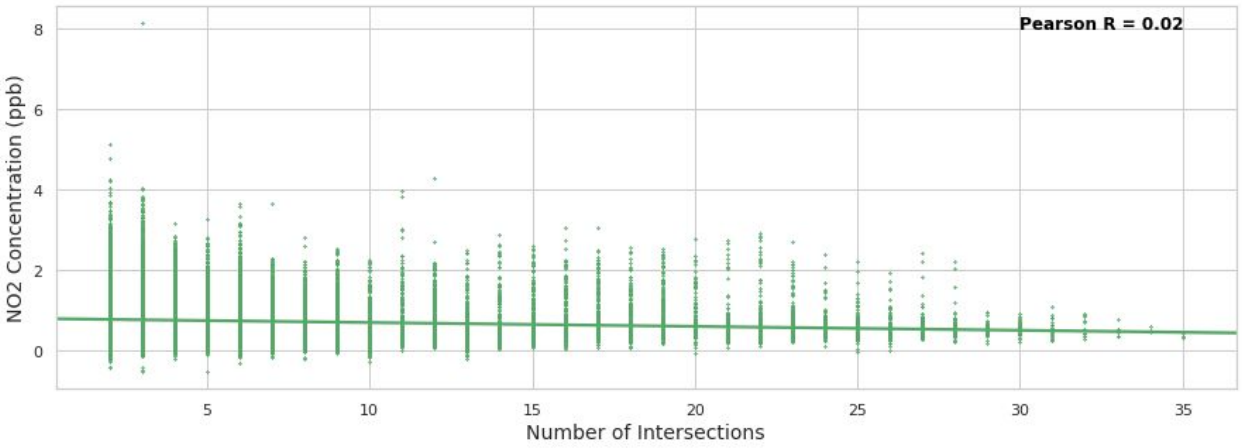

Figure 15: Concentration vs number of traffic intersections - NO2

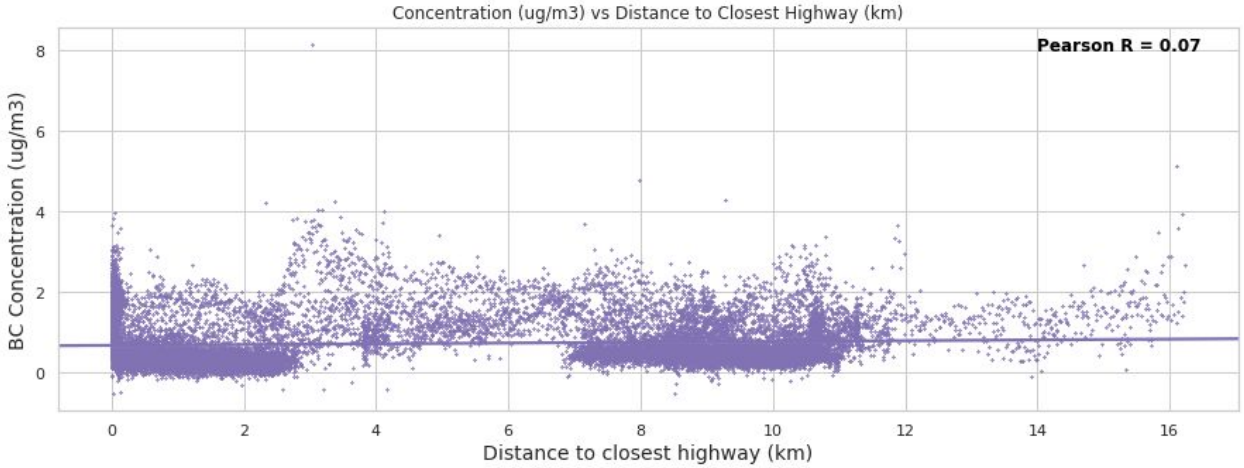

Figure 14: Concentration vs distance to closest highway - BC

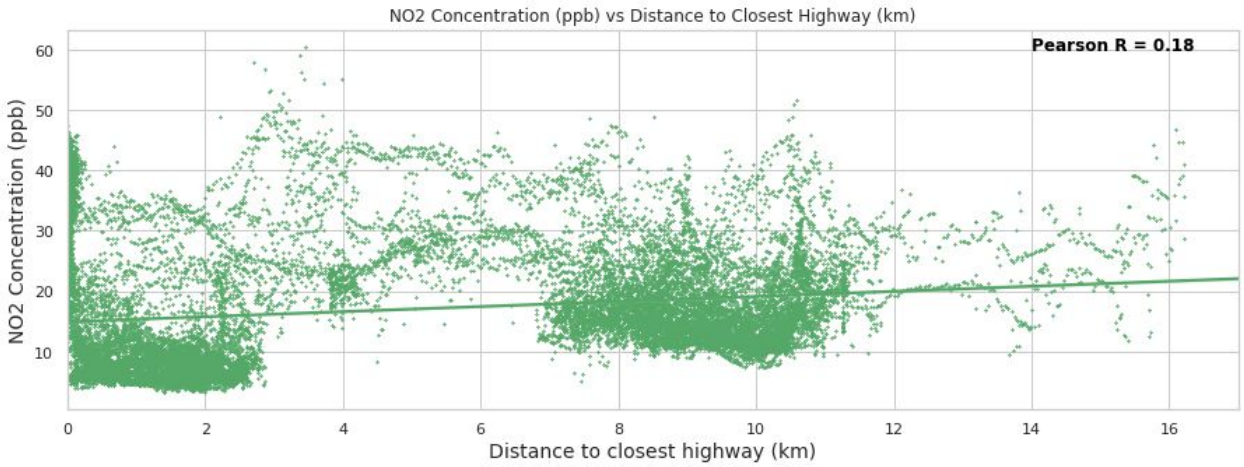

Figure 14: Concentration vs distance to closest highway - NO2

6) Next, we explore correlation between meteorological parameters and concentration . A plot of annual average precipitation and concentration indicates (*Figure 15*) that there is a negative correlation between concentration and precipitation. A plot of average annual radiation and concentration indicates that there is almost no correlation(*Figure 16*). Similarly, there seems to be a weak correlation between minimum or maximum temperature and concentrations (*Figures 17 and 18*)

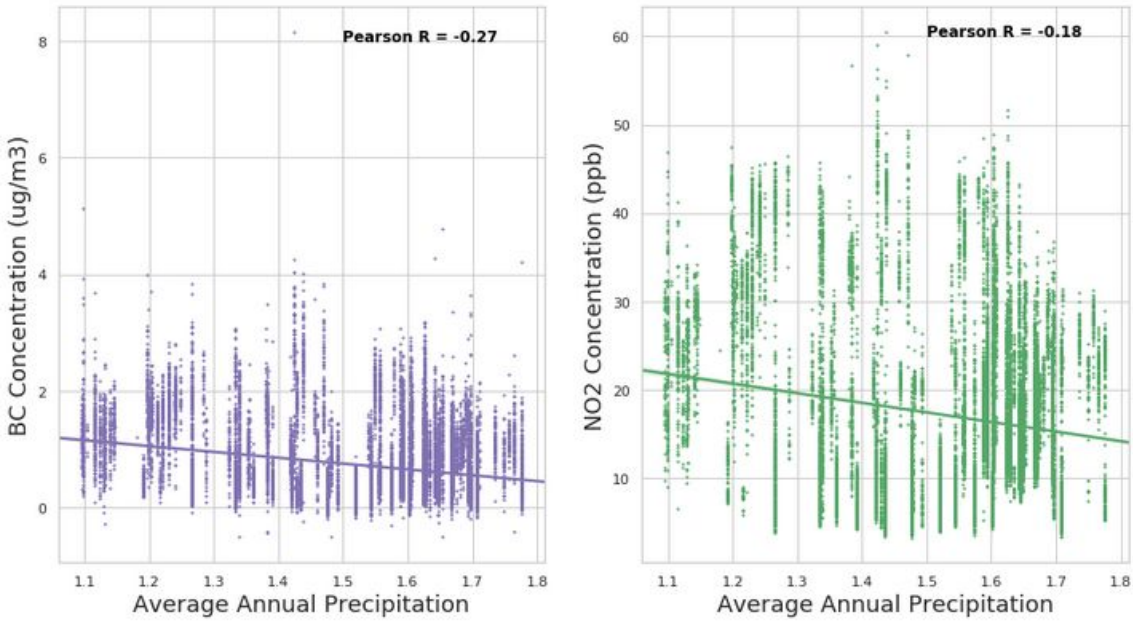

Figure 15: Concentration vs average annual precipitation

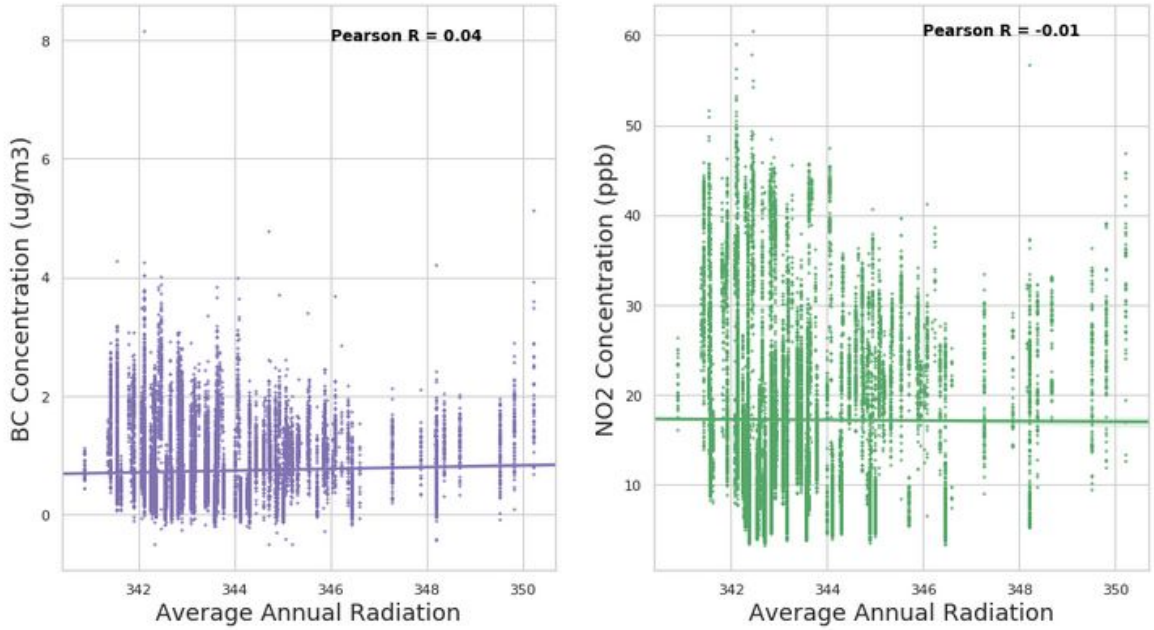

Figure 16: Concentration vs average annual radiation

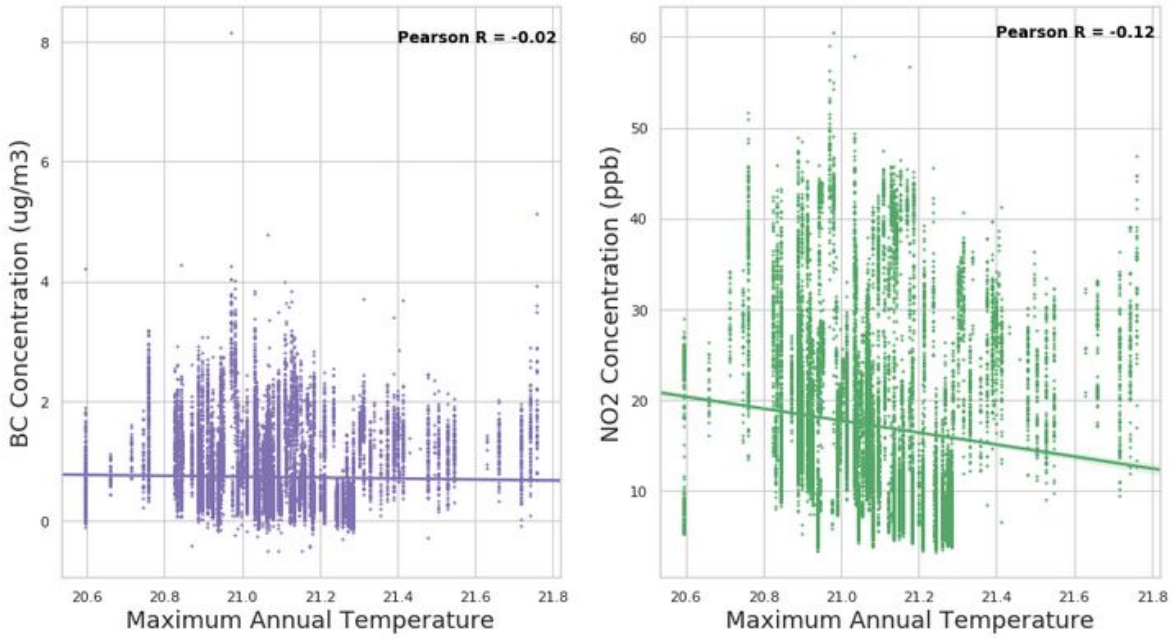

Figure 17: Concentration vs maximum temperature

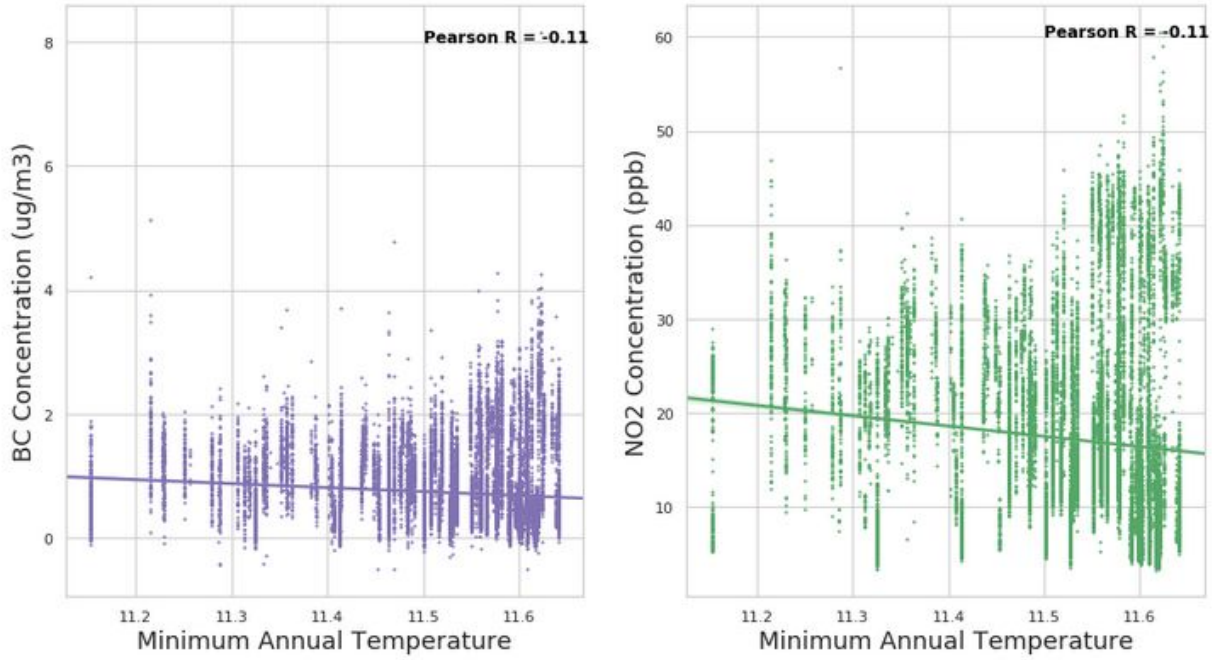

Figure 18: Concentration vs minimum temperature

Based on my above observations, I will explore the following hypothesis:

1) There is a correlation between air pollution concentration and types of emission sources.
2) There is a correlation between air pollution concentration and quantity of emissions from sources (low, medium and high)
3) There is a correlation between air pollution concentration and distance to emission source.
4) There is a correlation between concentration and number of traffic intersections.
5) There is a correlation between concentration and distance to closest highway
6) There is a correlation between concentration and meteorological parameters.