Milestone Report:

Author: Varsha Gopalakrishnan

Problem Statement

In densely packed cities like Oakland or San Francisco, air quality can vary wildly across neighborhoods, due to varying sources of emissions. Current methods to monitor air quality, while certainly very useful, are either too widely spread out (sparsely located monitoring systems set up by local agencies) or too localized (personal air quality monitors that have a small radius of detection). In this project, we plan to answer the following question:

> Can we predict air quality in different Oakland neighborhoods based on local meteorological conditions, local sources of emissions, and previously measured concentrations without having to rely on complex physical modeling?

We will predict air quality in the City of Oakland on a city-block basis where they are not measured using multiple sources of publicly available data such as:

- Oakland Air Pollution Monitoring Data measured by Environmental Defense Fund (EDF)
- Air Emissions data obtained from the National Emissions Inventory,
- Intersection Count from Open Street Maps,
- Distance to closest highway from each monitoring point from Open Street Maps,
- Local meteorological data on a 1kmx1km grid obtained from Oak Ridge National Lab

Monitoring pollutant concentration within cities is crucial for environmental management and public health policies in order to promote sustainable cities. Apart from being a resource to the general public, this project can be useful to:

- To detect if there is an anomaly in air quality on a given block.
- To detect extreme events in air quality due to anomalies.
- Local and state air agencies: Get air quality measurements on a near real-time basis
- Social fitness apps like Strava, Fitbit, Nike Run who can leverage this data to recommend routes for users to run, bike, hike, etc

Datasets

- The air pollution monitoring data is obtained from the EDF website. The data contains latitude, longitude (points where the measurements were taken), concentration of Nitric Oxide (NO), Nitrogen dioxide (NO2) and BC. A plot of the BC and NO2 concentration from this dataset are shown in *Figures 1 and 2*.

- Data on local sources of air emissions such as major industrial facilities were obtained from the National Emissions Inventory (NEI) for Alameda County. Data on Particulate Matter (PM2.5, PM10) and NOX were obtained from the NEI.
- Number of traffic intersections within 1,000 ft of each monitoring location obtained from Open Street Maps using the Overpass API
- Distance to the closest highway from each monitoring location obtained from Open Street Maps
- Local meteorological data is obtained from Oak Ridge National Lab's Daily Surface Weather and Climatological Summaries here. The dataset contains gridded estimates of daily weather data including total daily precipitation, minimum and maximum surface temperature, humidity, shortwave radiation, snow water equivalent and day length for the whole of North America.

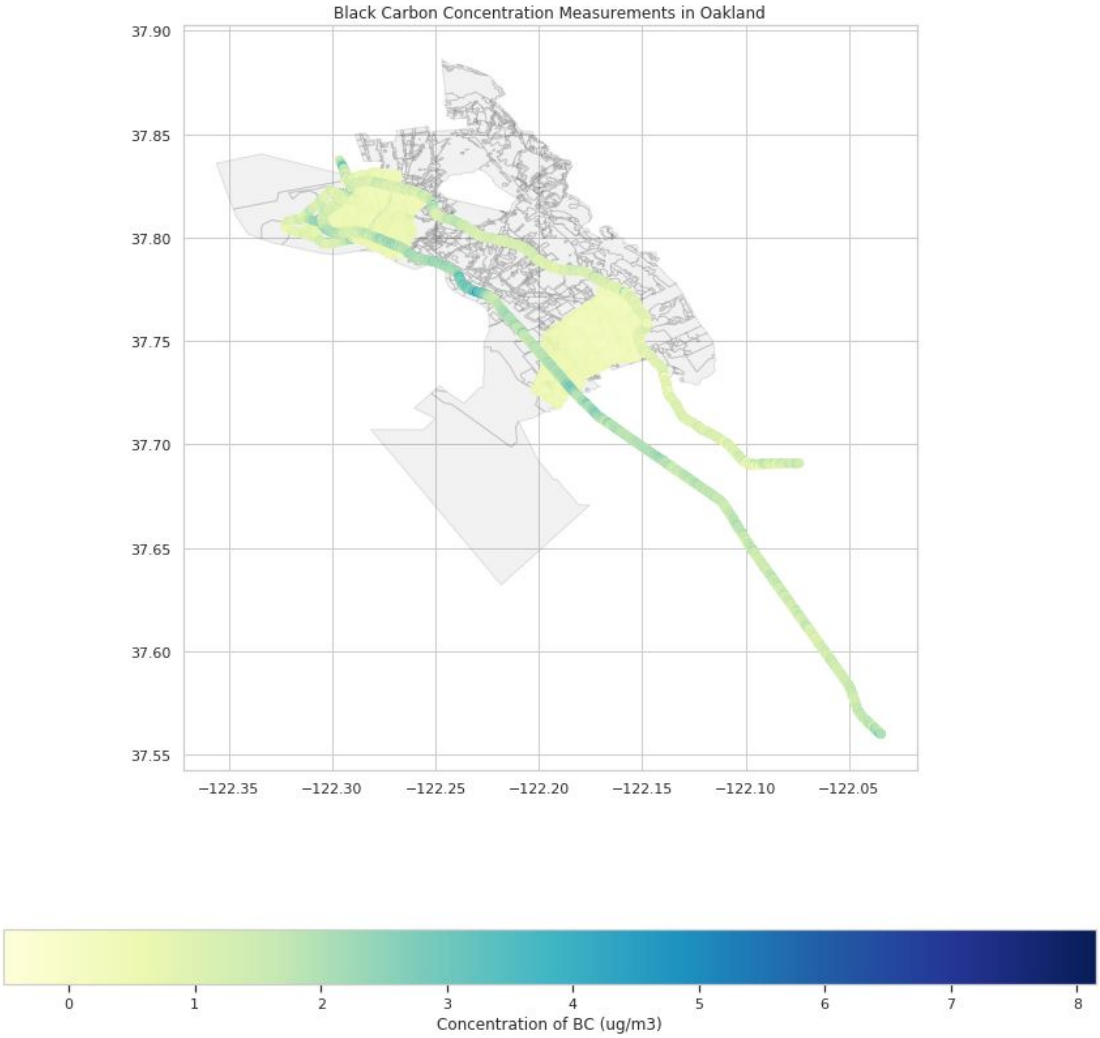

Figure 1: Concentration of Black Carbon Measured at Different locations

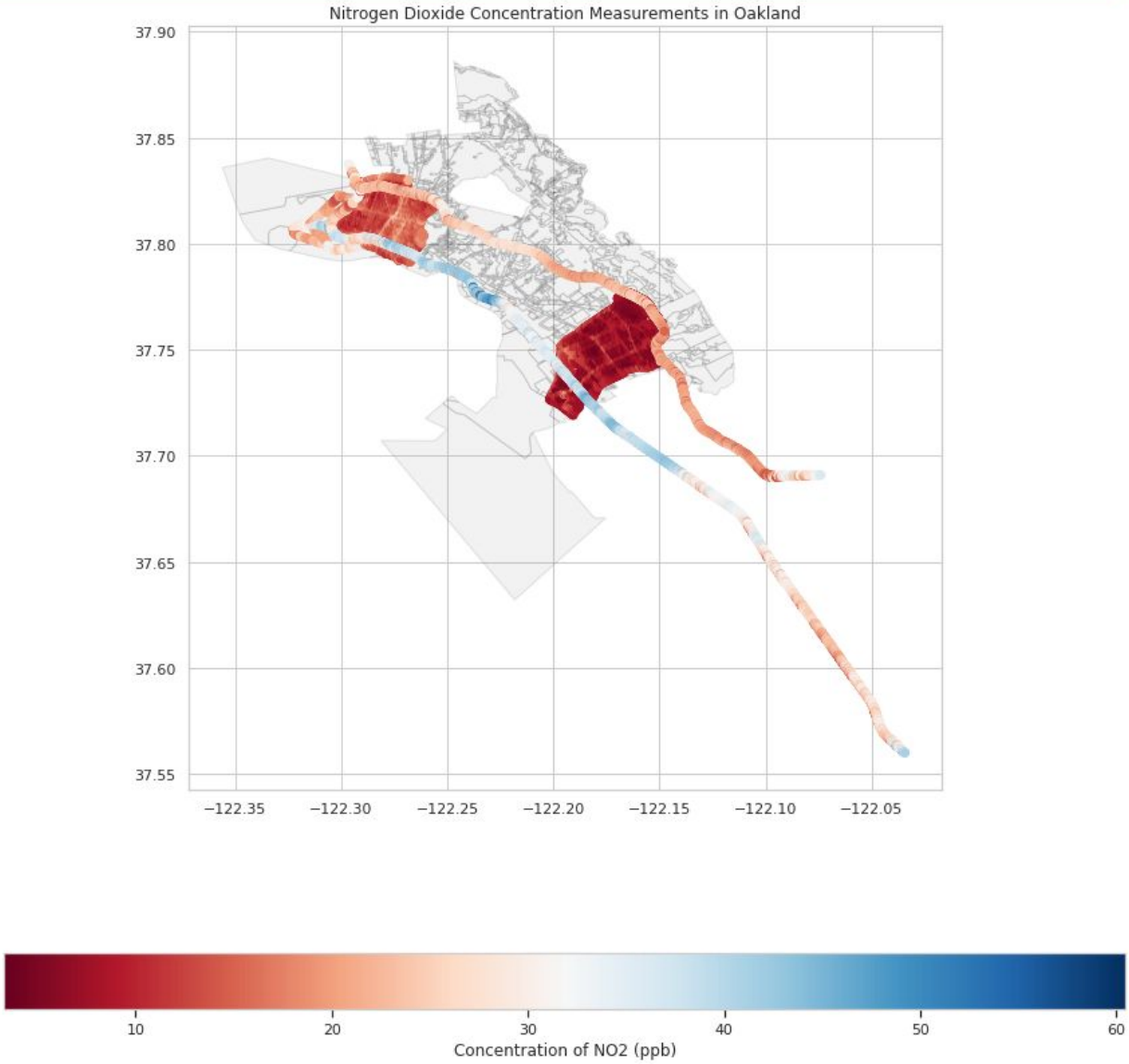

Figure 2: Concentration of Nitrogen Dioxide Measured at Different locations

Data Wrangling and Cleaning

Understanding the distribution of data

First histograms of the target variables, BC and NO2 concentrations were plotted to understand the distribution of sources. *Figure 3* shows the histogram for BC and NO2. The plots show that

the data don't follow a normal distribution, and hence we need to transform the dataset using Box-cox (log transformation) to convert the data to follow a more normal distribution

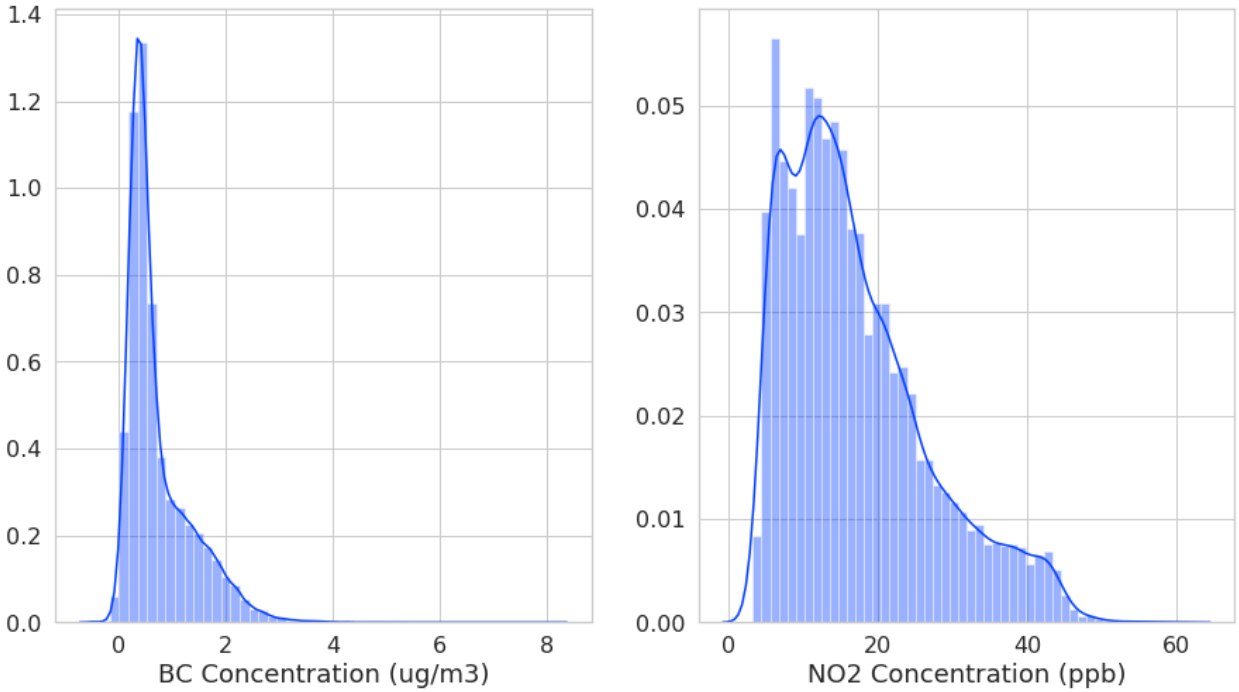

Figure 3: Histogram of BC and NO2 datasets

The transformed data is shown below in *Figure 4*.

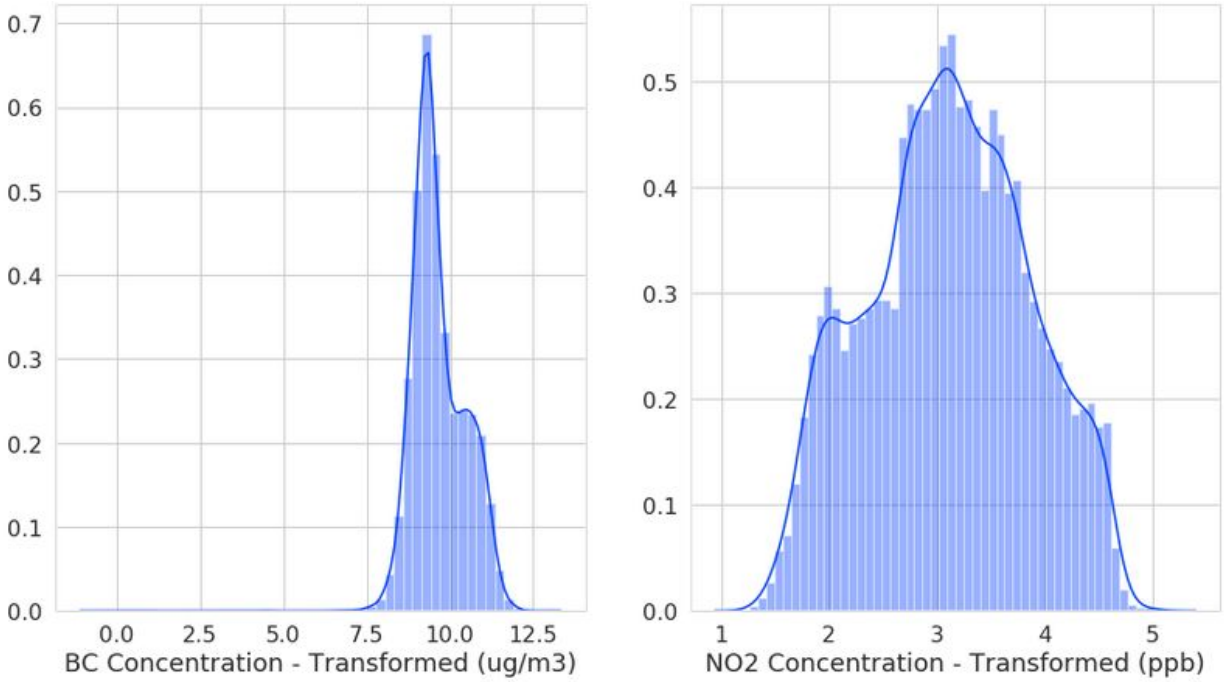

Figure 4: Histogram of BC and NO2 datasets after Box-cox transformation

Cleaning the National Emissions Inventory databases

The PM2.5 and NO2 datasets from national emissions inventory contained facilities that were categorized into several different types, resulting in a large number of independent features but with very few sources in each. Some of these facilities also had very low emissions. In order to minimize multicollinearity, some of the smaller sources (with emissions < 5 tons per year) were combined into larger source groups. Combining the sources into larger groups resulted in a boxplot as shown in *Figures 5 and 6*. 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. Only facilities categorized as 'high' were chosen in the analysis.

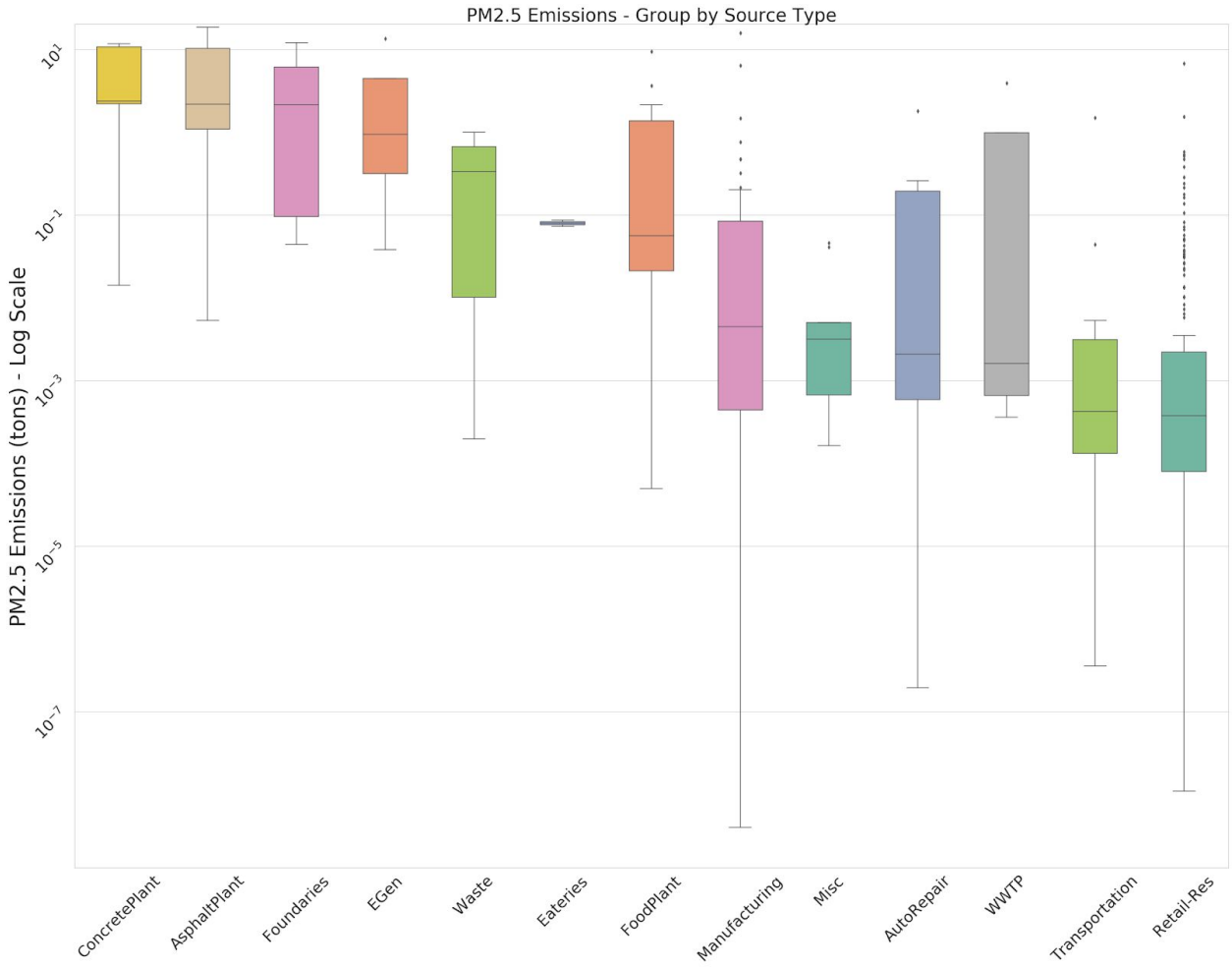

Figure 5: Boxplot of PM2.5 emissions by different source types - grouped - Log Scale

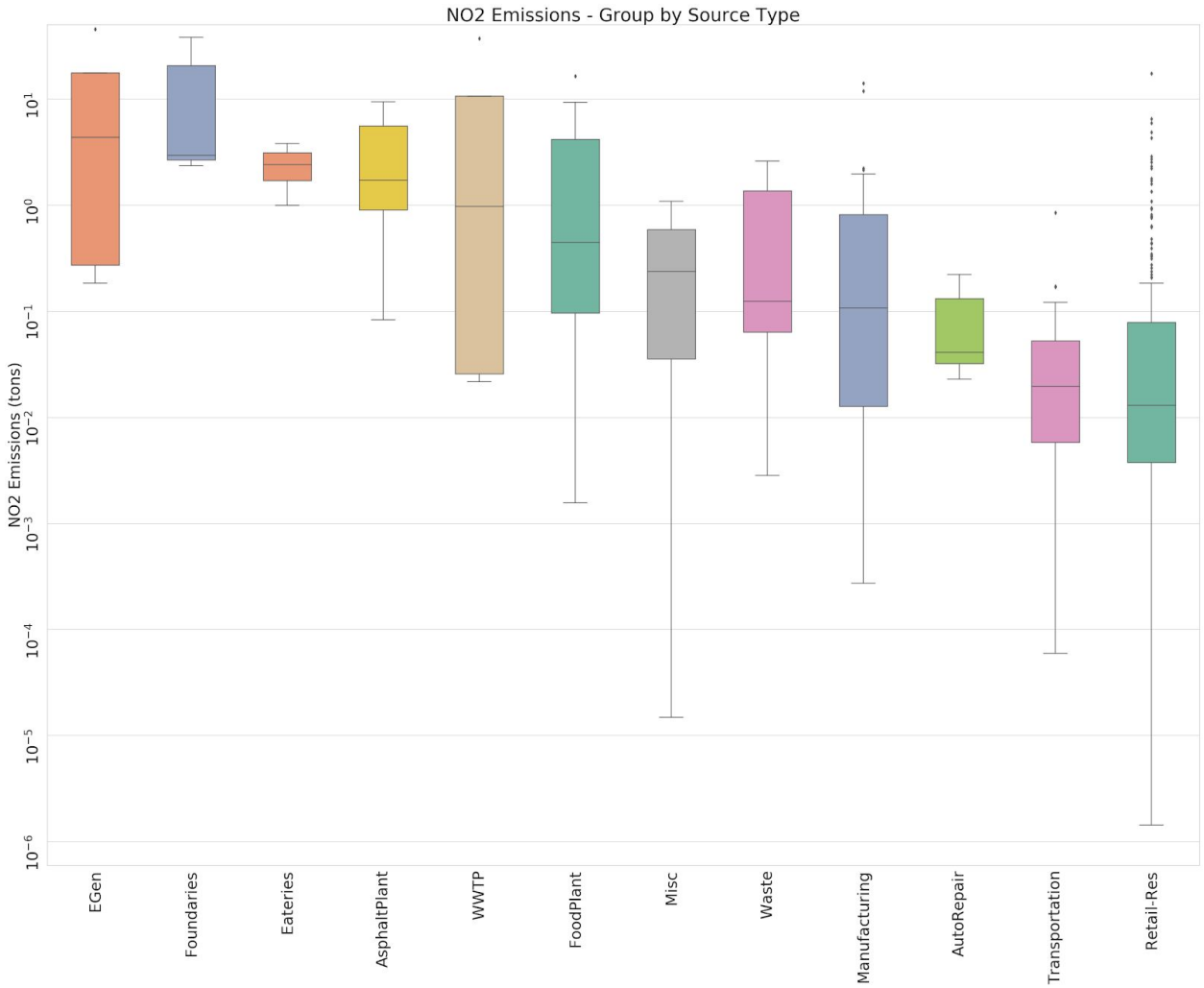

Figure 6: Boxplot of NO2 emissions by different source types - grouped - Log Scale

Exploratory Data Analysis

The hypothesis that I'm exploring here "is air pollution concentration in a given location 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 7* below.

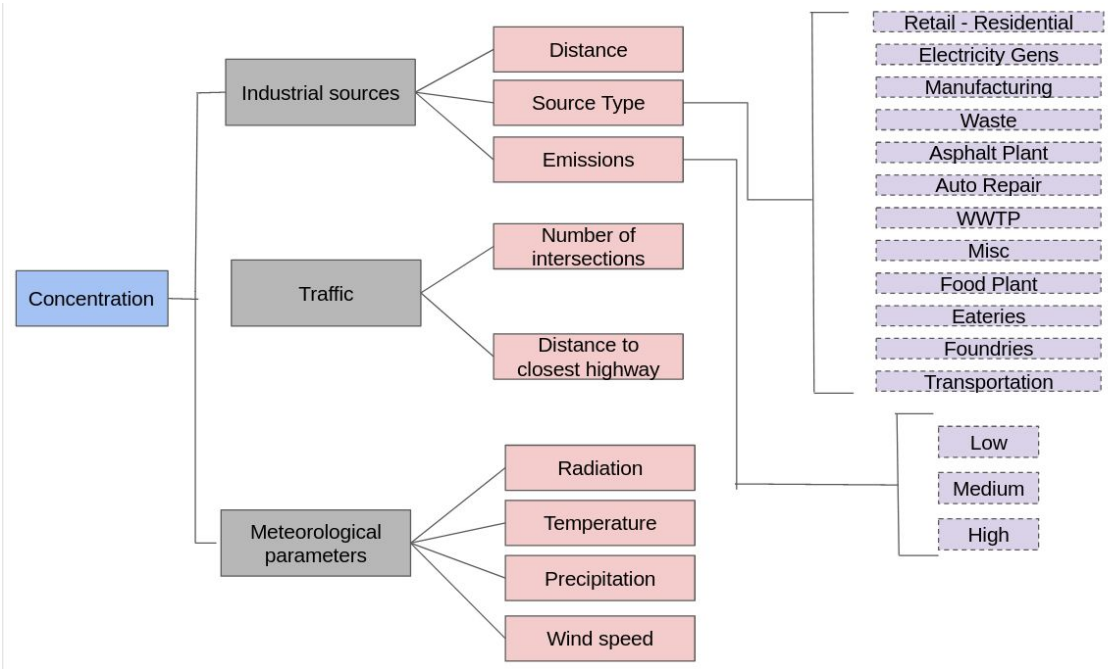

Figure 7: Hypothesis Tree

Some of the questions that were answered as a part of the exploratory data analysis include:

- ***How does air pollution concentration vary with distance to sources of emissions?***
- ***Is concentration directly correlation with distance, or to (distance)^2 or to emissions/distance?***
- ***Is there a correlation? between concentration and number of traffic intersections***
- ***Is there a correlation between concentration and distance to closest highway?***
- ***Is there a correlation between concentration and meteorological parameters?***

The graphs below (*Figure 8 and Figure 9*) show some of the correlations that were explored for the BC and NO2 datasets:

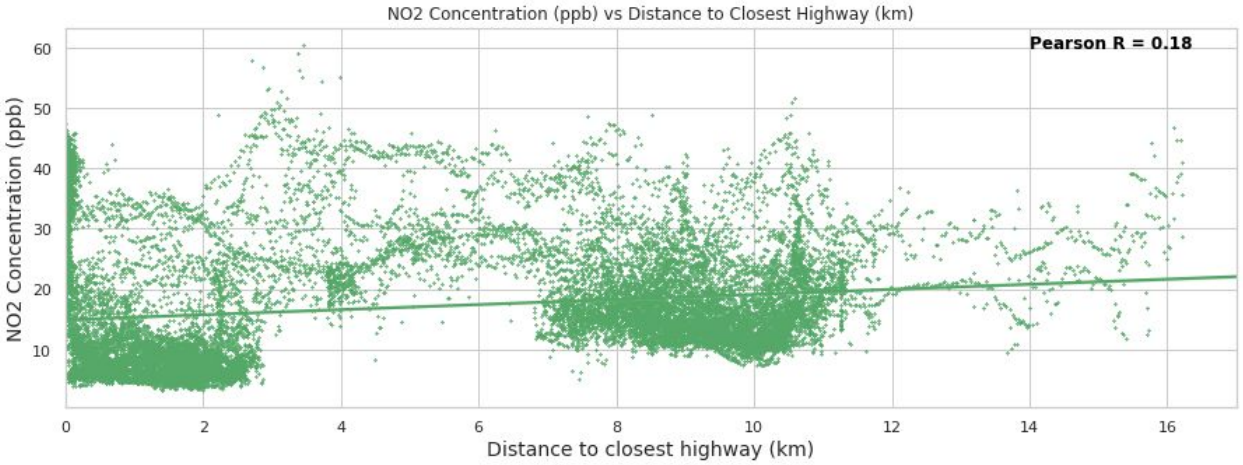

Figure 8: Concentration vs distance to closest highway - NO2

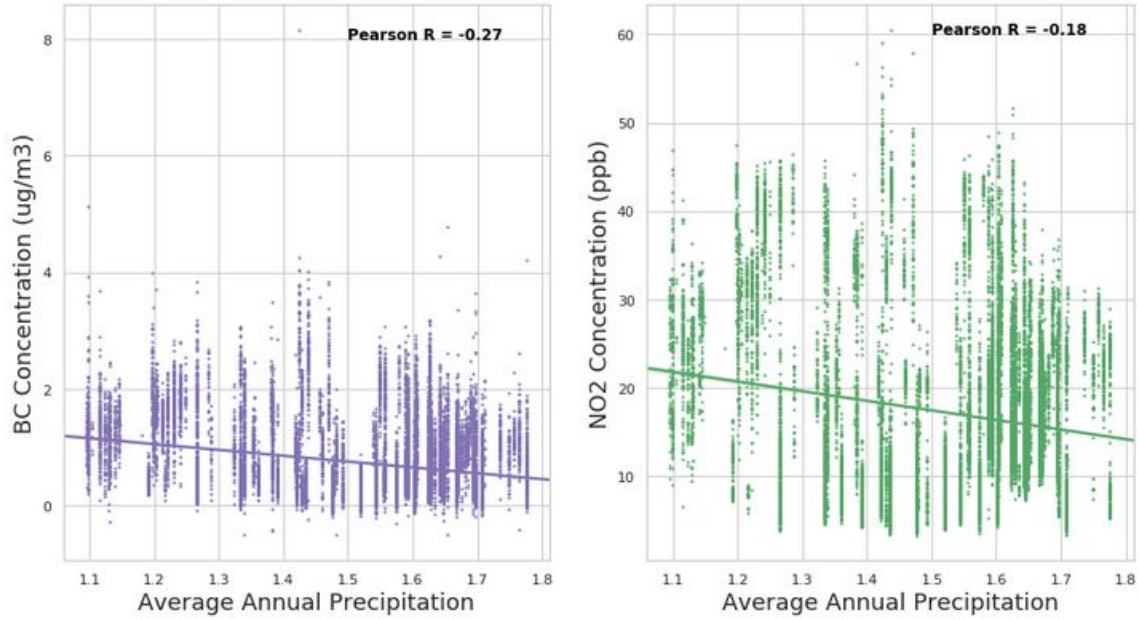

Figure 9: Concentration vs average annual precipitation

The above graphs show that there is only a small correlation between concentration and distance to the closest highway, as well as some of the meteorological parameters.

Applying Statistical Inference

Before building the machine learning model, some more statistical data analysis was performed in order to understand the dataset better.

Resolving Multicollinearity

A correlation matrix of the features in the BC and NO2 datasets revealed that several features were very highly correlated with each other, with a correlation coefficient as high as 0.99. This indicated that multicollinearity is a major issue in this dataset, given how spatially distributed the points in the dataset are. It is important to eliminate features that are multicollinear because multicollinearity can undermine the statistical significance of an independent variable. While multicollinearity does not necessarily affect a model's predictive accuracy, it affects the variance associated with the prediction, as well as, reduces the quality of interpretation of independent variables i.e. effect of the data on the model isn't trustworthy.

Several approaches were explored to remove features that were multicollinear, and to understand the importance of each feature on the model.

Simple Linear Regression on un-correlated features:

Based on the exploratory data analysis, several features were identified that were correlated with the Black Carbon (BC) and Nitrogen dioxide (NO2) concentrations. Strong positive or negative correlations were observed for some features, and weak correlations for this rest.

Next the effect of each feature on BC and NO2 concentrations using statistical inference techniques were studied. The first test was to plot a correlation matrix, and identify features that were highly correlated with each other. The features that were highly collinear were dropped, and a simple Ordinary Least Squares regression model was fit with the rest of the features.

The correlation matrices for the non-collinear features for BC and NO2 are shown below (*Figure 8*), along with the OLS results. While this method resulted in a model fit with R2 of 0.682 for BC and 0.818 for NO2, the Variance Inflation Scores (VIFs) for these features were still very high.

Thus, simply dropping features that have a high correlation coefficient in the correlation matrix is not enough to resolve multicollinearity.

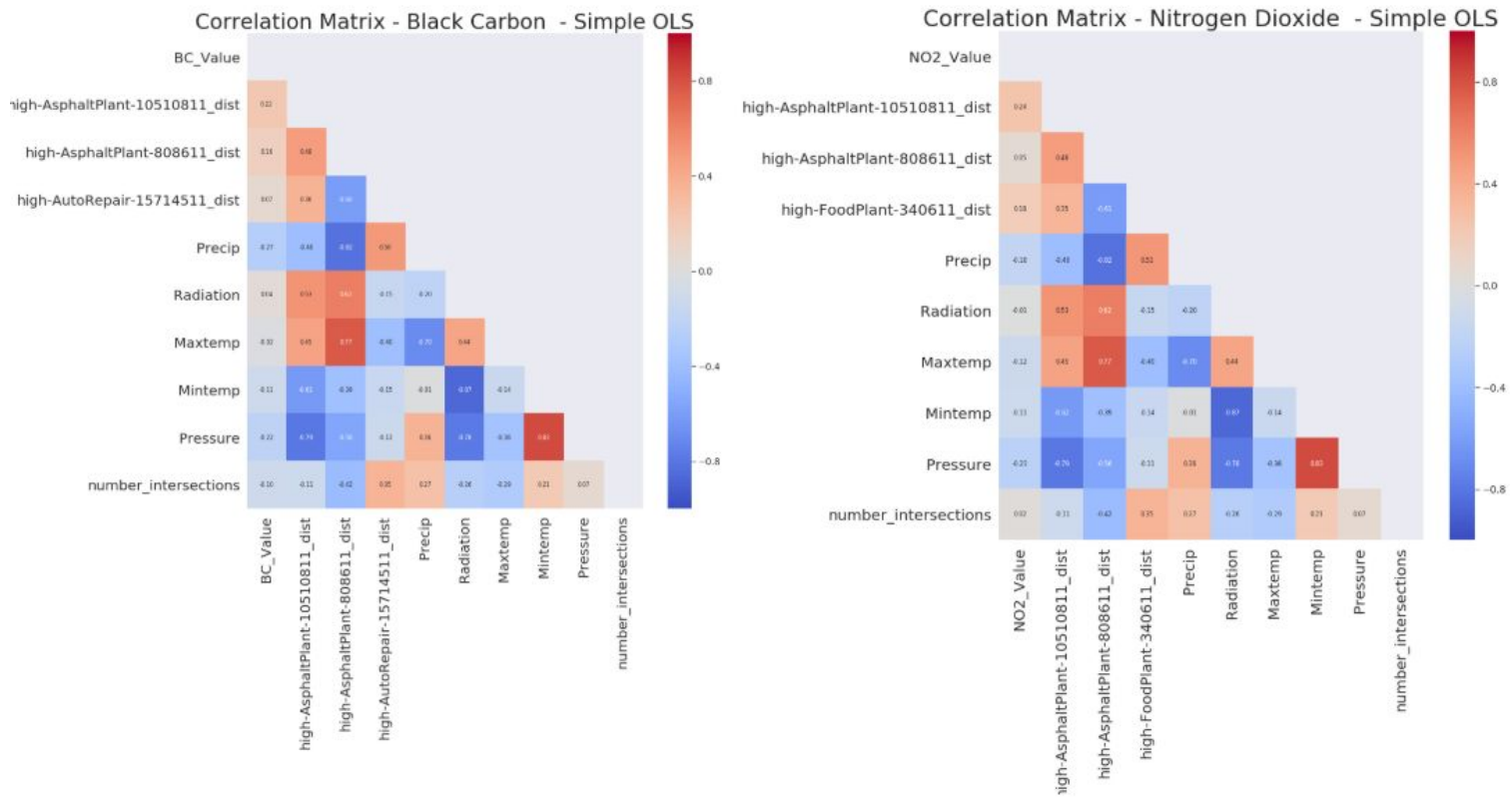

Figure 8: Correlation Matrix for BC and NO2 Datasets

Step forward approach with VIF and R2 estimation: Next, a step forward approach for feature selection with VIF score estimation was performed. In this approach, we first determine the feature that results in the maximum R2 value for a simple OLS fit between the target and predictor variable. Next, we keep adding features sequentially, each time calculating the R2 value and estimating if the newly estimated R2 is better than the previous R2. If the R2 is higher, then we calculate the VIF score with addition of each feature. If the VIF score of any of the features increases above threshold of 10, then we drop the newly added feature. This process is repeated until we are left with a set of features that results in a high R2 value, and low VIF score. The resulting features that were selected had a VIF score below 10, indicating that the features are not collinear.

For the BC dataset, only **'Radiation'** as a feature was selected and the OLS fit resulted in an R2 value of 0.593 for the training data and 0.584 for the test data. Similarly, for the NO2 dataset, only 'Radiation' as a feature was selected and the OLS fit resulted in an R2 value of 0.759 for the training data and 0.757 for the test data.

Lasso Regularization with Gridsearch combined with step forward approach with VIF and R2 estimation

Next the same step forward feature selection approach was used again, but this time performed a lasso regularization with gridsearch and cross validation, to identify only the most important predictor variables that have an effect on the target variable.

The Lasso regularization resulted in the selection of a total of 49 features out of the 68 features for the BC dataset and 42 features out of 50 features for the NO2 dataset.

Next, the step forward approach with VIF and R2 estimation was applied to the features selected by the Lasso regularization method, and only '**minimum temperature**' as a feature was selected. The resulting model R2 for the BC dataset was 0.592 for the training set, and 0.583 for the test set. Similarly, for NO2, only '**minimum temperature**' was selected as a feature and R2 value for the training set was 0.759, and test set was 0.756.

Feature Selection Using Random Forest and Cross Validation

Next, feature selection was performed using the Random Forest approach. Random forests are commonly used for feature selection because the tree-based models naturally rank features by how well they improve the purity of the node. Here, Random forest with a cross validation approach was applied for feature selection. The BC and NO2 datasets were split into test/train data, a 4-fold cross validation approach was applied on the training dataset to select the features. For each fold, the feature importance was calculated and then an average of all feature scores was calculated to determine feature importance in order to rank features. The following graphs (*Figure 9 and Figure 10*) show the feature importance for the BC and NO2 datasets.

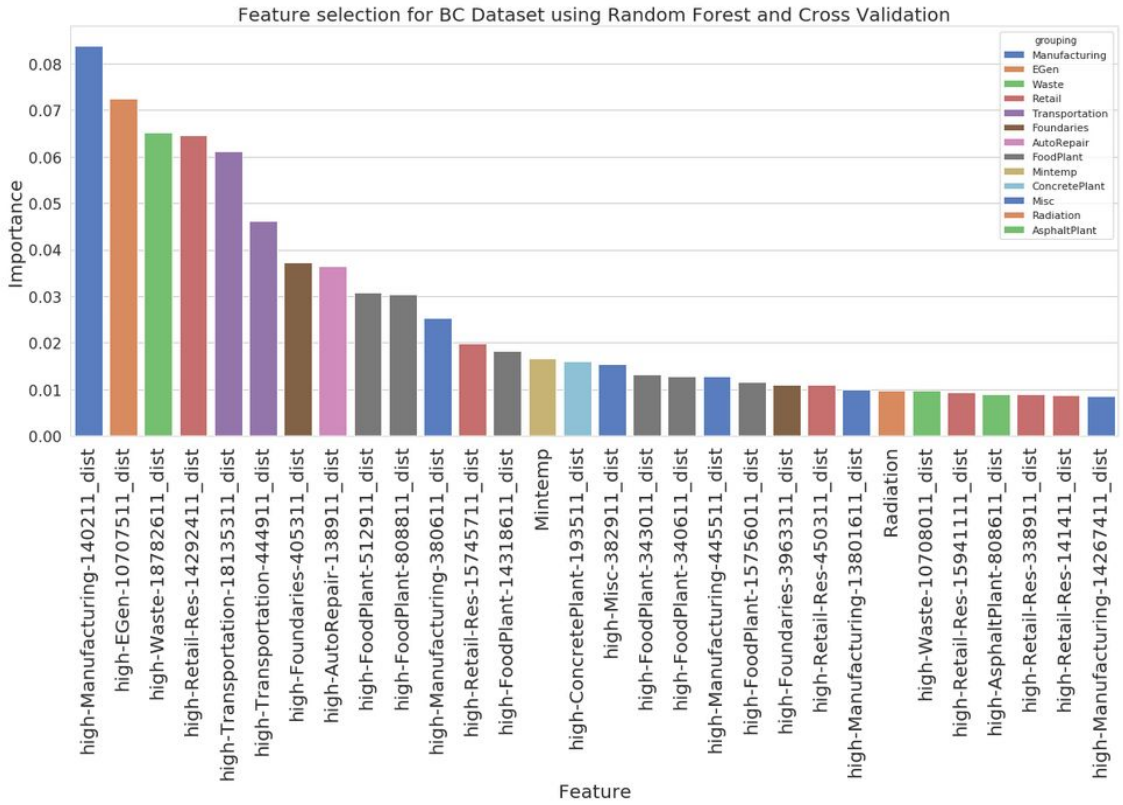

Figure 9: Feature importance for BC dataset

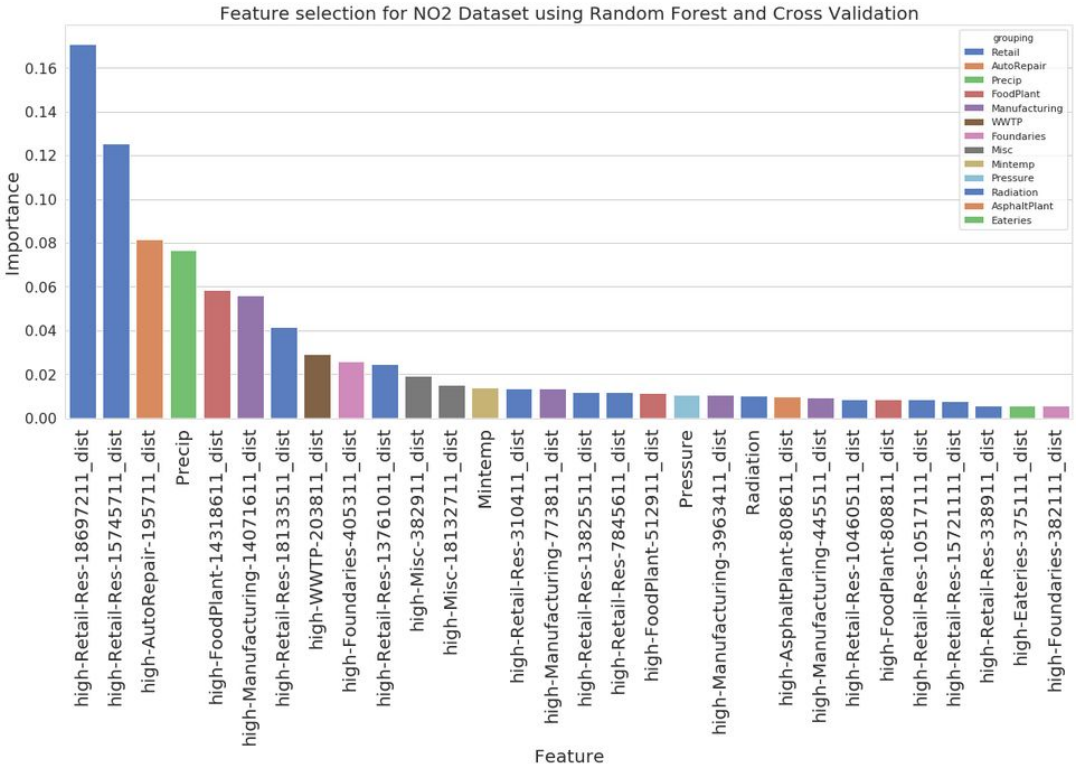

Figure 10: Feature importance for NO2 dataset

Figures 11 and 12 contain maps showing the location of the important features, in this case location of facilities that contribute most to air quality in the region. The size of the dots indicate the feature importance, with larger dots indicating higher importance. The color of the points show the sector to which each facility belongs.

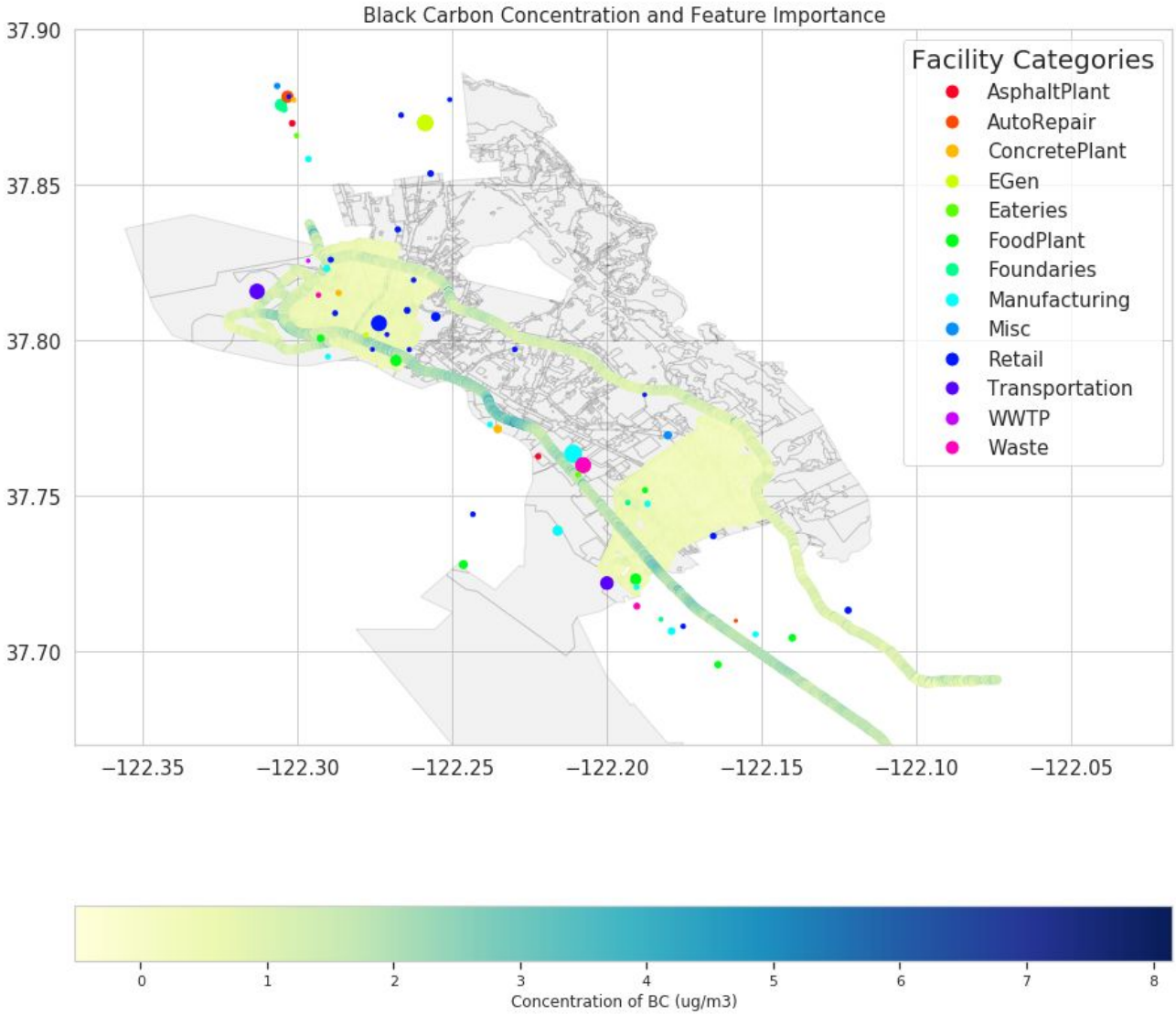

Figure 11: Location of Important Features for BC Dataset

For the BC dataset, the top five features that contribute to concentration in the region includes a soft drink manufacturing company (SVC manufacturing Inc, division of Pepsi Co), electricity generation units located in the Berkeley campus, a waste treatment facility, an office building located on Clay Street in Oakland, and the Port of Oakland.

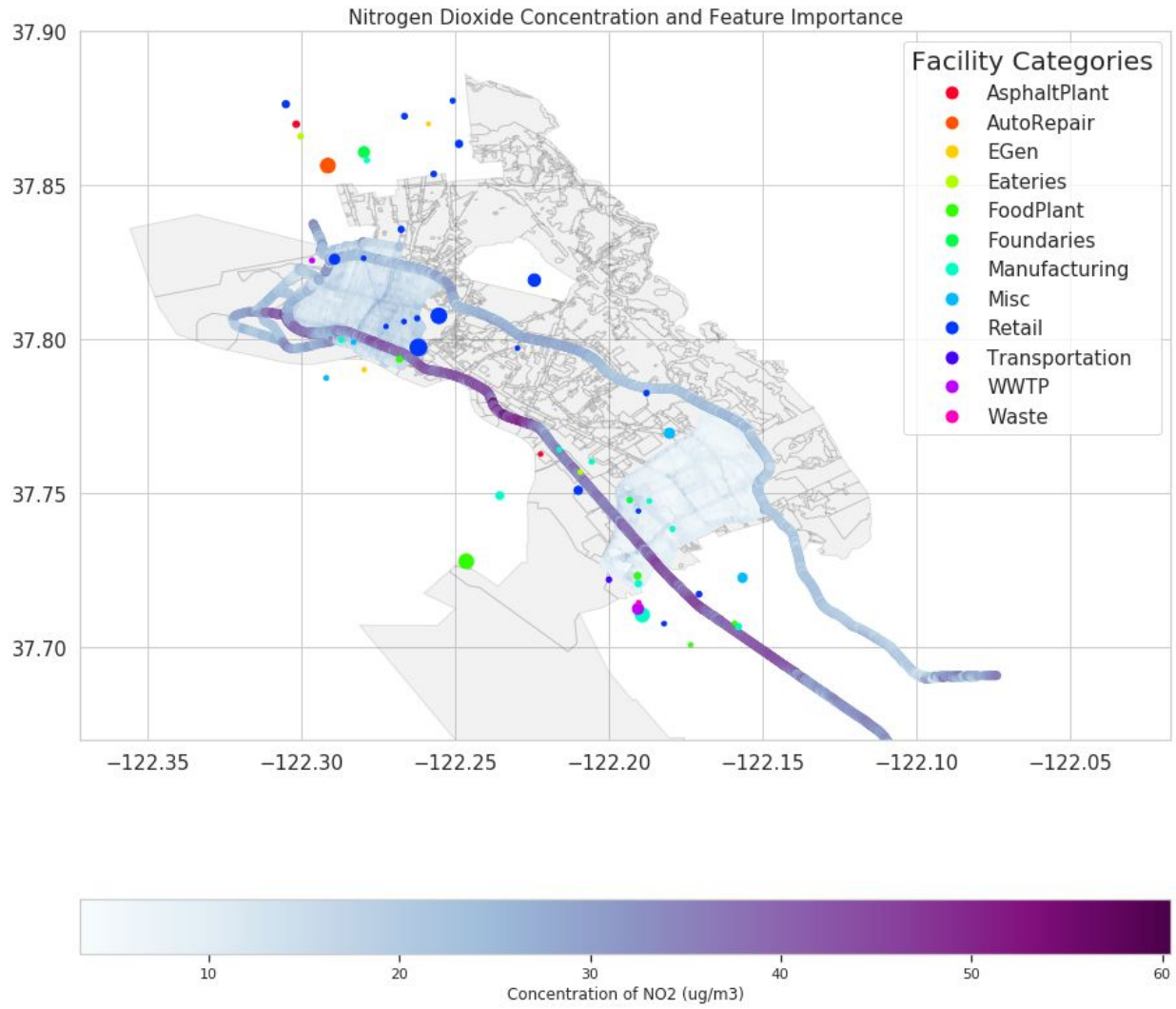

Figure 12: Location of Important Features for NO2 Dataset

For the NO2 dataset, the top five features that contribute to concentration in the region includes a wholesale facility, a residential complex with generators, a commercial printing facility (Consolidated Printers Inc.), annual precipitation, and a coffee roasting plant (Peets Coffee roasters).

While it's hard to view this on the map, the BC concentration at some points close to the facilities that are important are high indicating that these sources contribute to the concentration in that area. Similarly, for NO2, concentrations close to the 'Retail-Res' facilities are high indicating that these facilities contribute to the concentration in the area.

The features listed above are only the top five features in terms of feature importance. However, there are more features shown in the graphs in Figure 9 and 10 that actually contribute to

concentration of BC/NO2 near the areas where they are high. It is also important to note that most of the facilities that have a high feature importance are closely clustered around the I-880 highway, which has some of the highest concentration of BC and NO2.

Even though feature importance gives an insight into the features that contribute to concentration in the region, one of the challenges of this particular dataset has been the lack of data on emissions from traffic in the region. In other words, the lack of inclusion of emissions from traffic as a feature in the dataset has resulted in some challenges. Since traffic contributes highest to black carbon concentrations (which comes from partial combustion of fuel) and nitrogen dioxide concentrations, this analysis still does not give us a full picture on the major sources that contribute to air pollution in an area. Besides, the concentration in each location is so hyper-localized that it's hard to identify one single source that contributes to concentration in the entire region.

Summary of findings

The dataset for Capstone 1 is air pollution monitoring data of black carbon and nitrogen dioxide measured in Oakland and San Leandro between Jun 2015 - May 2016 obtained from the Environmental Defense Fund. The dataset contains annual average concentration of BC and NO2 at different locations and a map of the points showed that high concentrations occur pretty high close to highways.

The objective of my work is to build a machine learning model to predict air pollution concentration on a black-by-block basis based on input features such as major sources of emissions in the area including industrial sources, number of traffic intersections, proximity of each monitoring location to the closest highway and local meteorological data.

After a lot of exploration, data cleaning and statistical data analysis, I found that the dataset contains input features that are correlated with each other, resulting in severe multicollinearity issues. This was evident from the high VIF scores for all the features. I tried several different approaches to identify the features that had the highest effect on the target variable including Lasso regularization, step forward approach with VIF and R2 estimation, and random forest with cross validation.

Based on my findings, the features that had the largest effect on the target variable for the BC dataset included PM2.5 emissions from manufacturing facilities, an electricity generating unit, a waste disposal facility and the Port of Oakland. For the NO2 dataset, the facilities that resulted in a large contribution to NO2 concentration in the area includes a wholesale facility, a residential complex with generators, a commercial printing facility and precipitation.

While the Random forest approach gave us some insight into the features that are important, this still does not give us the major sources of emissions (or features) that contribute to air quality in a local area. The features we identified here only give us an insight into the main sources that contribute to concentration in the entire region, and not on a hyper-local level. Emissions from vehicles and trucks traveling on highways are actually one of the highest contributors to BC and NO2 emissions on a local-level, and lack of emissions from traffic is one of the limitations of this work. Based on my analysis so far, I observed that it is an aspirational goal to try and predict the major sources of pollution at a local level, even though the model was able to identify some features that are important.