Capstone 1 - Data Wrangling

Author: Varsha Gopalakrishnan

The following datasets were used in this project:

1) 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.
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.
3) Number of traffic intersections within 1,000 ft of each monitoring location obtained from Open Street Maps using the Overpass API
4) Distance to the closest highway from each monitoring location obtained from Open Street Maps
5) 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.

EDF monitoring dataset:

The first dataset, which is the air pollution monitoring dataset obtained from the EDF website was very clean, with missing values filled in. As a result, there was no data cleaning or wrangling done here. A sample of the dataset is shown below:

	Longitude	Latitude	NO Value	NO2 Value	BC Value	TimePeriod
0	-122.322594	37.806781	23.390071	17.539762	0.818032	Jun2015-May2016
1	-122.322310	37.806150	19.700000	19.956750	0.551475	Jun2015-May2016
2	-122.322301	37.806420	23.611111	23.967768	0.593712	Jun2015-May2016
3	-122.322299	37.805880	15.714285	18.435184	0.489898	Jun2015-May2016
4	-122.322267	37.806689	27.108695	25.797037	0.739341	Jun2015-May2016

National Emissions Inventory:

There were several data wrangling steps that had to be performed on the second dataset from the National Emissions Inventory. Firstly,

- Data was filtered for Alameda county in California
- Several unnecessary columns were dropped from the dataset, and filtered only for pollutants of interest
- Several columns had to be renamed to a shorter form, or spaces removed

- Several facilities had unknown names or categories. These had to be filled in by plotting them on a map and observing the type of facility, or looking up the facility based on the address.
- Some of these facilities also had very low emissions (as shown in *Figures 1 and 2*). In order to minimize the number of source categories, 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 *Figure 3 and 4*. Further, a categorical variable to each source as 'low','medium' and 'high' was assigned 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.

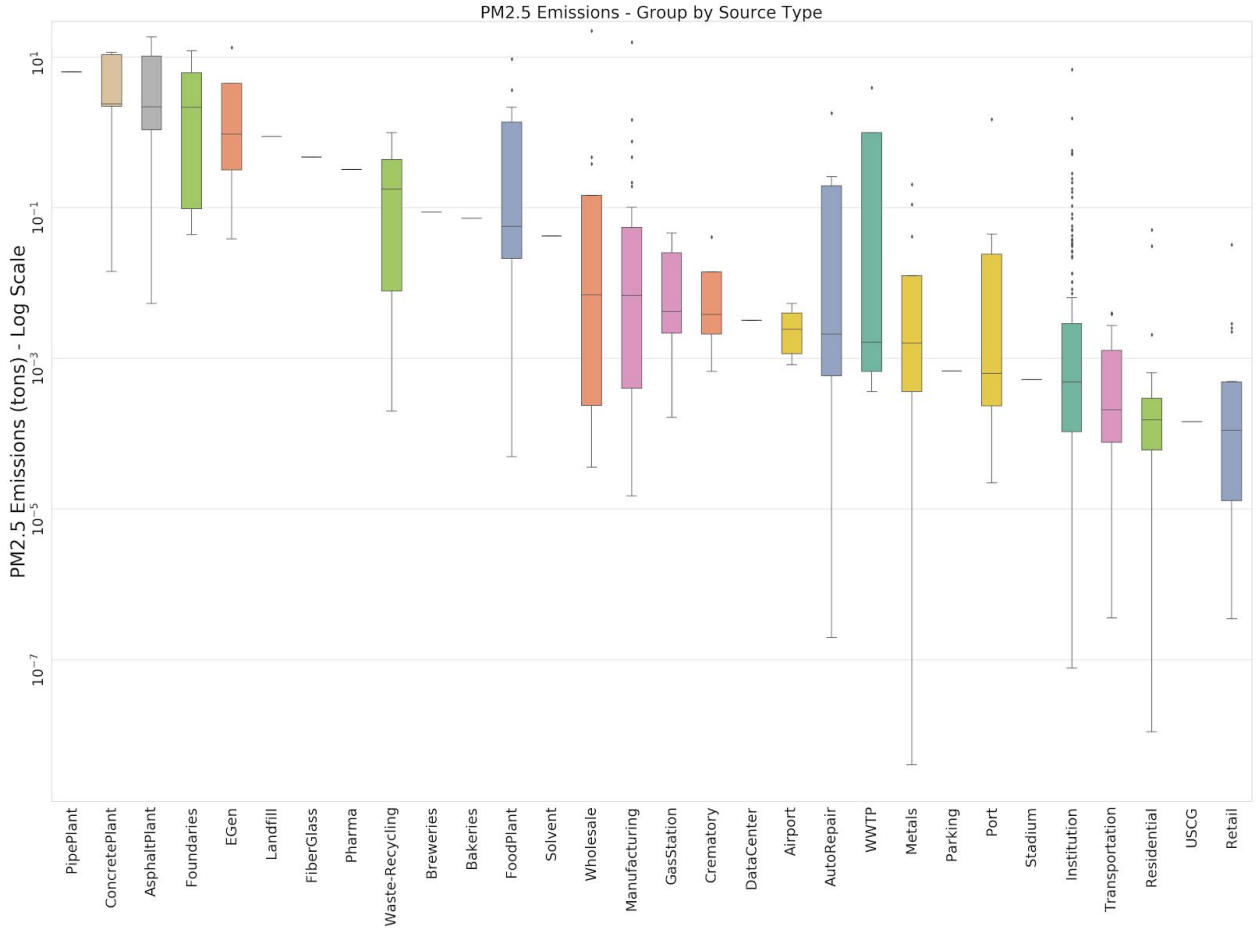

Figure 1: Boxplot of PM2.5 emissions by different source types - ungrouped - Log Scale

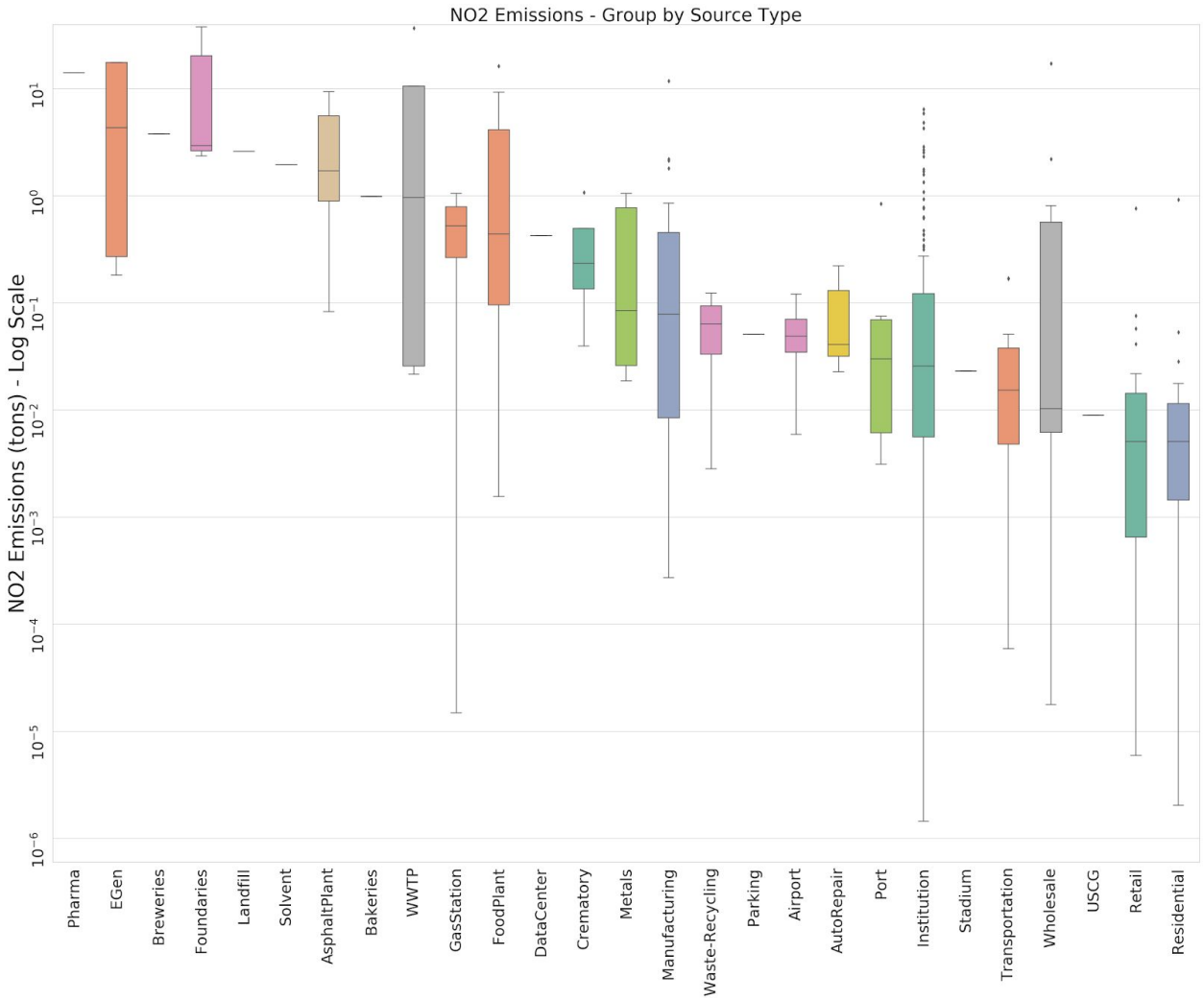

Figure 2: Boxplot of NO2 emissions by different source types - ungrouped - Log Scale

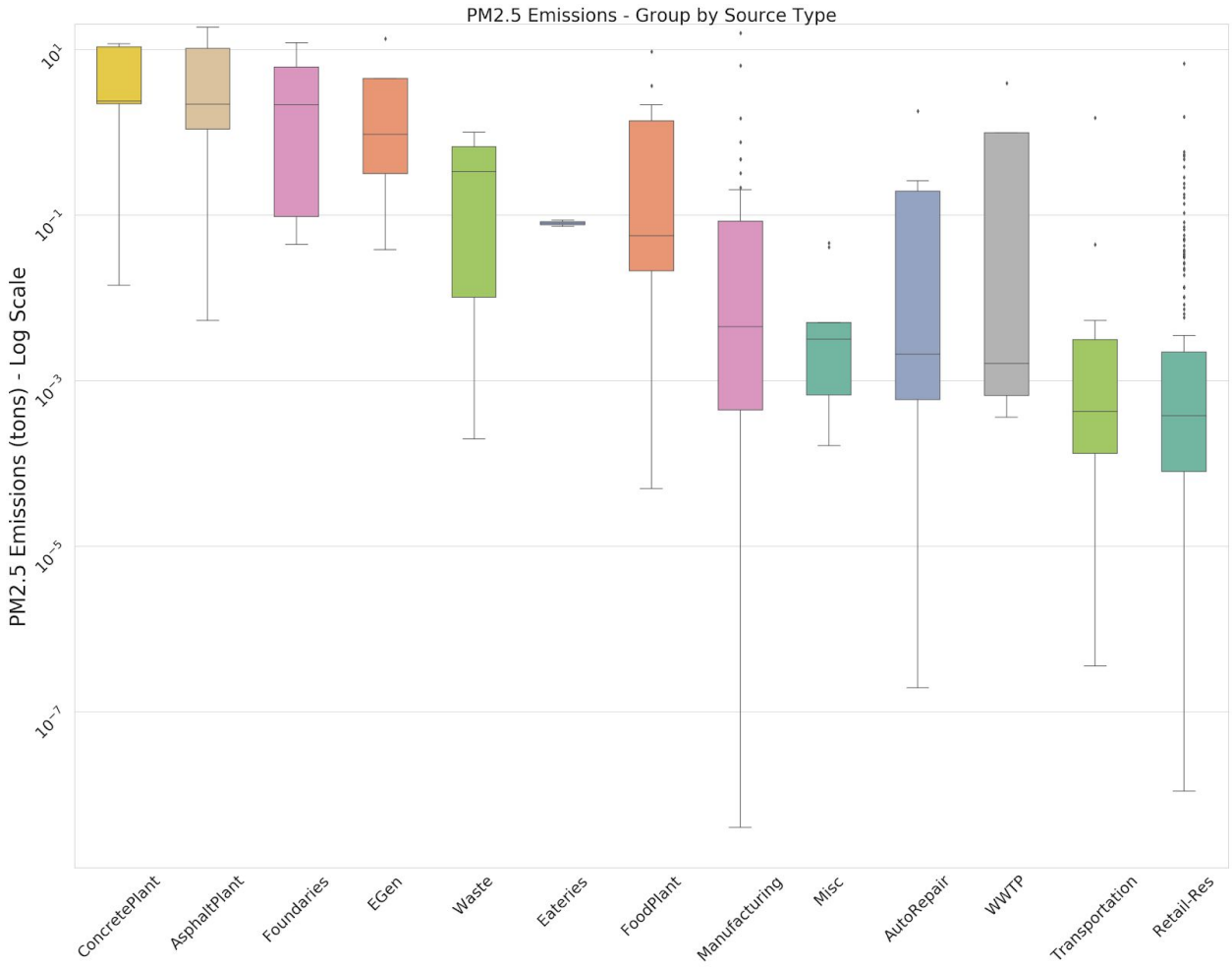

Figure 3: Boxplot of PM2.5 emissions by different source types - grouped - Log Scale

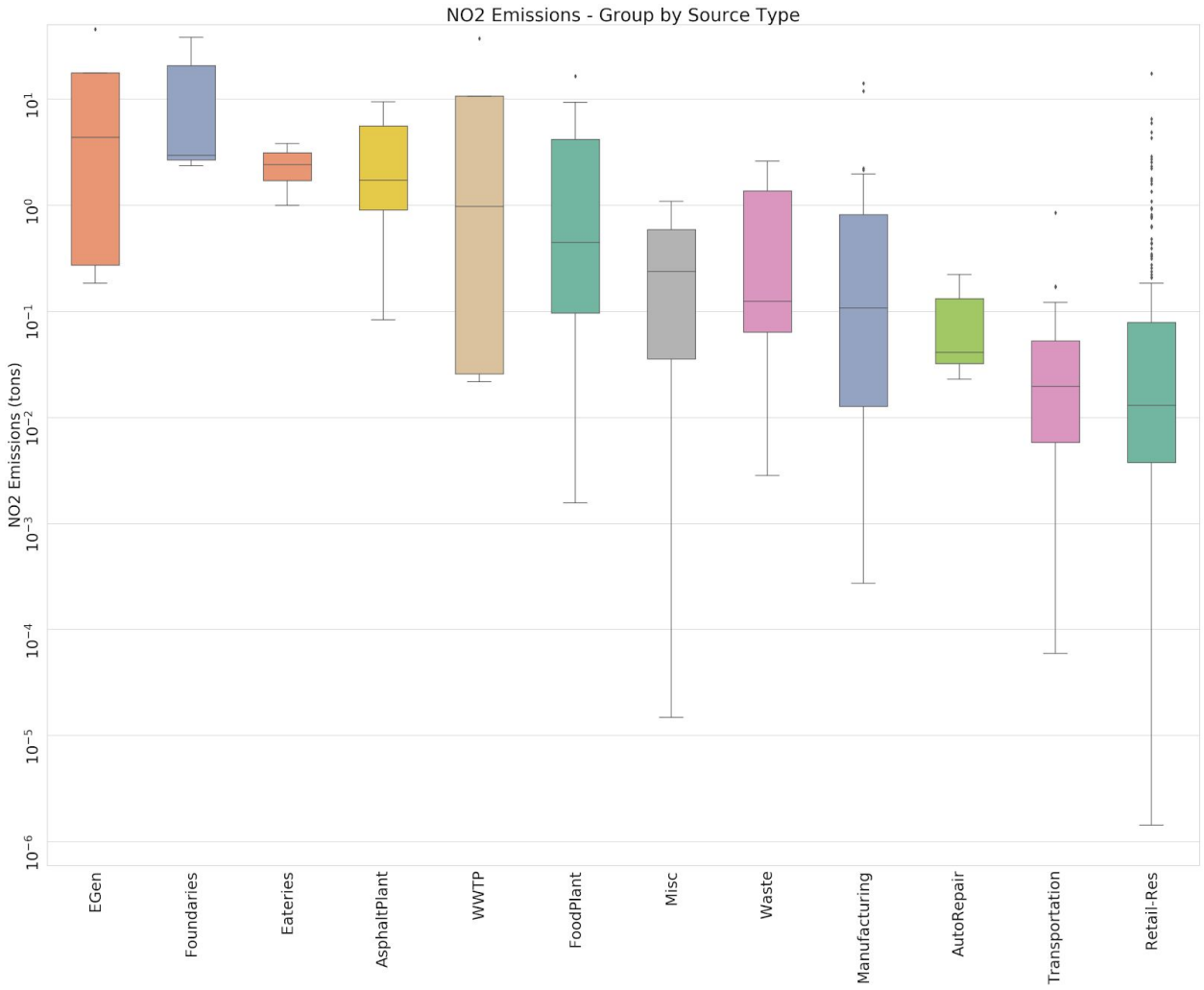

Figure 4: Boxplot of NO2 emissions by different source types - grouped - Log Scale

Traffic Data:

Traffic data in terms of intersection locations was obtained from Open Street Maps using the Overpass API. The API query takes in a bounding box (latitude and longitude) as inputs and returns the latitude and longitude of all traffic intersections within the box.

- The distance between the traffic intersections and each point in the monitoring dataset were measured, and the number of intersections with 1,000 ft from each monitoring point was measured using functions.

Similar to the traffic intersection data, data on distance to closest highway was obtained using the Overpass API within the bounding box.

- The distance from each monitoring point to the closest highway was then measured

There were not many data cleaning steps for the traffic data except to change some column names

Meteorological Data:
Meteorological data was obtained from Oakridge National Lab's Daymet dataset using the daymetpy package. The API call to Daymet call takes in arguments like latitude, longitude, start year, end year and output type as_dataframe. There were several missing rows of data, since data had to be obtained on a daily basis and then annualized. However, a multiprocessing threadpool option was used to parallelize the API calls, and the days with missing data were stored in a separate array. Another API call was made just for the missing days. As a result, no further data processing was needed.