**Import data**

Features: 1 year of hourly NO2, NO, O3, RH, and T from the Breathe Providence sensor co-located at the East Providence supersite

* Method: Alphasense and BME sensors
* Accessed via BEACO2N API

Target: 1 year of hourly reference NO2 from the East Providence supersite

* Method: FRM (Chemiluminescence Thermo Electron)
* Accessed via AQS API

Data were merged on the “date” column. 8760 total rows.

A “time since deployment” column was added, counting the hours since deployment.

Rows were dropped where the target variable was missing (6.7%)

8184 total rows.

**Exploratory data analysis**

EDA was performed to examine the distribution of each variable and examine the relationships between measurements as influenced by temperature, humidity, and ambient NO2 concentrations

Histograms

* Target variable has a non-normal distribution (log normal or exponential distribution)
* Features had more normal distributions

Heat map

* Reference NO2 has a relationship to temperature due to seasonal changes in emissions, meteorology, and chemistry
* NO2 is higher on average and more variable in the wintertime

Scatter plots

* At low temperatures, LCS and reference have a positive linear relationship
  + LCS performance degrades at high temps regardless of the ambient NO2 concentration
  + At medium-high temps, LCS is positively biased
  + At very high temps, LCS is negatively biased
* LCS NO and NO2 have a negative linear relationship at high temperatures/low NO2 concentrations, but at low temperatures have a slight positive correlation
  + Reference NO2 and NO have a strong positive correlation
  + Two different regimes for the LCS

Feature correlations and MI

* LCS NO2 and LCS NO have a strong negative correlation, even though reference NO2 and NO are highly positively correlated
* LCS NO2 and NO have a very high MI

**Splitting**

I created a function that divides the data into quintiles and uses train\_test\_split to stratify on those bins

* Data were divided into train, validation, and test sets
* Skewness was maintained across sets

**Preprocessing**

Missing data

* 1.7% of features were missing. Data were imputed using mean imputation as a placeholder until we learn more advanced methods

Feature engineering

* Time since deployment/sensor age column added before
* PolyomialFeatures function was used to generate interaction term features

Scaling

* StandardScalar was used