**Outlier elimination and missing value imputation:**

* Post the initial data prep stage (read: merging air quality & weather data and combining historical and API datasets), PM2.5, PM10, O3, humidity and wind\_speed variables had quite a few outliers. *For eg. PM10 variable for the Beijing dataset had the max value as 3000 which pulled the mean of the variable to 93.78 when it should have been 80.79 (without outliers).* The outliers were replaced by NA which then were imputed by the mean of the variable.
* Dealing with missing values was a critical part of the data prep as there were a large number of missing values and hence high impact on training the model. *Eg: Beijing had 105,684 missing values for PM10 variable before outlier elimination. After outlier elimination, there were 116,850 missing value observations.*
  + At the initial stages of submission, we tried to eliminate all rows with missing values (complete cases function in R) but that was giving poor model performance and wasn’t an option for the ARIMA model.
  + We approached missing value imputation with mean in a hierarchical approach.
    1. Replace outliers with NA
    2. Find the mean of each of the variables that contain missing values (pressure, humidity, wind\_speed, PM2.5, PM10 and O3) for the combination group [stationID, hour, month, weekend, holiday] (newly created variables explained in feature engineering section). The combination was chosen based on experimentation with various combinations and an assumption that these factors would affect the dependent variables the most.
    3. Impute this mean value for the missing values with the same combination group. As the missing value imputation was done at a group level, there were a few missing values still present in each of the variables
    4. In the second iteration, find the mean of each of the variables that still contain missing values (pressure, humidity, wind\_speed, PM2.5, PM10 and O3) for the combination group [stationID, hour, weekend, holiday]
    5. Impute the mean value for the missing values with the same combination group.
    6. Post the above imputations, a couple of stations in London didn’t have any O3 values and hence there were still a few missing values for O3 variable. So these were imputed by mean of [hour, month, weekend, holiday] combination (all London stations)