We decided to try an ARIMA model with this data since it is a time-series forecasting based model. Unlike a regression model where you get to choose what features to use in the model, and there’s a whole process to figure out which features would be best to use, ARIMA is simply based on time and the feature you are trying to predict. For example, to predict PM2.5, all we needed to put in the model was the timestamp and the history of PM2.5 up until the latest hour. We chose the past month of API data so that the model would be able to hopefully deduce some seasonality and trends. Another task to make sure to accomplish before running the model was to take care of the rows of data that had NAs. As you can imagine, NAs can be especially detrimental in time-series modeling due to the gaps of data causing the predictions to be skewed. Therefore, it was imperative to take care of the NAs. To accomplish this, we used the last hour at that station to fill in any future hours of data. We were assuming that a station would not be unavailable for long gaps of time. We were hoping to be able to figure out how to take the average of the last hour recorded and the next hour recorded, and smooth the NAs between, but it proved to be more complex than we thought. We settled with just carrying over the past hour recorded over the future NAs using na.locf in the zoo package.

Once we took care of NAs, it was time to decompose our data. The goal of this step was to be determine if there were seasonal components and/or trend components. Overall, most stations in the past month only had seasonal components based on time, and weekend/weekday. Therefore, each station seemed to have a seasonal component, and were relatively stationary. We used auto.arima to build the model with ‘seasonal = TRUE’. We had to loop each station through the set of code because each was unique. Then we had to duplicate that loop 5 times (once for each pollutant: PM2.5, PM10, and O3 for Beijing, and PM2.5, PM10 for London). Writing the R code and getting it set up in the correct format for submission was difficult since it was all new to us, but we eventually figured it out, and can chalk it up to a good learning experience. The results that we got back after waiting a few days was decent and performed just slightly better than the regression model used at that time. We would have liked to test other settings and parameters with ARIMA, and possibly ensemble it together with our regression model since that is known to be a worthwhile technique.