**Linear regression model:**

Post converting the time variable into various component features, for the 48 air quality stations in Beijing and London and each of air quality variables – PM2.5, PM10 and O3, we built a linear regression model (3 models each for Beijing and London) using caret package in R.

* **Train/Test data:** The Beijing/London dataset was partitioned into training and testing data set with training dataset containing 80% of observations/rows and testing data set containing 20% of observations/rows
  + Beijing dataset: Approx. 320,000 rows for training set and 80,000 for testing set
  + London dataset: Approx. 120,000 rows for training set and 32,000 for testing set
  + Used sample() function for random sampling
* **Features:** Each of the datasets had the following features apart from the air quality dependent variables – station name, lat and long data, corresponding nearest grid station, utc time at which the observation was recorded, hour, month, date, temperature, pressure, humidity, wind direction, wind speed in kph, whether weekend/not, whether the day was a public holiday or not.
* **Methods for improving model accuracy:** In order to better evaluate the model performance on unseen data (other than partition data as training and testing dataset), we used repeated 10-fold cross validation (repeat = 5 times).
* **Feature selection:** Recursive feature elimination, rfe(), using caret package was used to identify the variables used in the model. We used wrapper function such as RFE to optimize model performance effectively and automatically as there was high variability in the data available on a daily basis. We also experimented with variable subset size in the rfe package.
* Apart from using repeated k**-**fold cross validation and recursive feature elimination to automatically enhance the model, we were on the lookout to validate that the rmse doesn’t differ much for training set and the testing set to ensure that the model is not over fit.
* PM2.5 model was built using time and weather variables; For PM10, additionally predicted PM2.5 variable was used and for O3, predicted PM2.5 and PM10 variables were used.
* **Observations:** 
  + The model averaged 0.56 SMAPE over 27 submissions with minimum of 0.39 and maximum of 0.67
  + Before doing missing value imputation and outlier elimination, for Beijing the PM2.5 and PM10 model had rmse = ~65 whereas post the treatments the rmse reduced to ~36.
  + Recursive feature selection seemed to be working well even though it was computationally intensive as it reduced the rmse by atleast 5 for the Beijing PM2.5, PM10 and O3 models compared to a model without feature selection.
  + In case of Weather data for the future 2 days, we used caiyunapp’s data. There were some missing values in this data as it was real time. In order to impute for the missing values, we used the previous day’s data for the stationId, hour to match the predictor variables as closely possible.
* **Sample code & output :** 