Team Name: DS420-Factoria

* Kelsey Fertakis
* Jason Jakubiak
* Michelle Liu
* Vanethi Kathirvel

Ranking on the leaderboard: Haven’t submitted a legitimate model yet…so tied for last place

What you have done?

1. Data Discovery
   1. Perused the data on the KDD Website to see what we had available to us.
2. Data Clean-up
   1. Cleaned up the data we decided to use for our initial model
3. Joined Weather Data to the Air Quality Data
   1. Had to figure out how to easily join the data together and get the columns that we wanted. We have decided to use the historical file to train our model and the API data to test our model
   2. We decided to calculate the closest weather data station to each air quality station. This helped with processing times as it drastically decreased the time we had to wait for everything to run. Instead of millions of rows, we now are in the hundreds of thousands of rows. (We may end up using a weighted average of the top 5 closest weather stations.)
4. Added features to the data to help with our model (capturing time of year and time of day/traffic patterns)
   1. Weekday True/False
   2. Month
   3. Hour of Day
5. We are currently starting to build models now
   1. We are planning to try out Time-Series and Regression
6. Figured out how to upload to the KDD website with a sample submission

What you plan to do next?

1. We need to set up the automatic upload process
2. We need to figure out which of our models perform the best
   1. Moving forward from there, we may end up ‘ensembling’ the models together to improve our overall predictions
3. We need to find a proper way to cleanup ‘NA’s in the data
4. We haven’t yet explored the observed meteorology data. Need to explore this dataset to see if this strengthens our predictions