Capstone 1 Project Proposal

“A Machine Learning Model to Forecast PM 2.5 Concentration in Beijing” by Kevin Limkrailassiri

The poor air quality in Chinese cities such as Beijing is a well-known concern that has been drawing significant attention from news outlets, citizens posting photos of hazy skies on social media, and the Chinese government. In an effort to monitor the air quality, the US Embassy records hourly PM2.5 measurements to detect the concentration of atmospheric particulate matter with diameter less than 2.5 micrometers, and the Chinese government followed suit by setting up multiple measurement centers and began recording PM2.5 data in 2013.

The PM2.5 measurements reported by these measurement centers provide citizens a valuable gauge of the instantaneous air quality. However, citizens may benefit even more if they were provided with a forecast that can predict the air quality with reasonable accuracy several days into the future. In the same way that a weather forecast helps citizens to arrange their weekly plans based on the predicted weather, a forecast of the air quality can also provide citizens a means of responding to days when the air quality is predicted to be poor. Moreover, this study can provide an understanding of weather trends corresponding to higher PM2.5 level, which can help the government proactively curtail contributions from pollution emitted by factories and public transportation when air quality is predicted to be poor.

The PM2.5 data will be drawn from the UCI Machine Learning Repository. The dataset includes hourly PM2.5 data along with dew point, temperature, humidity, pressure, wind direction and speed, and precipitation data. This data will be supplemented by monthly passenger-kilometer data from the National Bureau of Statistics of China of several modes of transportation: railways, highways, waterways, and civil aviation. This data can be found on Quandl.

In brief, the approach of this study is first to perform data wrangling to assemble all the data into one DataFrame with each row containing an observation and each column containing a parameter. Next, we will perform exploratory data analysis to observe any possible trends and relationships between sets of data. The insights from this step will help us perform supervised machine learning in order to construct a model that can forecast PM2.5 concentration based on patterns in the weather and transportation usage. A training and testing split will be used to train the machine learning model and test its accuracy. All results will be delivered in the form of code, a paper, and slide deck.