CAPSTONE Project 1 Proposal

For my very first Capstone Project, I would like to analyze the Beijing PM2.5 Data Set from University of California Irvine Machine Learning Repository.

The goal of the project is to utilize the Time-Series based data of the measured PM2.5 particulate matter in the metropolitan area of Beijing from the United States Embassy and perform a time-series forecasting using linear regression. The results will shed lights on the pollution trend and variations based on the time and weather inputs that was attributed by the rapid urbanization with aids from fossil fuel and coal burning powerplants.

PM2.5 particulate matters are tiny particles that can lodge into the lungs of human body and stay there for long duration of points that cause a number of health issues. With a metropolitan population of about 37.5 million in Beijing, the pollution causes negative impact to both physiological and psychological health of every citizen within the vicinity. And hopefully with this project, information regarding the timing and severity of pollution outbreak can be approximated in such a way that actions can be taken to minimize the impact the pollution can cause to residents of Beijing.

The data set provided by UCI Machine Learning Repository will be in a CSV file, with columns on time divided into years, months, days and hours, followed by data on PM2.5 particulate matters, Dew Point, Temperature, Pressure, combined wind direction, wind speed, hours of snow and hours of rain.

The approach I use will be first to perform some data cleaning and create a time variable that can be used for linear regression, which I expect to be coming the Year, Month, Day, Hours column into one single time-dependent variable, then I will use that as a basis for time-series analysis using linear regression. One models I will use to perform this time-series analysis will be the Auto-Regressive Integrated Moving Average (ARIMA) model, and using Facebook’s opensource package, Prophet, to compare the results.

My deliverable of this project will be my python codes written in Jupyter Notebook and uploaded into my github repository.