Geoffrey Leigh

Capstone: Big Data and analytics

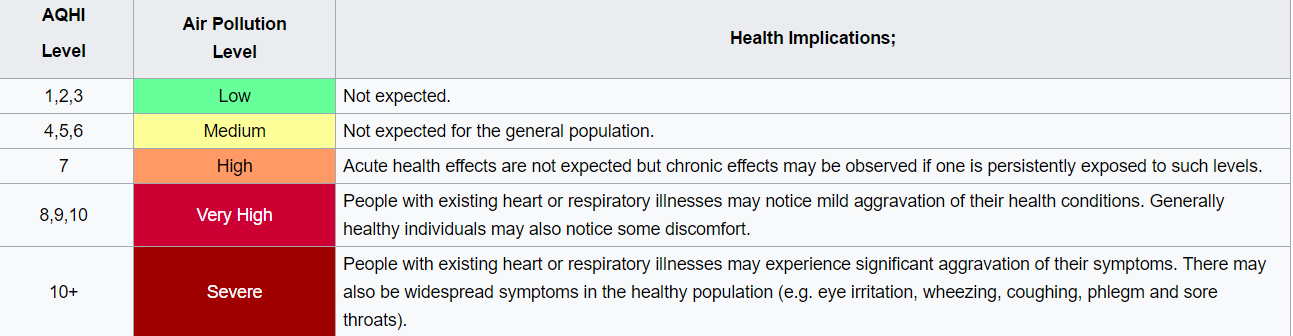
Prof. Jaume

January 28, 2017

Data Set Source: <http://iot.ee.surrey.ac.uk:8080/datasets.html>

Data focus: Pollution dataset with geospatial coordinates.

The data set that will be the focus of our project is the pollution dataset. The sponsors simulate one sensor for each of the traffic sensor at the exact location of this traffic sensor. The data values are labeled as carbon\_monoxide, nitrogen\_dioxide, sulfure\_dioxide, particulate\_matter and ozone index levels. These variables are defined and derived from the Air Pollution Index (API). The Air Pollution Index (API) is a way to describe the air quality of a particular region. This methodology uses several sets of air pollution data and gives a general level of the air conditions in the area. This is very important because regions of the world air quality is poor and can have health effect on the human body. Regions of the world that have a high API can develop asthma and some heart or respiratory illnesses. This index was formerly used in mainland China and Hong Kong. The API index and the air quality objectives started around 1987. The data of pollutant levels are measured over varying periods usually longer than 24 hours. Below is a chart of the levels of air pollution for a given city and some of the health implications.



Source: <https://en.wikipedia.org/wiki/Air_Pollution_Index>

About the data set:

The sponsored data set is from the City of Brasov in Romania and the timeline is from August 2014 - October 2014. The data set are in a CSV format and are broken down into measurements in a particular observation point. Pollution from urban atmosphere is mainly due to air pollutant emissions and transmissions. Emissions of sulfur dioxide is a main cause from the activities in the steel industry, the oil refineries, the motor vehicles, the thermo-electric power stations. Nitrogen oxide emissions are largely caused by electricity and thermal industry, road traffic and manufacturing industries. Engine emissions from road traffic are an important source of pollution around the world and the data produced can give great insight of the polluters in the area. Primary pollutants are those emitted directly into the atmosphere e.g. Sulphur dioxide, nitrogen oxides, soot particles, while secondary pollutants are produced by reactions between primary pollutants. The Ozone is formed over the urban areas by reactions between primary pollutants and normal builders of the atmosphere. Carbon monoxide and nitrogen oxides are the main pollutants emitted from burning these fuels. Soot and Sulphur dioxide are the primary byproducts produced mainly by burning fossil fuels in electric power stations, such as oil and coal. With each of the sensors, the data is streamed every five minutes and assigned a numerical value. There are 449 station points that collect five variables with the timestamp. To improve air quality and also protect the atmosphere are required control measures for pollutants emissions. To assess the pollution degree of the atmosphere the API calculates pollutants emissions to determine ambient air quality. Emissions are measured by appropriate methods of assessment, specific to each pollutant in part based on emission factors and activity indicators. Most sensitive air pollution control strategies involve methods that reduce, collect, capture or retain pollutants before they are released into the atmosphere. From an energy conservative perspective, the best method is to reduce emissions by burning a small amount of fuel. Therefore improving urban air quality can be obtained by converting people to use public transport instead of private cars

I uploaded the csv files and took the average of each variable. I plotted (see attached) the variables to get an insight on the data set. Going forward plan on creating a heat map of the area to include all 449 stations and their concentration of pollutants. This can yield great insight of the areas and reason that the areas are so highly polluted. Local governments can use this to focus their efforts to deploy resources to combat these high levels.