**Assignment – 11**

**DS-607-Capstone**

Neha Kumari

**Conclusion:**

K-means algorithm was applied to get the healthiest area in the smart city.

K means clustering, k = 9 with healthiest area at longitude=10.18032 and latitude=56.17061

The final minimum concentration levels of ozone, is 81.92253.

The data set consists of 221 files; each file represents the air pollutant levels of a place in a city during various time intervals and days (which means the latitude and the longitudes remain constant in each file). Since the data huge some data transformations was needed to work with the data and get some meaningful output from the algorithm.

So we aggregated each attribute in each file. We did this by taking the arithmetic mean of each attribute. Now this arithmetic mean represented the overall air pollutant level of each attribute at a certain location in the city.

X = ∑ xi / n

Where,

X = arithmetic mean of an air pollutant in certain location

n = number records in each file for that location

**Discussion**

In this study the main objective to use K-means is to find the healthiest area in the city. The clustering is done to find the minimum levels of the air pollutant. Since we are dealing with five air pollutants, the area having minimum concentration of ozone is considered as the healthiest area.

Note that we can use the levels of other air pollutant also (sulfur dioxide, nitrogen dioxide, carbon monoxide and particulate matter, it is just a matter of choice). Ozone is not emitted directly in the environment by some sources. It is created as a result of the reaction of nitrogen oxide with the sunlight and some volatile organic compounds in the environment. The reactions are encouraged by the sunlight and temperature. So the weather and amount of sunlight has a vital role in the concentration levels of ozone. Emissions from vehicles and other stationary sources can form high ozone concentrations, which can spread, over large regions. For these reasons the ozone level is taken as the deciding factor in finding the healthy and unhealthy area in this project.

This study can be further expanded to Smart city road Traffic. As the traffic and air pollution are closely linked, we can say that places with high air pollution are also high on traffic and vice versa. The Smart City is a connected city so these findings can be useful for other smart city projects.

Further, in this study ozone concentration level was used to determine if an area is polluted or not. In the data set there are other pollutant as well like particulate matter, nitrogen dioxide, sulfur dioxide etc. These pollutants were not taken into consideration in this study. Ozone is considered to be the prime pollutant among the given air pollutants in the data set so this study was restricted to ozone. Including other pollutants in determining the healthiest area can expand this study.

The K-means algorithm is a simple, unsupervised learning algorithm. It takes the input data set D and the input parameter, K. K is the number of clusters we want to group our data in. While any value of K can be chosen for the algorithm to run, but the results may not be useful. Ideally the value of K should neither be too small nor too large. In our study we will try for different values of K to understand the results.

The grouping of the data in K-means clustering depends on the similarity basis. The partition of the data in K clusters is done in such a way that the inter cluster similarity is low but the intra cluster similarity is high. The K-means algorithm works as follows:

First it randomly choses K objects, each of which initially defined as cluster mean or cluster centroid. For the remaining of the objects each object is assigned to these K objects to which it is most close (The closeness is measured in terms of Euclidean distance). It then re-calculates the cluster mean for each cluster also called the centroid. The process is repeated until there is no major change in the mean value of cluster. This phenomenon is called convergence.

In our study, the data set met most conditions for the K-means to work well. The data is clean, structured and complete.

We are using Zeppelin platform for this implementation, it supports many languages, we used R, SQL and Pyspark for this study. So far the arithmetic mean of each of the five-air pollutant is taken in each file. This will represent the pollution level at a certain location. Next the task was to combine this aggregated data into a single frame.

To begin with, there has been exploratory data analysis done. Starting with the box plots. The box plots gave the insights of each air pollutant. It helps in understanding the median value and the presence of outliers in the data set if any. The initial study has revealed that the particulate matter and sulfur dioxide to be highly skewed. Also there are some outliers in the data set. Mainly in ozone, particulate matter, carbon monoxide and sulfur dioxide. In general the range of ozone and sulfur dioxide is high in the environment compared to other air pollutants. The K means algorithm have been applied and we got the final result as below. Also the location has been graphically represented on google maps in Zeppelin.

K =9

longitude latitude avg\_ozone particulate

1 10.18442 56.16306 116.07195 123.69064

2 10.16102 56.17356 136.82826 125.90213

3 10.18583 56.15928 115.45212 96.45877

4 10.18759 56.17984 133.05192 107.95164

5 10.15406 56.16399 95.34281 100.19310

6 10.17816 56.17024 117.89435 110.68795

7 10.17013 56.17043 101.31702 123.55242

8 10.18032 56.17061 81.92253 121.50843

9 10.18094 56.15773 106.83063 111.85106

With k = 9, for cluster 8(highlighted above), the ozone concentration is minimum, thus we find the healthiest area at longitude 10.18032 and latitude 56.17061 and ozone concentration level 81.92253