**Assignment – 10**

**DS-607-Capstone**

Neha Kumari

**Results:**

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 there is some data transformations needed to work with the data and get some meaningful output from the algorithm.

So we need to aggregate each attribute in each file. We do this by taking the arithmetic mean of each attribute. Now this arithmetic mean will represent 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

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 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.

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

We are using Zeppelin platform for this implementation. 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 results are shown below:

K = 4

longitude latitude avg\_ozone particulate

1 10.17102 56.16822 94.32454 117.74445

2 10.18229 56.16382 113.46153 117.07341

3 10.17705 56.17164 130.43995 111.18956

4 10.17615 56.16417 109.35957 99.12508

K = 5

longitude latitude avg\_ozone particulate

1 10.17102 56.16822 94.32454 117.74445

2 10.18229 56.16382 113.46153 117.07341

3 10.17705 56.17164 130.43995 111.18956

4 10.17615 56.16417 109.35957 99.12508

K =6

longitude latitude avg\_ozone particulate

1 10.18324 56.17770 132.73874 113.9244

2 10.17545 56.16624 89.56244 123.9848

3 10.17683 56.16007 106.34493 114.8368

4 10.18525 56.16414 116.16317 121.0616

5 10.17839 56.16568 116.44108 101.5405

6 10.15835 56.16847 96.50405 100.9700

K = 7

longitude latitude avg\_ozone particulate

1 10.17647 56.16176 105.60458 115.36039

2 10.18355 56.15940 116.92120 96.96951

3 10.18582 56.16181 115.38675 124.21001

4 10.17874 56.17044 117.82135 110.40380

5 10.17218 56.16558 88.92402 124.20092

6 10.15835 56.16847 96.50405 100.97003

7 10.18290 56.17964 135.92837 115.61803

K = 8

longitude latitude avg\_ozone particulate

1 10.18023 56.17058 118.01701 112.23172

2 10.15081 56.16349 95.25544 99.63985

3 10.17543 56.16871 87.21657 123.12371

4 10.17732 56.16306 105.34921 113.45741

5 10.18570 56.16142 114.37760 98.11670

6 10.17707 56.17570 132.45073 124.08558

7 10.18096 56.15995 110.75201 125.16890

8 10.18214 56.17365 132.71768 105.81515

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

K=10

longitude latitude avg\_ozone particulate

1 10.17052 56.17031 100.98485 123.39569

2 10.18366 56.16245 115.32841 124.68005

3 10.17410 56.15099 114.33345 89.32272

4 10.18570 56.16588 111.94914 102.53156

5 10.18032 56.17061 81.92253 121.50843

6 10.18209 56.17875 137.23991 119.00751

7 10.15168 56.16368 94.99104 101.03512

8 10.17870 56.16995 127.04528 103.73866

9 10.18328 56.17163 118.22879 112.93678

10 10.17883 56.15726 106.61785 112.47661

With k = 10, for cluster 5 the ozone concentration is minimum, thus we find the healthiest area at longitude 10.18032 and latitude 56.17061