**Assignment – 10**

**DS-607-Capstone**

Neha Kumari

**RESULTS**

**Algorithm**: K-means clustering

**Data Set structure**: ozone, particulate matter, carbon monoxide, sulfur dioxide, nitrogen dioxide, longitude and latitude

**Data Description**: Air pollutant levels of 221 different locations in Denmark.

**Data preparation**: Aggregate each pollutant in each file to get the mean of each air pollutant, and then group the data by longitude and latitude. Get all these values in a single file; these values will now represent over all pollution level of each pollutant at a certain location in the city.

**Data set link**: [http://iot.ee.surrey.ac.uk:8080/datasets.html - pollution](http://iot.ee.surrey.ac.uk:8080/datasets.html#pollution)

**Platform:** Zeppelin

**Programming Language**: Pyspark, sql, R

The initial findings were:

* The pollutant levels were around the same range. The Air Quality Index (AQI) for all the pollutants was around the range of 70 to 150.
* Slight Co-relation between sulfur dioxide and nitrogen dioxide, for other variables the co-relation was very less (<0.5).
* The data was almost normally distributed with mean as well as median around 110-113 AQI
* The data was very slightly skewed for some of the variables (mainly particulate matter, right skewed).
* There were only few outliers in the data. Amongst all Ozone had the most (around 11), particulate matter and carbon monoxide had the minimum (only 2).

Two approaches were used to apply K-means clustering on the data set.

Approach1: Full Features

In this approach the data set consisted of all the 5 air pollutants, since we don’t know the initial number of clusters, so started with initial value of 3 till 10 and looked for the minimum concentration of ozone

Pollutants: Ozone, Particulate Matter, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide

Results are shown below:

K = 3

avg\_ozone particulate monoxide sulfur nitrogen

1 103.3060 114.6799 107.7228 117.2473 118.1678

2 110.3113 111.8108 112.2988 101.4899 103.1753

3 122.0689 109.1919 116.3790 115.5235 112.0790

K=4

avg\_ozone particulate monoxide sulfur nitrogen

1 119.6578 102.0581 108.9062 116.3143 108.18656

2 103.8930 114.2321 108.8146 118.4781 119.08962

3 121.2742 117.4912 116.9557 107.3068 114.56501

4 103.0570 112.6837 115.9608 102.2227 98.91799

K=5

avg\_ozone particulate monoxide sulfur nitrogen

1 130.8930 117.6205 112.3691 107.6710 109.2275

2 115.0258 100.0200 109.4065 118.6586 113.6807

3 109.2033 108.9456 111.7973 100.1192 101.3911

4 108.4146 114.9965 125.0470 115.6794 110.9432

5 104.8908 118.3312 104.7244 114.9681 119.3882

K=6

avg\_ozone particulate monoxide sulfur nitrogen

1 112.93882 117.35925 102.8370 111.5981 116.73902

2 120.91786 112.58709 124.3464 112.1318 117.03663

3 123.38678 112.99449 109.8004 104.0503 99.92181

4 113.21394 97.09827 108.8874 118.3516 110.80682

5 97.58487 112.75821 118.7827 106.0150 101.66312

6 98.95310 116.68341 114.1469 120.9267 120.83930

K=7

avg\_ozone particulate monoxide sulfur nitrogen

1 115.45835 108.1195 114.3776 101.6584 96.40714

2 94.58225 113.9676 118.6713 106.9700 102.04157

3 102.85383 118.5240 106.6239 118.7502 122.39703

4 112.77864 100.8235 103.8428 117.7609 111.64093

5 134.59073 115.8818 115.1955 111.3208 107.96933

6 116.13599 118.5576 106.1965 101.8605 112.93617

7 114.30093 110.2111 124.6781 117.1969 115.75666

K=8

avg\_ozone particulate monoxide sulfur nitrogen

1 110.63875 113.57679 104.4940 113.72944 103.9318

2 120.77631 101.32774 113.9208 119.91294 115.3013

3 111.14871 113.27778 125.6396 117.54208 115.1267

4 130.42081 118.56950 111.9069 106.07908 109.1151

5 103.13010 94.31367 101.4331 112.85116 116.6343

6 107.61763 118.92139 105.3398 115.11473 123.9977

7 91.13839 117.00705 118.2684 108.89922 102.1843

8 113.82606 106.30684 116.0804 97.39338 101.8633

K=9

avg\_ozone particulate monoxide sulfur nitrogen

1 107.29453 103.8855 104.0500 115.32283 113.66056

2 115.87110 110.4960 109.1231 100.86032 97.73822

3 115.25104 123.0852 103.8103 112.25667 116.90096

4 93.06211 117.7377 120.5376 106.80234 99.49122

5 135.17136 117.8236 123.2362 103.67201 110.02300

6 113.04748 111.0406 122.7118 117.47901 114.96446

7 126.69651 106.8628 105.1469 122.85601 104.10338

8 111.35458 100.0188 114.7872 96.10198 116.17782

9 96.14344 116.8084 109.1165 119.23561 125.33828

K=10

avg\_ozone particulate monoxide sulfur nitrogen

1 106.93403 98.82974 102.8302 116.12671 118.26981

2 85.06180 125.69522 111.1593 100.80956 97.05194

3 116.35451 107.57061 115.8282 98.64193 96.15474

4 115.52903 107.40190 103.9809 117.67564 102.26439

5 113.37976 121.60003 105.7415 113.11510 116.78425

6 136.89854 120.92265 115.2964 108.97840 102.40764

7 119.59255 107.59357 123.5841 117.58886 116.02013

8 98.68995 117.36842 113.6513 121.99025 122.17094

9 120.76275 109.99191 107.9365 96.68356 117.24993

10 100.62492 110.05658 121.4600 110.97163 105.65457

We got minimum ozone concentration at K=10, cluster=2 as highlighted above.

Approach2: Reduced Features

In this approach, we used reduced data set with ozone and particulate matter in the data set

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.

Approach1 gave us minimum ozone levels as 85.060180 and maximum ozone levels as 137.23991

Aproach2 gave us minimum ozone levels as 81.92253 and maximum ozone levels: as 136.89854

Comparing the two approaches we found with reduced features data set results were better so we used the results of approach 2.

Thus we find the healthiest area at longitude 10.18032 and latitude 56.17061 where ozone level is found minimum, 81.92253 on AQI.