Are You Breathing In Clean Air?

An Exploration of air pollutants

Neha Kumari
Saint Peter's University and Data Science Institute
nkumari@saintpeters.edu

Abstract

The data produced by IoT (internet of things) is enormous and data mining techniques can be used to get hidden information, which is of high business value. Smart cities are completely based on IoT.Air pollution is increasing rapidly in the smart cities and has adverse effects on human health. The sources of pollution are many including road traffic, industrial gases and others.

Introduction

In this study we try to find the healthiest areas, which are suitable for leaving, in the smart cities by using K-means clustering. The dataset is generated from the City Plus project. The data is enormous and dynamic due to the number of sensors deployed in the same location and their measurement frequency. This data consists of 5 air pollutants namely ozone, sulfur dioxide, nitrogen dioxide, carbon monoxide and particulate matter. There are 3 more fields in the data set namely- Longitude, latitude and timestamp. The main objective of the CityPulse project is to be able to use this real time data for building real time applications.[5]

Aliq, The analysis is based on the air pollution data collected from the city plus project. There are five air pollutants in the data set Ozone, Carbon Monoxide, particulate matter, Sulphur Dioxide and Nitrogen Dioxide. In addition there is longitude, latitude and timestamp field. There are total 449 files in the data set; each file has the measure of these five air pollutants at a specific location but during different hours of the day.

Main Objectives

1. Exploratory data analysis through visualizations.
2. Levels of air pollutants in different areas.
3. Identify the healthy and unhealthy areas in the city.
4. Clustering algorithm, K-means.
5. Map Visualizations.

Materials and Methods

The data is enormous and dynamic due to the number of sensors deployed in the same location and their measurement frequency. This data consists of 5 air pollutants namely ozone, sulfur dioxide, nitrogen dioxide, carbon monoxide and particulate matter. There are 3 more fields in the data set namely- Longitude, latitude and timestamp. The K-means algorithm is an 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. 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 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.

K-means Clustering

1. Let X = x1, x2, x3, x4xn be the data points in the data set.
2. Randomly assign objects in the data point to these K clusters. These data points are the initial centroids of each k clusters.
3. For the remaining objects, calculate the distance between each data point and the cluster centers.
4. Assign the data point to that cluster whose distance between the cluster center and the data point is minimum..
5. Now again calculate the centroid of each cluster.
6. Continue this process until no data point is reassigned to a new cluster
7. Since K is an input parameter, we will try with different values of K that gives us more meaningful and clear results. Usually the value of K should neither be very large nor too small

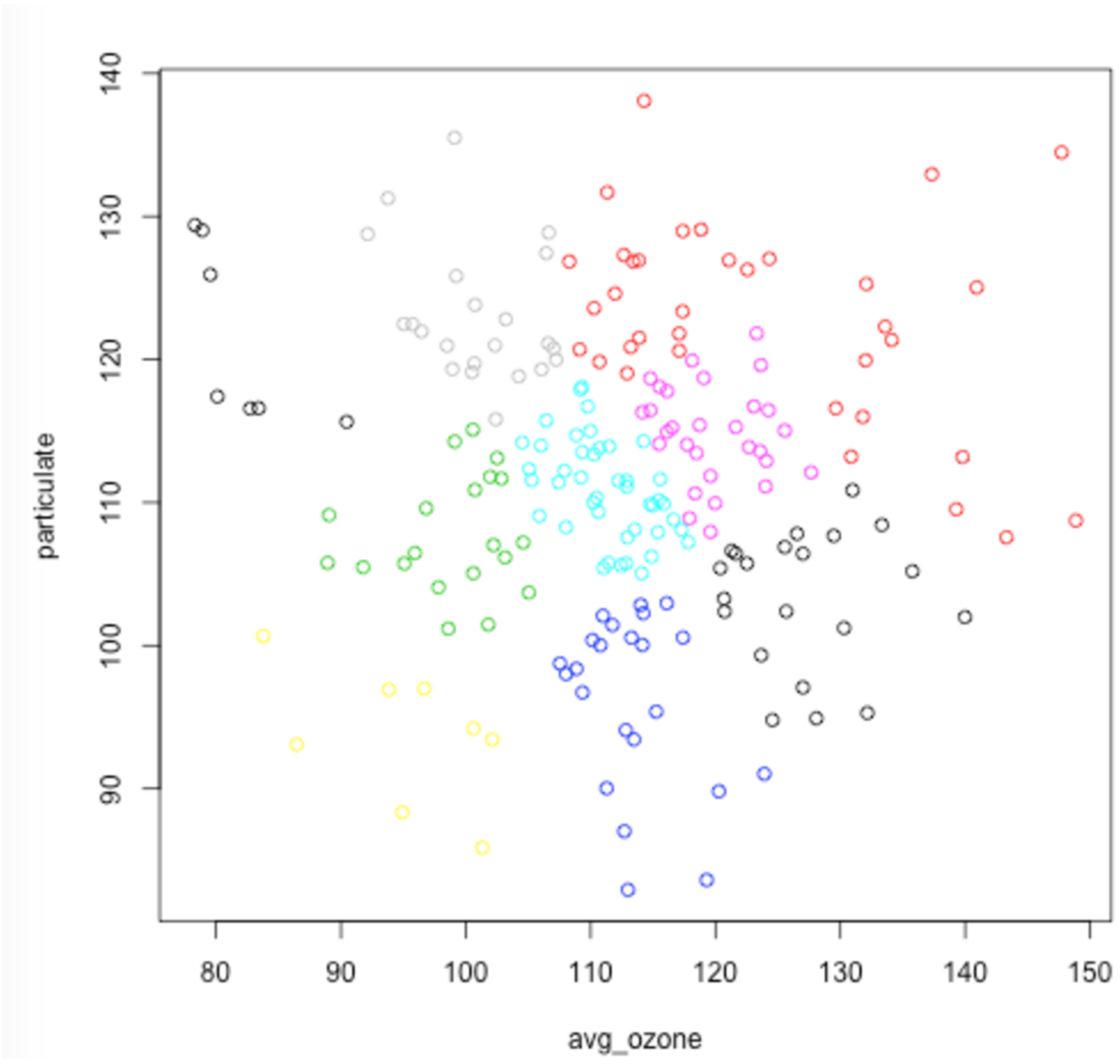

Figure 1: Clustering on Pollution levels of ozone in the city

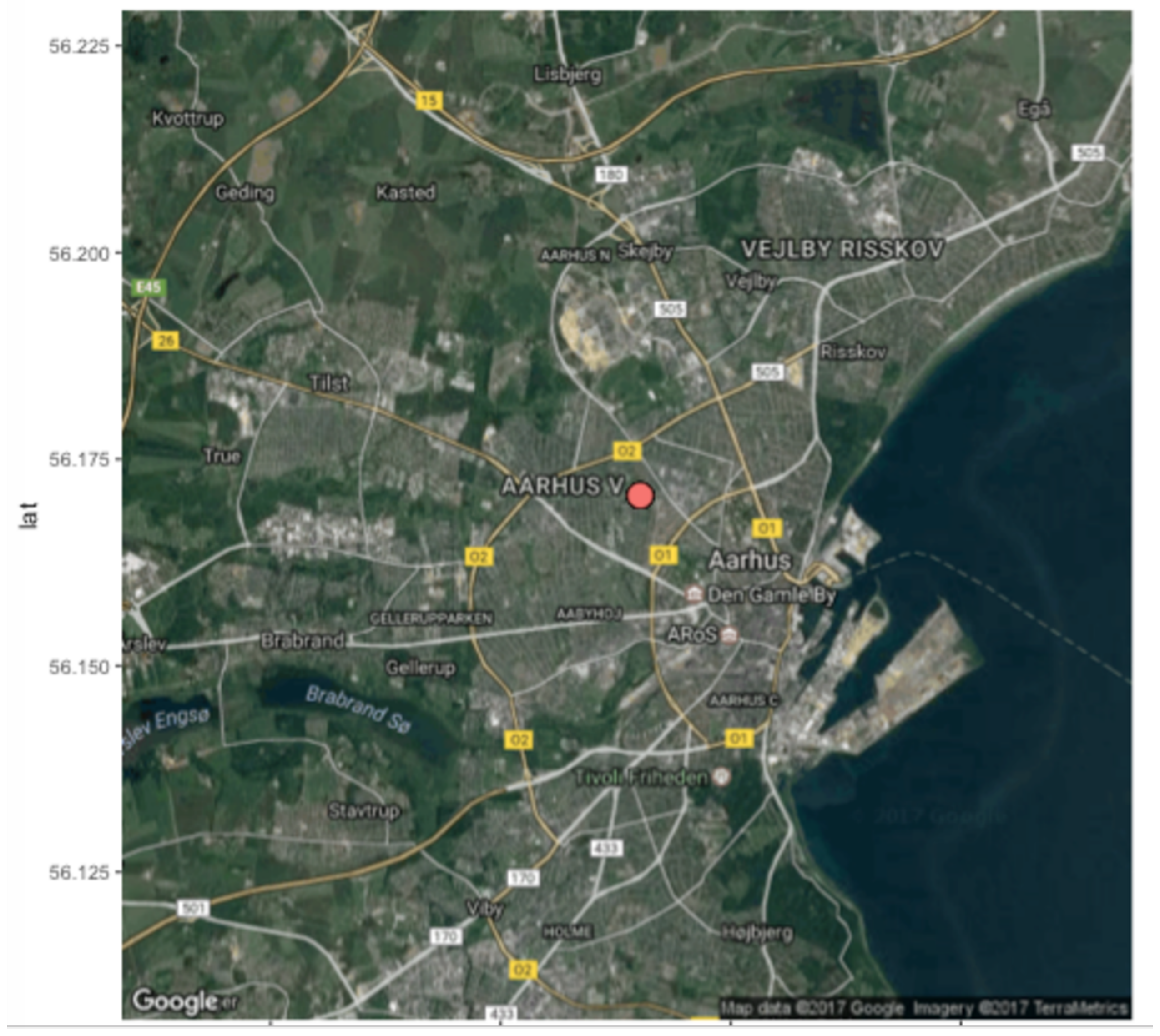

Figure 2: Visualizing the most healthy area in the city

Conclusions

- K-means clustering technique had been applied on the air pollution data set from City Pulse project. K value changed from 3 to 10.
- The cluster analysis is focused on the Ozone (O3) average concentration.
- Healthy and unhealthy location in the City Pulse project has been determined which helps in getting smart environment in smart city.

References

[1] Denmark Aarhus. City pulse data set collection. http://iot.ee.surrey.ac.uk:8080/datasets.html.

[2] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7):881–892, 2002.

[3] Nikhil R Pal, Kuhu Pal, James M Keller, and James C Bezdek. A possibilistic fuzzy c-means clustering algorithm. *IEEE Transactions on Fuzzy Systems*, 13(4):517–530, 2005.

[4] Anthony Seaton, D Godden, W MacNee, and K Donaldson. Particulate air pollution and acute health effects. *The lancet*, 345(8943):176–178, 1995.

[5] Weiqiang Wang and Ying Guo. Air pollution pm2. 5 data analysis in los angeles long beach with seasonal arima model. In *Energy and Environment Technology, 2009. ICEET'09. International Conference on*, volume 3, pages 7–10. IEEE, 2009.