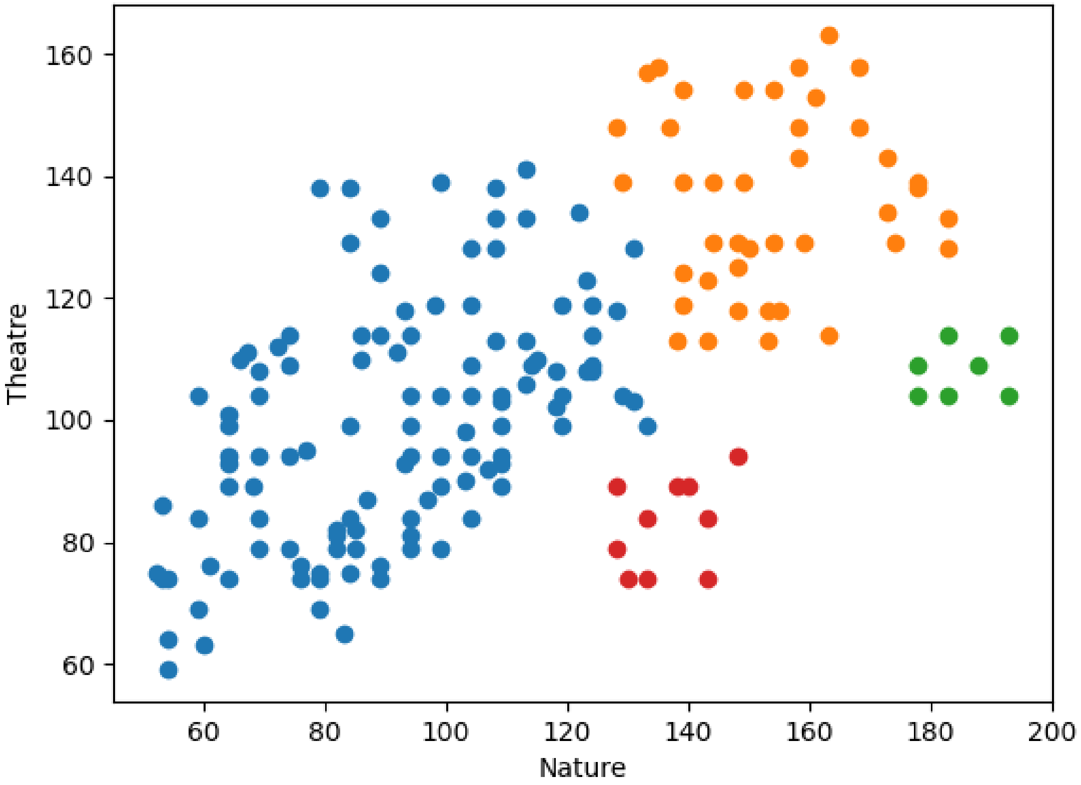
DATA MINING HOMEWORK 1 REPORT

1) ***About data:*** User interest information extracted from user reviews published in holidayiq.com about various types of point of interests in South India.

   Parameters:

D:	data set from "buddymove_holidayiq.csv"
Radius:	10
MinPts:	5
Considered Attiributes:	Nature and Theatre
Source:	https://archive.ics.uci.edu/ml/datasets.php

   Here you can see extracted clusters:

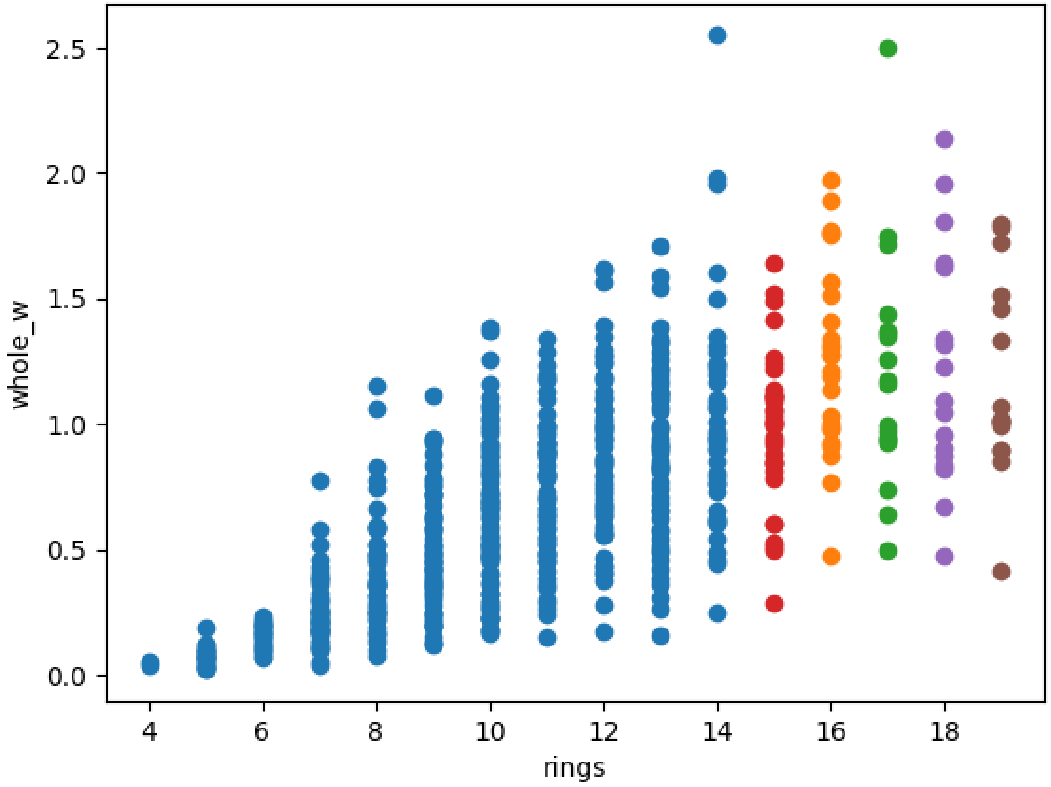


2) ***About data:*** Abalones' physical measurements.

Parameters:

D:	data set from "abalone.csv"
Radius:	1
MinPts:	15
Considered Attiributes:	rings and whole_w
Source:	https://archive.ics.uci.edu/ml/datasets.php

Here you can see extracted clusters:

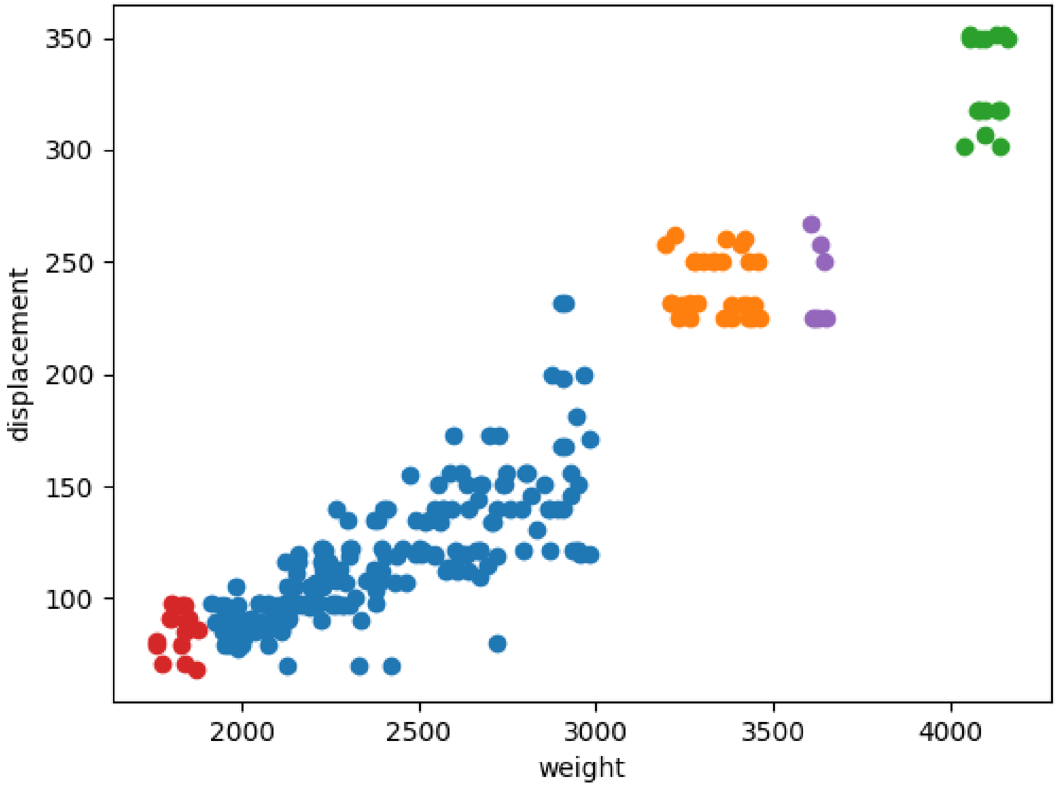


3) ***About data:*** Revised from CMU StatLib library, data concerns city-cycle fuel consumption.

Parameters:

D:	data set from "abalone.csv"
Radius:	40
MinPts:	6
Considered Attiributes:	displacement and weight
Source:	https://archive.ics.uci.edu/ml/datasets.php

Here you can see extracted clusters:



How parameters effect the results:

epsilon or radius: If radius is too high it can wrap 2,3 or more clusters into 1 cluster which is inappropriate. There is a higher chance of finding only 1 cluster at the end which is useless is most situations. But if it is too low it can break an actual cluster into more clusters which is also inappropriate. If radius is too low you probably found clusters that distance between clusters are unbalanced such as 3 and 9000 unit distance.

Minimum points(minPts): if minPts is too high you may end up finding too much noise data in the set which cause loss valuable data. And it may even end up not finding a cluster which means failure of this algorithm most of the times. But if it is too low noise datas may create their own clusters which means that exceptional data distract your model. So this cause hih number of clusters also.

How to automatically decide on the parameters of DB-Scan?

radius:

Here is a pseudocode of an algorithm to decide the “radius” parameter:

```
for i to n:
    for j to n:
        find distance( i , j )
    save (lets say 4) minimum distances among all
plot the points
ciritical point of curve of plotted points is the “radius”
```

By benefitting from distances among points this algorithm derives a radius value based on serious change in distances.

MinPts:

Here is a pseudocode of an algorithm to decide the “MinPts” parameter:

```
initialize MinPts to 0
for i to n:
    for j to n:
        save distance( i , j ) to Distances
sort Distances in ascending order
for i to (n^2)-1:
    save Distances[i] - Distances[i+1] to Differences
for each peak in Differences
    increment MinPts
```

By benefitting distance differences between points I figured out an algorithm to pick a nice MinPts value. High number of noise points may hurt this algorithm as long as this algorithm only checks for peaks in defferences.

Yunus Gedik

141044026