Assignment 1

STK 802

Connor McDonald
u16040725

Behavioural Analytics

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Department of Computer Science
University of Pretoria
Date: 2 September 2021

1 k-NN density estimate

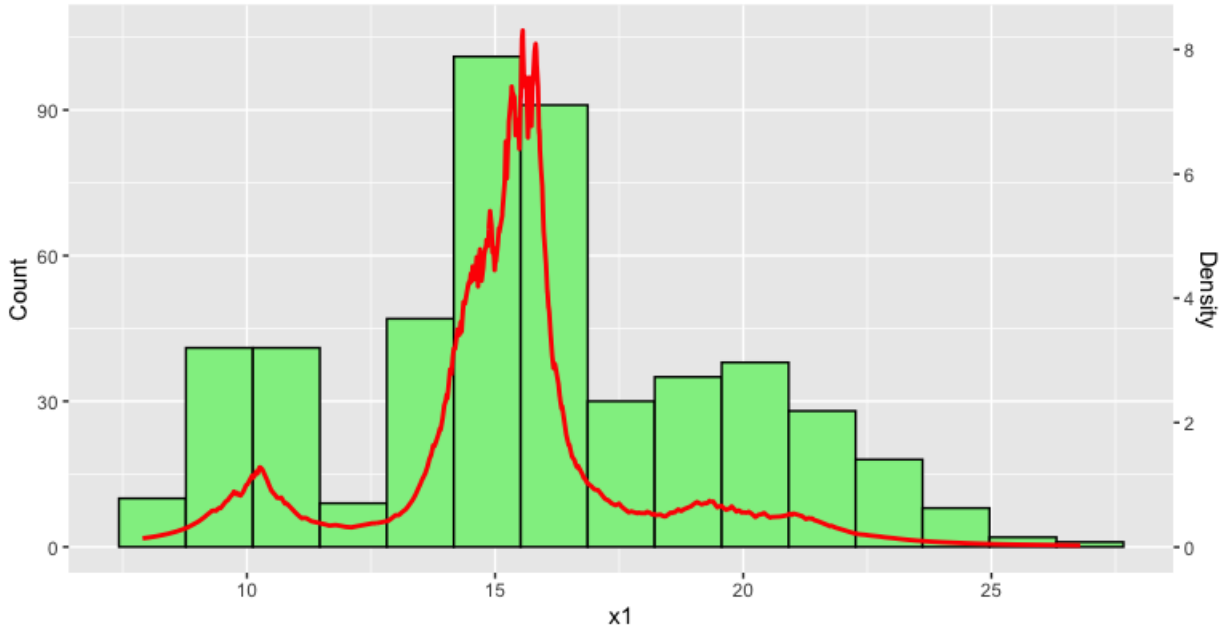

Figure 1

2 Effect of k On The Number of Modes and Clusters

The histogram in figure 1 suggests that there are three modes within the data. Furthermore when experimenting with different values of k, it was observed that this was also the most common number of modes found with k-values between 10 and 100 (see table 1).

Excessively high k-values lead to an overly smooth distribution making it difficult to identify modes, whilst excessively small k-values lead to very 'spikey' distributions and did not generalise the data well enough. However, when looking at figures 2 and 3, we can see that the 70-nn density estimate achieves a balance between identifying multiple clusters whilst also presenting an adequately smooth distribution of points.

Therefore, 70-nn can be declared the optimal k-value for this data set and thus the optimal number of clusters is **3**.

Table 1: Number of Modes Found at Different Values of K

K	**Modes Found**
10	25
20	10
30	7
40	6
50	4
60	3
70	3
80	3
90	2
100	1

3 Cluster Solution in Graphical Form

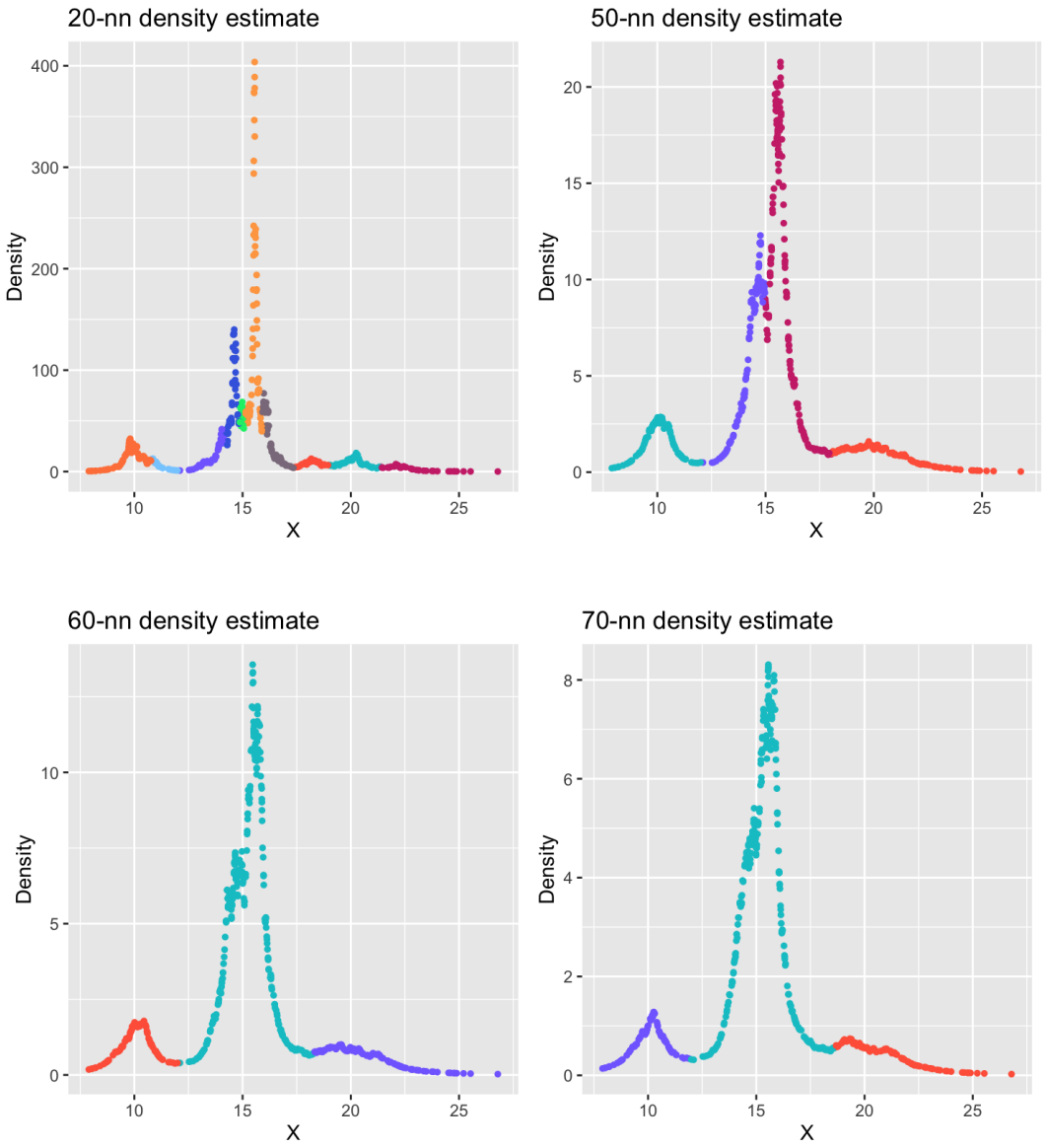

Figure 2

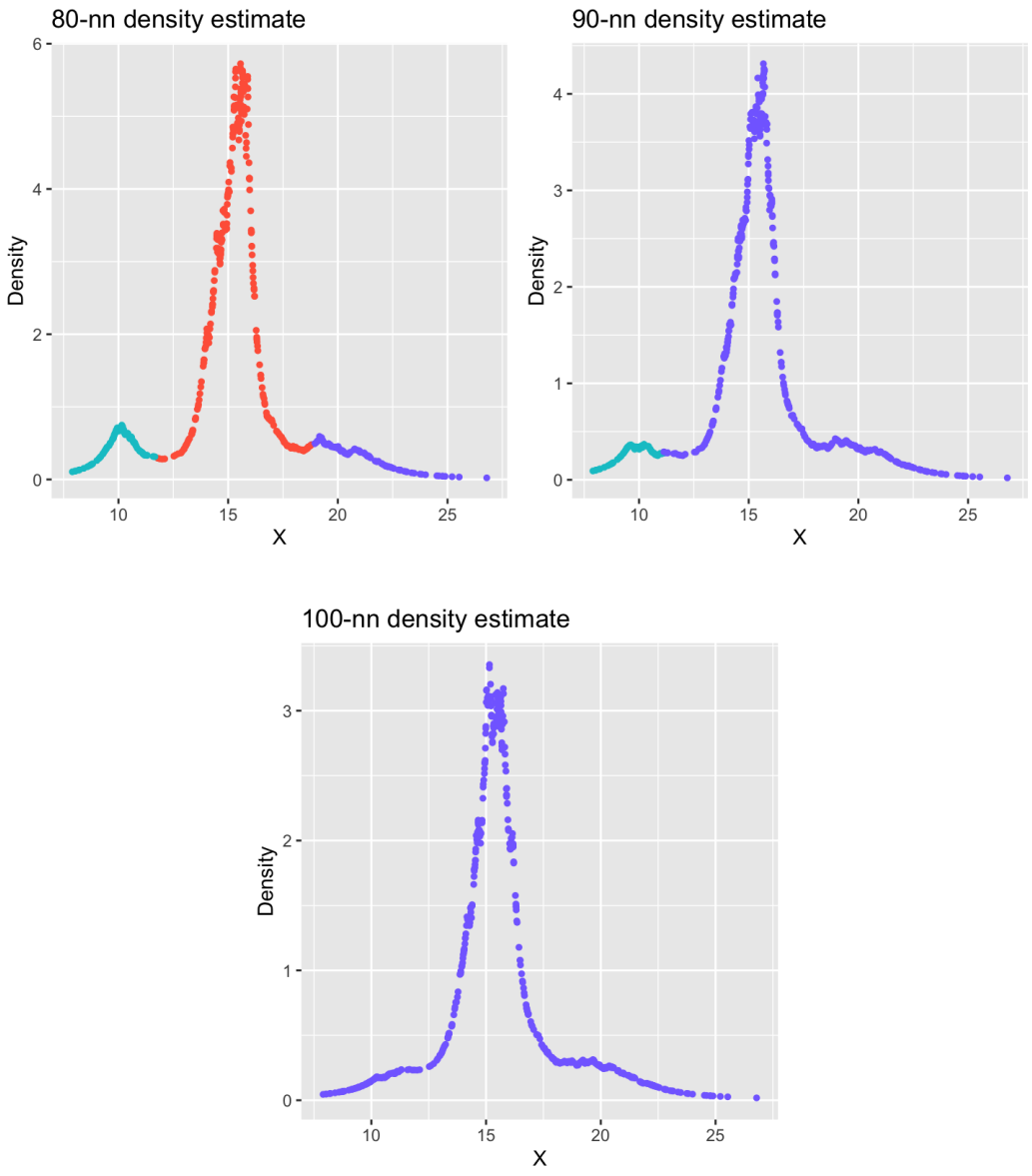
80-nn density estimate
90-nn density estimate
100-nn density estimate
Density
X
0
1
2
3
4
6
10
15
20
25

Figure 3