Olabode Alamu

Guide to Engineering Data Science

1498663

Dataset

$X = [61, 10, 32, 19, 22, 5, 100, 29, 36, 14, 49, 3]$

$Y = [2.3, 2.7, 1.7, 1.9, 2.1, 2.8, 1.8, 2.4, 5.9]$

(a) Find the outliers using Tukey method

Arrange in order

$X = [3, 5, 10, 14, 19, 22, 29, 32, 36, 49, 61, 100]$

median $= \frac{22 + 29}{2} = 25.5$

first quartile at $[3, 5, 10, 14, 19, 22]$

$Q1 = \frac{10 + 14}{2} = 12$

Third quartile at $[29, 32, 36, 49, 61, 100]$

$Q3 = \frac{36 + 49}{2} = 42.5$

Interquartile Range $IQR = Q3 - Q1$

$= 42.5 - 12 = 30.5$

Lower bound of box plot $= Q1 - 1.5(IQR)$

$= 12 - 1.5(30.5)$

$= -33.75$

Upper bound of box plot $= Q3 + 1.5(IQR)$

$= 42.5 + 1.5(30.5)$

$= 88.25$

Outlier present $= 100$

$$Y = [2.3, 2.7, 1.7, 1.9, 2.1, 2.8, 1.8, 2.4, 5.9]$$

Arrange in order

$$Y = [1.7, 1.8, 1.9, 2.1, 2.3, 2.4, 2.7, 2.8, 5.9]$$

Median = 2.3

first quartile in $[1.7, 1.8, 1.9, 2.1, 2.3]$

$$Q1 = 1.9$$

Third quartile in $[2.3, 2.4, 2.7, 2.8, 5.9]$

$$Q3 = 2.7$$

$$IQR = Q3 - Q1 = 0.8$$

lower band of box plot $= Q1 - 1.5(IQR) = 0.7$

Upper band of box plot $= Q3 + 1.5(IQR) = 3.9$

Outlier point $= 5.9$

LOF

Data points

A(15.03393, 15.67469)

B(12.75117, 15.75761)

C(13.87044, 15.52042)

D(19.03499, 12.02895)

E(28.54179, 21.59978)

k = 2

Using manhattan distance

dist(A,B) = 2.36568

dist(A,C) = 1.317755

dist(A,D) = 7.6468

dist(A,E) = 19.43294

dist(B,C) = 1.3564625

dist(B,D) = 10.0124826

dist(B,E) = 21.63278

dist(C,D) = 8.6566

dist(C,E) = 20.750697

dist(D,E) = 19.077625

$dist_k(0)$

$dist_2(A) = 2.36568$

$dist_2(B) = 2.36568$

$dist_2(C) = 1.35646$

$dist_2(D) = 8.65602$

$dist_2(E) = 19.43294$

k - distance neighborhood of 0

$N_2(A) = (B, C) = 2$

$N_2(B) = (A, C) = 2$

$N_2(C) = (B, A) = 2$

$N_2(D) = (A, B) = 2$

$N_2(E) = (B, D) = 2$

$Lrd_k(A) = ||N_2(A)||$

$$lrd_2(A) = \frac{2}{2\text{-}36568 + 1\text{-}35646} = 0.537325$$

$$lrd_2(B) = \frac{||N_2(B)||}{reachdist_2(A \leftarrow B) + reachdist_2(C \leftarrow B)}$$

$$reachdist_2(A \leftarrow B) = max\left[dist_2(A), dist(A,B)\right]$$

$$= max\left[2\text{-}3656811, 2\text{-}3656511\right]$$

$$= 2\text{-}3656811$$

$$reachdist_2(C \leftarrow B) = max\left[dist_2(C), dist(C,B)\right]$$

$$= max\left[1\text{-}3564625, 1\text{-}3564625\right]$$

$$= 1\text{-}3564625$$

$$lrd_2(B) = 0.537325$$

$$lrd_2(C) = \frac{||N_2(C)||}{reachdist_2(B \leftarrow C) + reachdist_2(A - C)}$$

$$reachdist_2(B \leftarrow C) = max\left[dist_2(B), dist_2(B,C)\right]$$

$$= max\left[2\text{-}36568, 1\text{-}3564625\right]$$

$$= 2\text{-}3656811$$

$$reachdist_2(A \leftarrow C) = \max\left[dist_2(A), dist(A,C)\right]$$

$$= \max\left[2.38568, 1.3177546\right]$$

$$= 2.3656811$$

$$lrd_2(C) = 0.4227112$$

$$lrd_2(D) = \frac{\|N_2(D)\|}{reachdist_2(A \leftarrow D) + reachdist_2(B \leftarrow D)}$$

$$reachdist_2(A \leftarrow D) = \max\left[dist_2(A), dist(A,D)\right]$$

$$= 7.64680$$

$$reachdist_2(B \leftarrow D) = \max\left[dist_2(B), dist(B,D)\right]$$

$$= 10.01248$$

$$lrd_2(D) = 0.11325487$$

$$lrd_2(E) = \frac{\|N_2(E)\|}{reachdist_2(B \leftarrow E) + reachdist_2(D \leftarrow E)}$$

$$reachdist_2(B \leftarrow E) = \max\left[dist_2(B), dist(B,E)\right]$$

$$= 21.63278$$

$$reachdist_2(D \leftarrow E) = 19.07763$$

$$lrd_2(E) = 0.0491275$$

$$LOF(A) = \left(lrd_2(B) + lrd_2(C)\right) \cdot \left[reachdist_2(B \leftarrow A) + reachdist_2(C \leftarrow A)\right]$$

$$= 3.57339$$

$$LOF(B) = lrd_2(A) + lrd_2(B) \cdot \left[reachdist_2(A \leftarrow B) + reachdist_2(C \leftarrow B)\right]$$

$$= 3.57339$$

and for the remaing factors

$$LOF(E) = 26.48537$$
$$LOF(D) = 18.97755$$
$$LOF(C) = 5.08456$$
$$LOF(A) = 3.57339$$
$$LOF(B) = 3.57339$$