24-02-22

# Machine learning
## CSPE-65
## Cycle test - 1

10611910O
Rajneesh.

## Question ①

(A) supervised.

---

## Question ②

(i) values → 4, 7, 9, 8, 12, 80, 15

Ascending order = 4, 7, 8, 9, 12, 15, 80

↓ ↓ ↓

$Q_1$ $Q_2$ $Q_3$.

as, 1st quantile = 7

2nd quantile = 9

3rd quantile = 15.

smallest = 4

highest = 80.

Inter-Quantile Range

$IQR = Q_3 - Q_1 = 15 - 7$

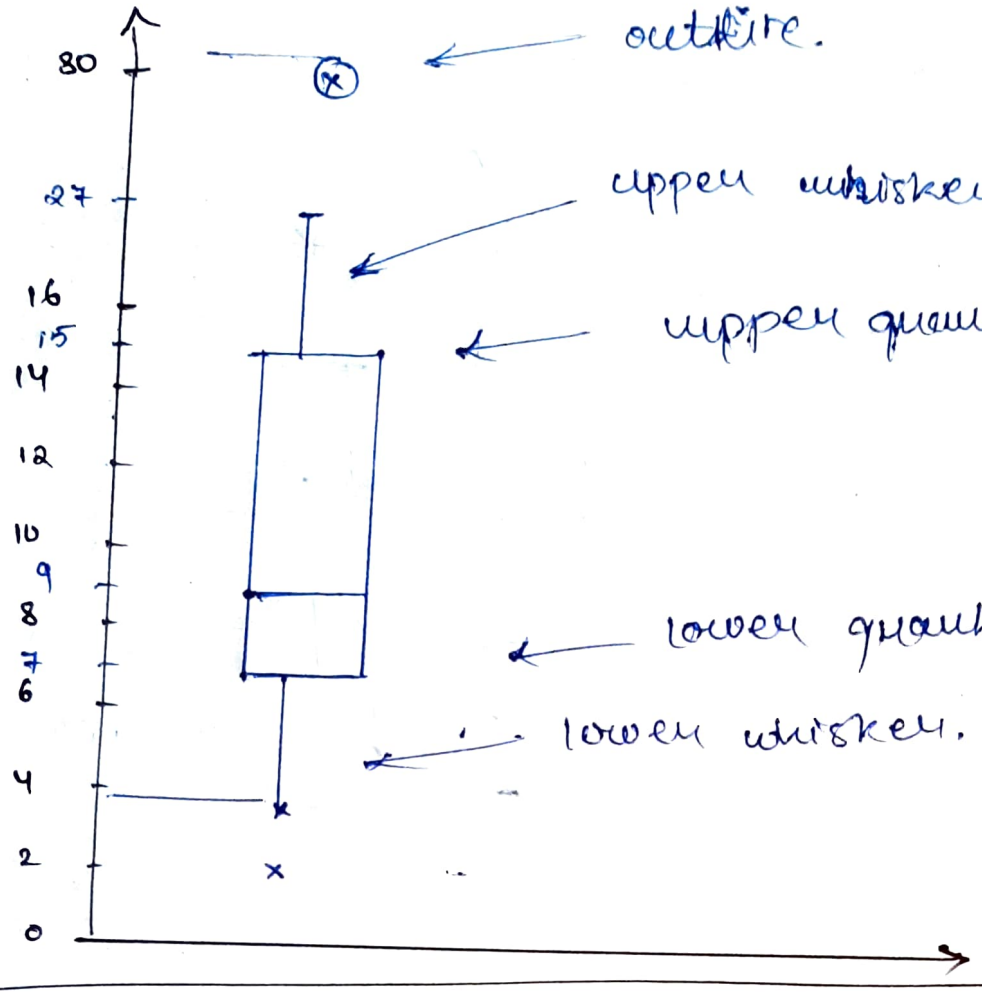
$\boxed{IQR = 8}$

lower whisker $= Q_1 - 1.5 * IQR = 7 - 3/2 * 8$

$= \underline{5}$

upper whisker $= Q_3 + 1.5 * IQR = 15 + 3/2 * 8$

$= \underline{27}$.

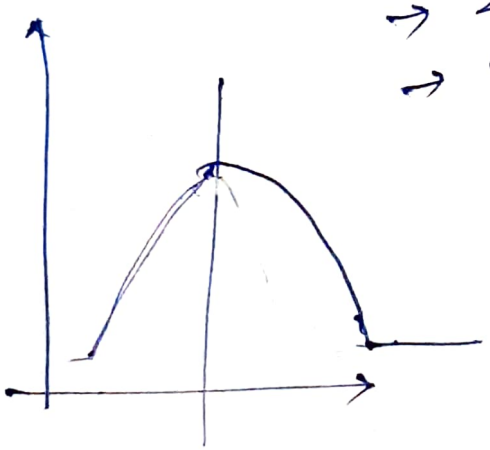



Data distribution → 25% (15 and 27)
→ 25% of data b/w 4 – 7
→ 50% of data b/w 7 and 15

Mean.
$= \frac{4+7+9+8+12+80+15}{7} = 19.28.$

Median = 9

(ii)



as, Mean > Median

80,
the data is positively-skewed or Right-skewed.

$Q_1 = 7$
$Q_2 = 9$.
$Q_3 = 15$
and IQR = 8.
but max = 80 and min = 4.
so, data have outlier point.

## Question 3

(D) A → (iv) ; B → (iii) ; C → (ii) ; D → (i)

---

## Question 4

(i) Binning methods

(a) unsupervised Binning
   1) Equal width binning
   2) Frequency width binning

(b) supervised Binning
   1) Entropy based Binning.

---

(ii)

value : 15, 21, 45, 6, 11, 17, 45, 19, 12, 4, 9, 5

sorting : 4, 5, 6, 9, 11, 12, 15, 17, 19, 21, 45, 45.

(a) Apply Equal width binning :

Equal width bining.

len(value) = 12

$w = (max - min)/3 = \frac{45-4}{3} = 13.67$

bin 1 - range (4, 17.67)

bin 2 - range (17.67, 31.33]

bin 3 - range (31.33, 45).

Bin 1 : [4, 5, 6, 9, 11, 12, 15, 17]

Bin 2 : [19, 21]

Bin 3 : [45, 45]

(ii) Binning by Frequency (equal frequency)

Bin size = 3

Bin-1 [4, 5, 6]

Bin 2 - [9, 11, 12]

Bin 3 - [15, 17, 19]

Bin-4 [21, 45, 45]

---

## Question (5)

(i)

As we know, feature scaling, ~~because~~ require when data is missing, to remove outlire, or pruned redundant raws

Yes, the feature "Age" in the data set requires feature scaling, Because the other dataset are 25, 35, 30, 20 with difference of min 5 & max 15.

but the age value = 1000.

the difference between the values are much higher which is outlier point in the data.

- we need to make the data in same scale. so, that each feature is equally important.

---

(ii) using person correlation

Speed (S)

$$\bar{S} = \frac{\Sigma S}{n} = \frac{100+120+140+160+180}{5} = 140.$$

$$\sigma_S = \sqrt{\frac{\Sigma(\bar{S}-S)^2}{n}} = \sqrt{\frac{(40)^2+20^2+20^2+40^2}{5}}$$

$$= 10\sqrt{\frac{40}{5}} = \boxed{\sigma_2 = 20\sqrt{2}}$$

Acceleration (a)

$$\bar{a} = \frac{\Sigma a}{n} = \frac{60+70+75+85+100}{5} = \frac{390}{5} = 78.$$

$$\sigma_a = \sqrt{\frac{18^2+8^2+3^2+7^2+22^2}{5}} = \sqrt{\frac{930}{5}}$$

$$= \boxed{\sqrt{186} = \sigma_a}$$

Covariance (S, a)

$$= \frac{1}{n} \sum_{i}^{n} (s_i - \bar{s}) * (a_i - \bar{a})$$

$$= \frac{1}{5} ((-40)(-18) + (-20)(-8) + 0(3) + 20(7) + 40(22))$$

$$= \frac{1900}{5} = 380 = \text{cov}(S, a)$$

Pearson correlation

$$P_{s,a} = \frac{\text{cov}(s,a)}{\sigma_s \sigma_a} = \frac{380}{20\sqrt{2} \times \sqrt{186}}$$

$$= \frac{380}{20\sqrt{4 \times 93}} = \frac{19}{2\sqrt{93}}.$$

$$= 19/19.2873$$

$$\boxed{P_{S,a} = 0.98510}$$ ≈ nearly 1.

Rough plot

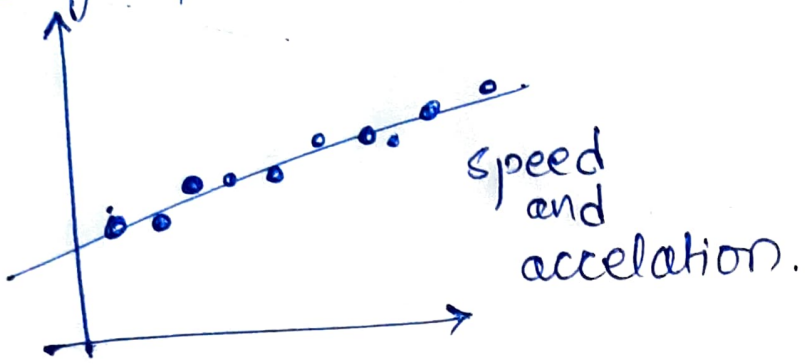
→ Since p is positive, they change in the same direction.

→ since, $p \approx 1$, they are strongly correlated.

Rough plot

106119100

(7)




$\rho \approx 1$, so most point on the line