DWM
ASSIGNMENT - 1
PRE-PROCESSING

JUNAID · GIRKAR
60004190057
TE COMPS A-4

29/10/21

Q1 Arranged data : 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

a what is the mean of the data? What is the median?

ANS

$$\text{Arithmatic mean } (\bar{x}) = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{13 + 15 + 16 + 16 + 19 + \dots + 46 + 52 + 70}{27}$$

$$= \frac{809}{27}$$

$$\boxed{\bar{x} = 29.96}$$

∵ Number of values in the data are odd

∴ Median $= \left(\frac{n+1}{2}\right)^{th}$ value

$= \left(\frac{27+1}{2}\right)^{th}$

$= 14^{th}$ value

$\boxed{= 25}$

b what is the mode of the data? Comment on the data's modality

ANS

13, 15, 19, 21, 30, 36, 40, 45, 46, 52, 70 occurs once in the dataset

16, 20, 22, 33 occurs twice in the dataset

25, 35 occurs four times in the dataset

∴ The dataset is bimodal with $\boxed{25 \text{ and } 35}$ as the modes

c What is the midrange of the data?

ANS

$$\text{Midrange} = \frac{\text{Max Value} + \text{Min Value}}{2}$$

$$= \frac{70 + 13}{2}$$

$$\boxed{= 41.5}$$

d Can you find the first quartile (Q_1) and the third quartile (Q_3) of the data?

ANS

First Quartile (Q_1) = 25^{th} percentile of data

$$= \frac{25}{100} \times 27$$

$$= 6.75$$

$\approx 7^{th}$ term of dataset

$$\boxed{= 20}$$

Third Quartile (Q_3) = 75^{th} percentile of data

$$= \frac{75}{100} \times 27$$

$$= 20.25$$

$\approx 20^{th}$ term of dataset

$$\boxed{= 35}$$

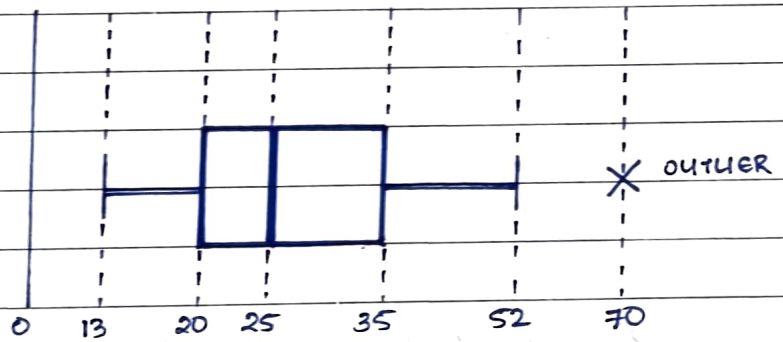
e Give the five-number summary of the data

ANS.

Minimum = 13
Q1 = 20
Median = 25
Q_3 = 35
Maximum = 70

f Show a boxplot of the data

ANS



g How is a quantile-quantile graph different from a quantile plot.

ANS

A quantile plot is a graphical method used to show the approximate percentage of values below or equal to the independent variable in an univariate distribution. Thus, it displays quantile information for all the data, where the values measured for the independent variable are plotted against their corresponding quantile

A quantile-quantile plot graphs the quantile one

unvariate distribution against the corresponding quantiles of another unvariate distribution. Both axes display range of values measured for their corresponding distribution, and points are plotted that correspond to the quantile values of the two distributions.

Q2

AGE	FREQUENCY
1 - 5	200
6 - 15	450
16 - 20	300
21 - 50	1500
51 - 80	700
81 - 110	44

compute an approximate median value for the data

ANS

From the table, $N = 3194$

Median value $= \left(\frac{N}{2}\right)^{th}$ value

$= 1597^{th}$ value

∴ the median lies in the age range of 20.5 to 50.5

$$\text{Median} = L_1 + \left(\frac{N/2 - (\Sigma \text{freq})_1}{\text{freq}_{(\text{median})}}\right) . \text{width}$$

$L_1 = 20.5$

$N = 3194$

$\Sigma(\text{freq})_1 = 200 + 450 + 300 = 950$

$\text{Freq}_{(\text{median})} = 1500$

width = 30

$\therefore$ Median = $20.5 + \left[\frac{1597 - 950}{1500}\right] \times 30$

= 33.44 years

Q3 Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

age	23	23	27	27	39	41	47	49	50
% fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2

age	52	54	54	56	57	58	58	60	61
% fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

a Calculate the mean, median and SD of age and % fat

ANS

Mean = $\frac{\sum_{i=1}^{n} x_i}{n}$

Median = middle term after arranging

Standard Deviation (SD) = $\sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$

FOR AGE : 23, 23, 27, 27, 39, 41, 47, 49, 50, 52, 54, 54, 56, 57, 58, 58, 60, 61

mean $= \frac{836}{18}$

$= 46.44$

median $= \frac{50 + 52}{2}$

$= 51$

SD $= \sqrt{\frac{2972.2}{18}}$

$= 12.85$

For fat : 7.8, 9.5, 17.8, 25.9, 26.5, 27.2, 27.4, 28.8, 30.2, 31.2, 31.4, 32.9, 33.4, 34.6, 34.6, 35.7, 41.2, 42.5

mean $= \frac{518}{18}$

$= 28.78$

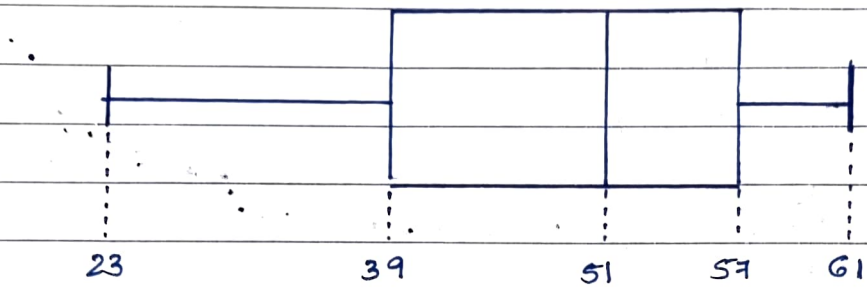
median $= \frac{30.2 + 31.2}{2}$

$= 30.7$

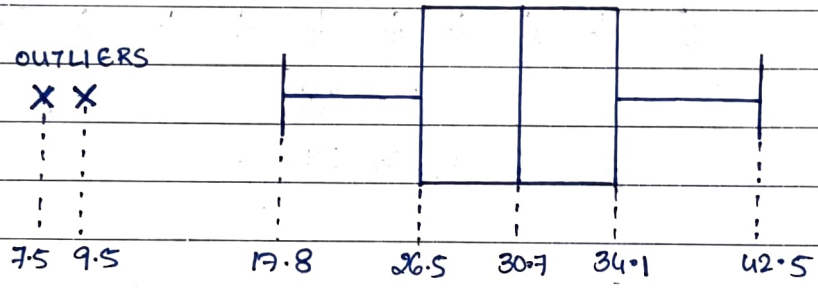
SD $= \sqrt{\frac{1454.76}{18}}$

$= 8.99$

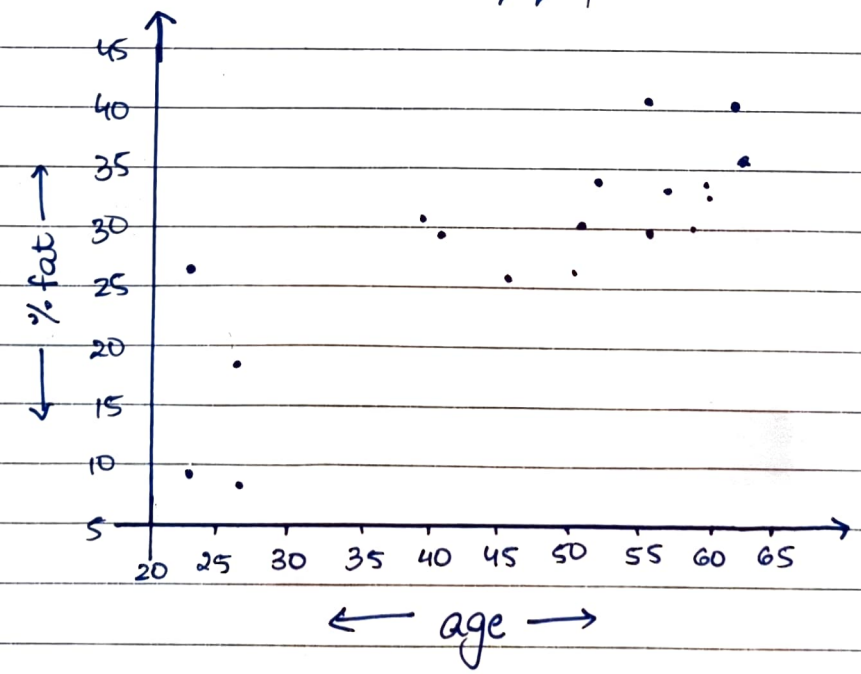
b) age : Min = 23, Q_1 = 39, Median = 51, Q_3 = 57, Max = 61



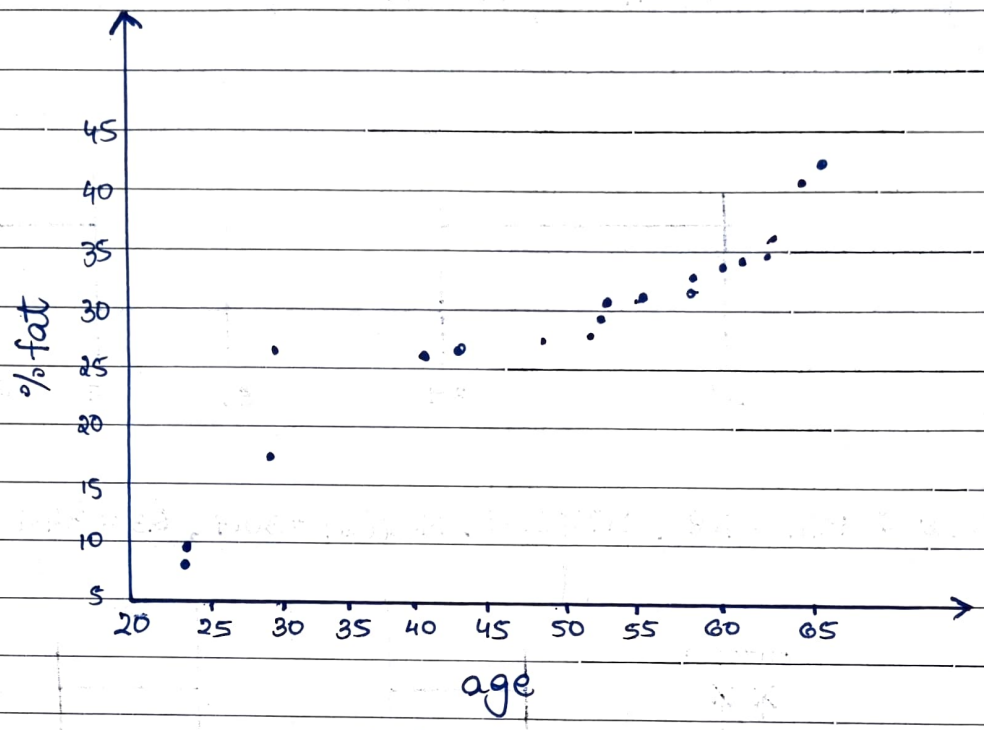
%fat : Min = 17.8, Q1 = 26.5, Median = 30.7, Q3 = 34.1, Max = 42.5



c) Draw a scatter plot and a q-q plot based on these two variables

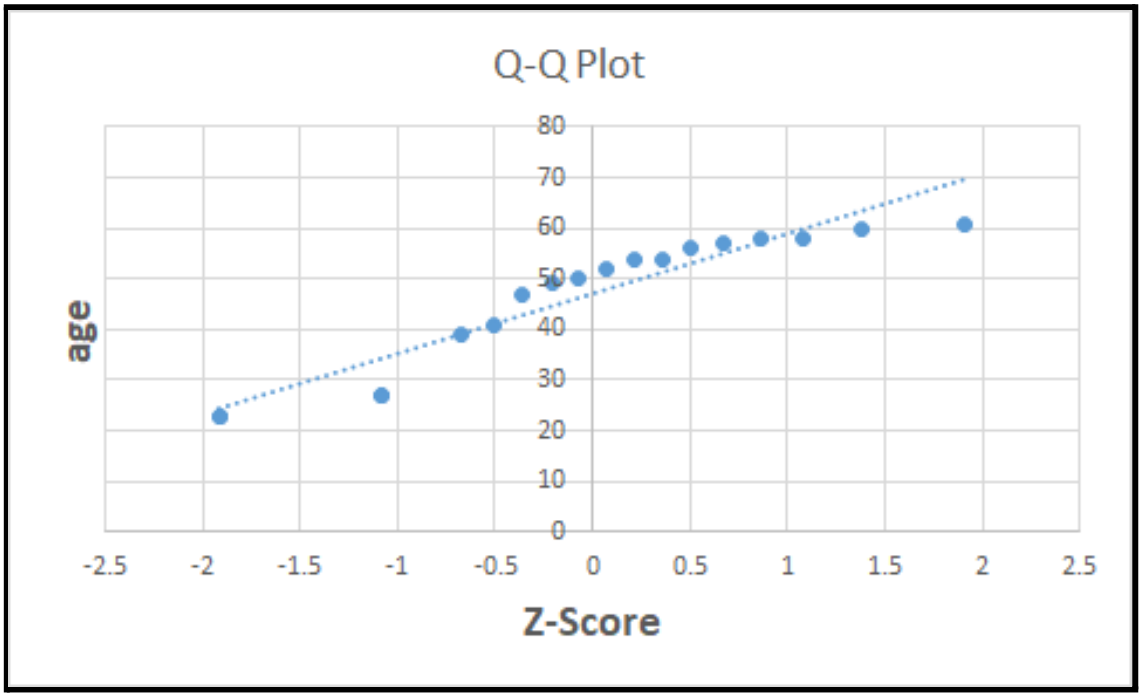


SCATTER PLOT

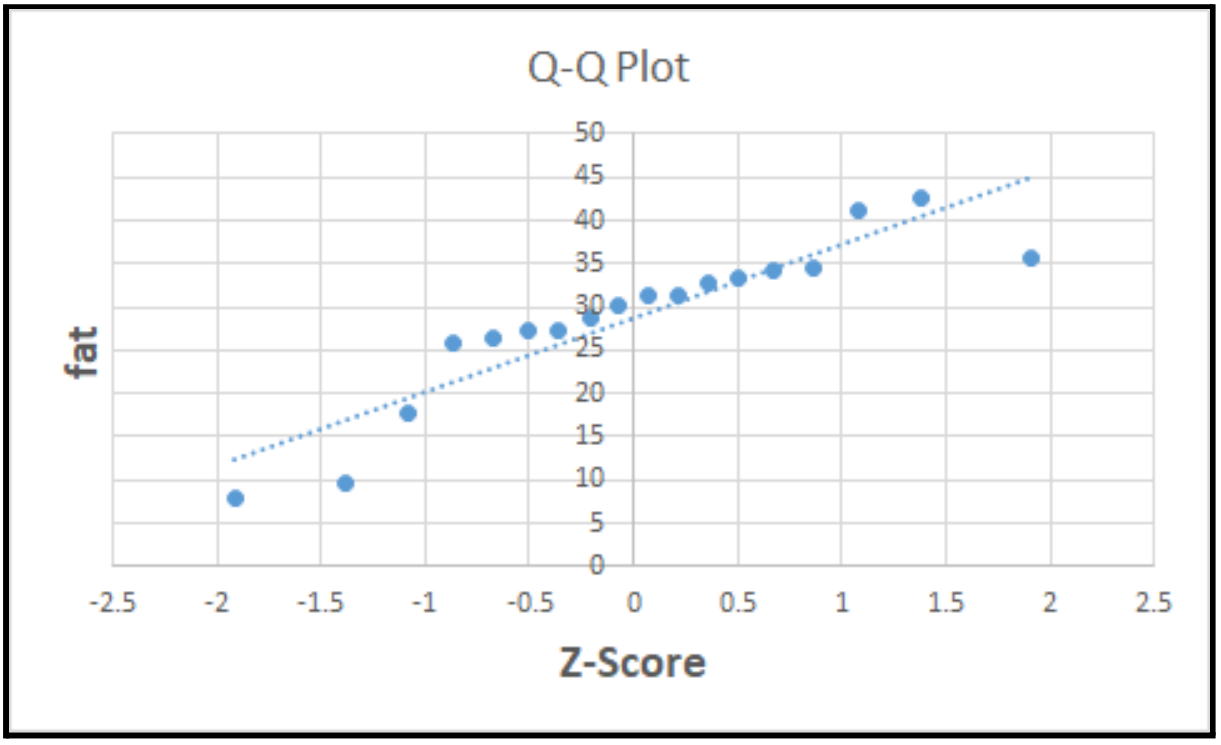


Q-Q PLOT.

Data	Rank	Percentile	Z-Score	Data
23	1	0.027778	-1.9145	23
23	1	0.027778	-1.9145	23
27	3	0.138889	-1.0853	27
27	3	0.138889	-1.0853	27
39	5	0.25	-0.6745	39
41	6	0.305556	-0.5085	41
47	7	0.361111	-0.3555	47
49	8	0.416667	-0.2104	49
50	9	0.472222	-0.0697	50
52	10	0.527778	0.06968	52
54	11	0.583333	0.21043	54
54	12	0.638889	0.35549	54
56	13	0.694444	0.50849	56
57	14	0.75	0.67449	57
58	15	0.805556	0.86163	58
58	16	0.861111	1.08532	58
60	17	0.916667	1.38299	60
61	18	0.972222	1.91451	61



Data	Rank	Percentile	Z-Score	Data
7.8	1	0.027778	-1.9145	7.8
9.5	2	0.083333	-1.383	9.5
17.8	3	0.138889	-1.0853	17.8
25.9	4	0.194444	-0.8616	25.9
26.5	5	0.25	-0.6745	26.5
27.2	6	0.305556	-0.5085	27.2
27.4	7	0.361111	-0.3555	27.4
28.8	8	0.416667	-0.2104	28.8
30.2	9	0.472222	-0.0697	30.2
31.2	10	0.527778	0.06968	31.2
31.4	11	0.583333	0.21043	31.4
32.9	12	0.638889	0.35549	32.9
33.4	13	0.694444	0.50849	33.4
34.1	14	0.75	0.67449	34.1
34.6	15	0.805556	0.86163	34.6
35.7	18	0.972222	1.91451	35.7
41.2	16	0.861111	1.08532	41.2
42.5	17	0.916667	1.38299	42.5



5

Q.4 Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8)

a Compute the Euclidean distance between the two objects

ANS

$$\text{Euclidean distance} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} + x_{j2})^2 + \dots + (x_{in} - x_{jn})^2}$$

$$= \sqrt{(22-20)^2 + (1-0)^2 + (42-36)^2 + (10-8)^2}$$

$$= \sqrt{45}$$

$$= 6.708$$

c Compute the Minkowski ~~Manhattan~~ distance between the two objects

ANS

$$\text{Minkowski distance} = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \dots + |x_{in} - x_{jn}|^h}$$

where $h \geq 1$

Here, $h = 3$

$$= \sqrt[3]{|22-20|^3 + |1-0|^3 + |42-36|^3 + |10-8|^3}$$

$$= \sqrt[3]{233}$$

$$= 6.153$$

b Compute the Manhattan distance between the two objects

ANS

$$\text{Manhattan distance} = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{in} - x_{jn}|$$

$$= |22-20| + |1-0| + |42-36| + |10-8|$$

$$= 11$$

d Compute the supremum distance between the two objects.

ANS

$$\text{Supremum distance} = \lim_{n\to\infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^2 \right)^{1/n}$$

$$= \max_{f}^{p} |x_{if} - x_{jf}|$$

$$= \max(2, 1, 6, 2)$$

$$= 6$$

Q5 Suppose we have the following 2-D data set

	A_1	A_2
x_1	1.5	1.7
x_2	2	1.9
x_3	1.6	1.8
x_4	1.2	1.5
x_5	1.5	1.0

a Consider the data as 2-D data points. Given a new data point, $x = (1.4, 1.6)$ as a query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, supremum distance and cosine similarity.

ANS

$$\text{Euclidean distance} = \sqrt{(x_{i1} - x_{j1})^2 + \dots + (x_{in} - x_{jn})^2}$$

$$\text{Manhattan distance} = |x_{i1} - x_{j1}| + \dots + |x_{in} - x_{jn}|$$

$$\text{Supremum distance} = \max_{f}^{n} |x_{if} - x_{jf}|$$

cosine similarity $= \frac{x^t \cdot y}{\|x\| \|y\|}$

x^t = transposition of vector x

$\|x\|$ = Euclidean norm of vector x

$\|y\|$ = Euclidean norm of vector y

From points no (1.4, 1.6) , we get

	EUCLIDEAN	MANHATTAN	SUPREMUM	COSINE SIMILARITY
x_1	0.1414	0.2	0.1	0.99999
x_2	0.6708	0.9	0.6	0.99535
x_3	0.2828	0.4	0.2	0.99997
x_4	0.2236	0.3	0.2	0.99903
x_5	0.6083	0.7	0.6	0.96536

RANKINGS :

Euclidean : x_1, x_4, x_3, x_5, x_2

Manhattan : x_1, x_4, x_3, x_5, x_2

Supremum : x_1, x_4, x_3, x_5, x_2

Cosine Similarity : x_1, x_3, x_4, x_2, x_5

b Normalize the data set to make the norm of each data point equal to 1. Use Euclidean distance on the transformed data to rank the data points.

ANS The normalized query is (0.65850, 0.75258).

The normalized dataset is given by the following table :

	A_1	A_2
x_1	0.66162	0.74984
x_2	0.72500	0.68875
x_3	0.66436	0.74741
x_4	0.62470	0.78087
x_5	0.83205	0.55470

Recomputing Euclidean distances before yields.

	Euclidean Distance
x_1	0.00415
x_2	0.09217
x_3	0.00781
x_4	0.04409
x_5	0.26320

$\therefore$ final ranking : x_1, x_3, x_4, x_2, x_5 //