Assignment 1

Shashwat Shah
60004220126

Pre-Processing

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

a) Mean & Median

$$\therefore \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{13 + 15 + \ldots 46 + 52 + 70}{27}$$

$$= 809 / 27 \qquad \boxed{\bar{x} = 29.96}$$

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

$$= \frac{27+1}{2}$$

$$= \frac{28}{2} = 14\text{th value}$$

$$\boxed{\text{Median} = 25}$$

b) Mode (Also comment on modality)

13, 15, 19, 21, 30, 36, 40, 45, 46, 52, 70 occur 1 time
16, 20, 22, 33 occur 2 times
25, 35 occur 4 times, Hence the dataset is bimodal with 25 & 35 as modes

c) Midrange

$$= \frac{\text{Max} - \text{Min}}{2}$$

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

$$= 41.5$$

d) Find Q_1 & Q_3

Q_1 (First Quartile) = 25% of data

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

$$= 6.75$$

$\simeq$ 7th term

$Q_1 = 20$

Q_3 (Third Quartile = 75% of data

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

$$= 20.25$$

$\simeq$ 20th term

$Q_3 = 35$

e) Five Number summary

Min = 13

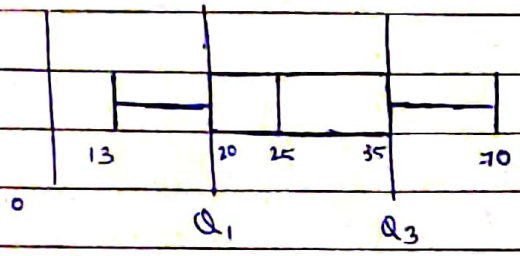
Q_1 = 20

Median = 25

Q_3 = 35

Max = 70

f) Boxplot



g) How is quantile quantile graph different from quantile plot

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 variables are plotted against their corresponding quantile

A quantile quantile plot graphs the quantiles on univariate distribution against the corresponding quantiles of another univariate distribution. Both axes display range of values measured for their variable corresponding distribution and points are plotted that correspond to the quantile value 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.

From the table $N = 3194$

Median $= \frac{N}{2} = 1597^{th}$ value $\therefore$ Median $= 20.5 - 50.5$

$$\text{Median} = L_1 + \left(\frac{N/2 - (\Sigma f_{rea})}{f_{rea}(\text{median})} \right) \cdot \text{width}$$

$L_1 = 20.5$

$N = 3194$

$\Sigma (f_{rea}) = 200 + 456 + 300$

Frea median $\sim$ 1500

width $= 30$

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

$$= \underline{33.44 \text{ years}}$$

3

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 mean, median, SD. of age & fat

$$\text{Mean for age} = \frac{836}{18} = 46.44$$

$$\text{Median} = \frac{50 + 52}{2} = 51$$

$$SD = \sqrt{\frac{\Sigma (x_1 - \bar{x})^2}{n}} = \sqrt{\frac{2972.2}{18}} = 12.85$$

For fat

$$\text{Mean} = \frac{518}{18}$$

$$= 28.78$$

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

$$= 30.7$$

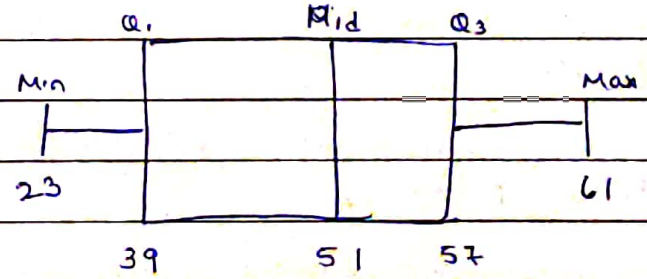
$$SD = \sqrt{\frac{1456.76}{18}}$$

$$= 8.99$$

b) Boxplot :

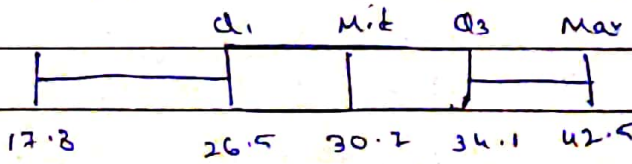
Max : 61

Age : min = 23 Q_1 = 39 Med = 51 Q_3 = 57



max = 42.5

Fat : min = 17.8 Q_1 = 26.5 median = 30.7 Q_3 = 34.1



4 (22, 1, 42, 10) & (20, 0, 36, 8)

∴ Euclidean Distance

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

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

$$= \sqrt{45}$$

$$= 6.708$$

∴ Minowski distance

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

h = 3 given

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

$$= \sqrt{8 + 1 + 216 + 8}$$

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

$$= 6.153$$

∴ Manhattan Distance

$$= |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots$$

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

$$= 2 + 1 + 6 + 2$$

$$= 11$$

∴ Supremum Distance

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

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

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

$$= 6$$

5 Suppose we have 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

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.

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

x^t - transpose of x

$\|x\|$ - Euclidean norm

$\|y\|$: Euclidean norm

From points (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	0.99575
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

∴ Ranks.

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 : x_1, x_3, x_4, x_2, x_5

→ Normalise the data set

∴	A1	A2
x_1	0.661682	0.74984
x_2	0.72500	0.68875
x_3	0.66436	0.74741
x_4	0.62470	0.78087
x_5	0.83250	0.55470

∴ Recompute Euclidean

∴ Rank

x_1, x_3, x_4, x_2, x_5

	Eud Dist
x_1	0.00415
x_2	0.09217
x_3	0.00781
x_4	0.04409
x_5	0.26320