DWM

Name: Ayush Jain
SAP ID: 60004200132
TE Comps B 2

Assignment 1 - Preprocessing

Q. 1) Suppose that the data for analysis includes the attributes age. The age values for the data tuples are:
13, 15, 16, 16, 19, 20, 20, 21, 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?

→

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

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

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

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

∵ Number of values in the data are odd

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

$$= \left(\frac{27+1}{2}\right)^{th} = 14 \text{ value}$$

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

(b) What is the mode of the data? Comment on the data's modality.

→ 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 time in the dataset.
∴ The dataset is bimodal with 25, 35 as the modes.

(c) What is the midrange of the data?

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

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

$= 41.5$

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

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

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

$= 6.75$

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

$= 20$

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

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

$= 20.25$

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

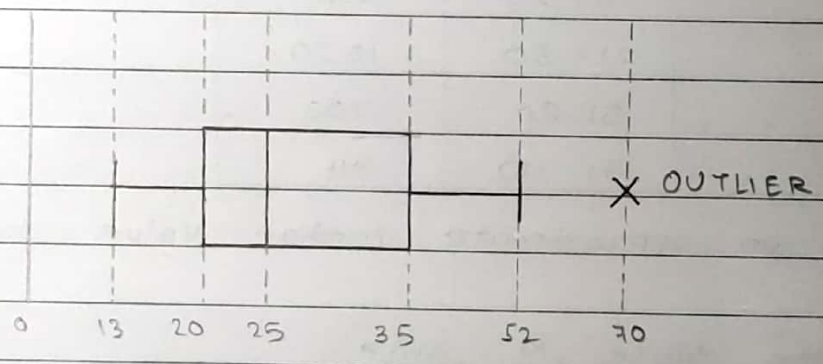
$= 35$

(e) Give the five-number summary of the data?

→ Minimum = 13
Q_1 = 20
Median = 25
Q_3 = 35
Maximum = 70

(f) Show a boxplot of the data

→



(g) How is a quantile-quantile plot different from a 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 variable are plotted against their corresponding quantile.

A quantile-quantile plot graphs the quantile one

univariate distribution against the corresponding quantiles of another univariate 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.

Q. 2)

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

$= 1597^{th}$ value

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

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

$L_1 = 20.5$

$N = 3194$

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

freq (median) = 7500

width = 30

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

$$= 33.44 \text{ years}$$

Q. 3) Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following units.

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 standard deviation of age and % fat

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

Median = middle term after arranging

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

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 $= \sqrt{\frac{2972.2}{18}} = 12.85$

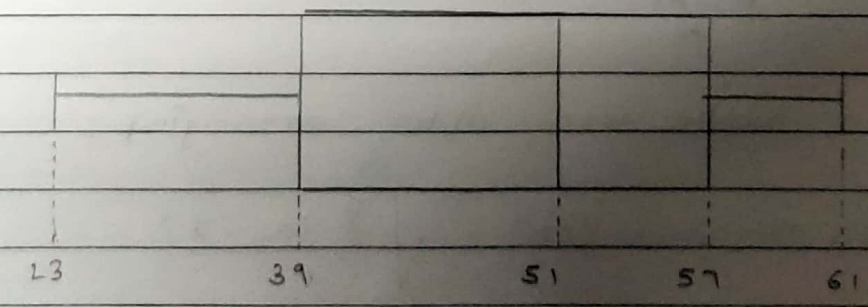
For fat : 7.8, 9.5, 17.8, 25.9, 27.2, 27.4, 28.8, 30.2, 31.2, 31.4, 32.9, 33.4, 34.1, 34.6, 35.7, 41.2, 42.5

Mean $= \frac{518}{18} = 28.78$

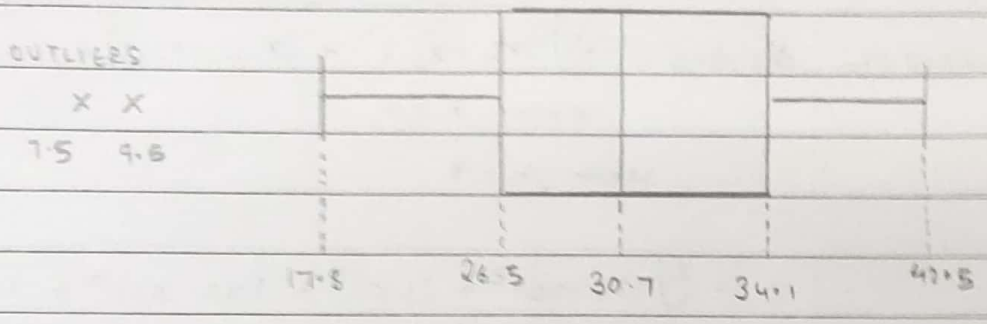
Median $= \frac{30.2 + 31.2}{2} = 30.7$

SD $= \sqrt{\frac{1454.76}{18}} = 8.99$

b) age : Min = 23, Q1 = 39, Median = 51, Q3 = 57, Max = 61



% fat = Min = 17.8, Q_1 = 26.5, Median = 30.7, Q_3 = 34.1, Max = 42.5



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.

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

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

$= \sqrt{45}$

$= 6.708$

(b) Compute the Manhattan distance between the two objects.

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

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

$= 11$

(c) Compute the Minkowski distance between the two objects.

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

where $h \geq 1$

Here, $h = 3$

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

$= \sqrt[3]{233}$

$= 6.153$

(d) Compute the supremum distance between two objects.

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

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

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

$= 6$

Q. 5) Suppose we have following 2-D data sets.

	A1	A2
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) Euclidean distance $= \sqrt{(x_{i1} - x_{j1})^2 + \dots + (x_{in} - x_{jn})^2}$

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

Supremum distance $= \max_f |x_{if} - x_{jf}|$

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

where 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	0.99575
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

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) 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

∴ final ranking : x_1, x_3, x_4, x_2, x_5