Experiment No: 03	TE AI&DS
Date of Performance:	Roll No: 9696
Aim: Apply data Exploration techniques on given data to organize data (Tutorial)	
CO3: Apply data exploration and Data preprocessing techniques to organize and prepare data for data mining	

Rubrics for assessment of Experiment:

Sr. No	Parameters	Exceed Expectations(EE)	Meet Expectations (ME)	Below Expectations (BE)
1	Timeline (2)	Early or on time (2)	One session late (1)	More than one session late (0)
2	Preparedness (2)	Knows the basic theory related to the experiment very well. (2)	Managed to explain the theory related to the experiment. (1)	Not aware of the theory to the point. (1)
3	Effort (3)	Done expt on their own. (3)	Done expt with help from other. (2)	Just managed. (1)
4	Documentation(2)	Lab experiment is documented in proper format and maintained neatly. (2)	Documented in proper format but some formatting guidelines are missed. (1)	Experiments not written in proper format (0.5)
5	Result (1)	Specific conclusion.(1)	Partially specific conclusion. (0.5)	Not specific at all. (0)

Assessment Marks:

Timeline(2)	Preparedness(2)	Effort(3)	Documentation(2)	Result(1)	Total(10)

NAME:- CHAVARATTIL SHIVPRASAD PREMARAJAN.
ROLLNO: 9696.

Tutorial 1. DWM Experiment 3.

1]

→ 18, 20, 21, 21, 24, 25, 25, 26, 27, 27, 29, 29, 29, 29, 35, 38, 38, 40, 40, 40, 40, 41, 45, 50, 51, 57, 75.

a) mode = 33.15

median = 29.

b)

Modes are 29 & 40.

∴ Data's modality = bimodal.

c) mid range = (18 + 75)/2.
mid range = 46.5

d)

→ Since the data is already sorted

$Q_1 = (27+1)/4 = 7$ th value

$Q_3 = 3 * (27+1)/4 = 21$ st value.

$Q_1 = 25$
$Q_3 = 40$

c) minimum = 18 $\quad Q_2 = \frac{n \times 50}{100} = \frac{27 \times 1}{2} = 13.5 \approx 14^{th}$

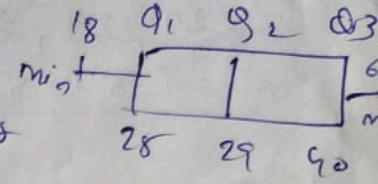
$Q_1 = 25$ $\quad Q_3 = 40$ $\quad IQR = Q_3 - Q_1 = 40 - 25 = 15$

Median = 29 $\quad$ Lower limit $= Q_1 - 1.5 \times IQR$

$Q_3 = 40$ $\quad = 25 - 1.5 \times 15 = 2.5$

Maximum = 75 $\quad$ Upper limit $= Q_3 + 1.5 \times IQR$

$= 40 + 1.5 \times 15 = 62.5$



9) A quantile-quantile (Q-Q) plot is used to compare the quantiles of data distribution with quantiles of a theoretical distribution like a normal distribution. It helps to determine if the data follows a specific distribution. A quatile plot on the other hand, typically refers to a plot of the quantiles of data itself.

2].

→

age	frequency	c.f.
1 - 5	300	300
6 - 15	550	850
16 - 20	450	1300
21 - 50	1200	2500
51 - 80	800	3300
81 - 110	65	3365

$n = 110$, $n/2 = 110/2 = 55$.

55 th item lies in the class of 51 - 80.

$\therefore$ $L_1 = 51$.

$n = 110$.

$(\Sigma f_{req})_L = 2500$

$\Sigma (f_{req})_m = 800$.

$$\text{median} = L_1 + \left(\frac{n/2 - \Sigma (f_{req})_L}{\Sigma (f_{req})_m} \right) \times \text{width}$$

$$= 51 + \left(\frac{55 - \cancel{800}\ 2500}{800} \right) \times 29$$

$$= 51 - 29.07 = 21.93$$

3]

→ a)

mean = (23 + 23 + 27 + 27 + 39 + 41 + 47 + 49 + 50 + 52 + 54 + 54 + 56 + 57 + 58 + 58 + 60 + 61) / 18

= 45.06.

Median = 50.

standard deviation = $\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$.

$= 2(23-45.06)^2 + 2(27-45.06)^2 + (39-45.06)^2 + (41-45.06)^2 + (47-45.06)^2 + (49-45.06)^2 + (50-45.06)^2 + (52-45.06)^2 + 2(54-45.06)^2 + (56-45.06)^2 + (57-45.06)^2.$

Standard deviation = $\sqrt{\frac{2970.44}{18}} = 12.84$.

~~Ans~~ Fact:

Mean = $\frac{516.1}{18} = 28.67$.

median :

$N = 18 \Rightarrow \frac{N}{2} = \frac{18}{2} = 9$

Median :- ~~N = 18~~ $\frac{(9^{th} + 10^{th})}{2} = \frac{30.2 + 33.6}{2} = 31.9$

fat	fat − 2867	(fat − 2867)²
7.5	−21.17	448.16
24.5	−4.17	17.38
5.8	−22.87	523.03
15.8	−12.87	165.62
33.4	4.73	22.37
28.9	−4.77	22.75
31.4	2.73	7.45
29.2	0.53	0.2809
30.2	1.53	2.3409
33.6	4.93	24.3049
44.5	15.83	250.58
28.8	0.13	0.0169
31.4	2.73	7.4529
33.2	4.53	20.52
32.1	3.43	11.76
34.9	6.23	38.81
40.2	11.53	132.94
35.7	7.1	50.41

$\Sigma \Rightarrow 1746.1827$

$\sigma = \sqrt{\frac{1746.1827}{18}}$

$= 9.84$

b) box plot

for age

$Q_1 = \frac{N \times 25}{100} = \frac{18 \times 25}{100} = 4.5 \approx 5^{th} = 39$

$Q_2 = \frac{N \times 50}{100} = 9 = \frac{50+52}{2} = 51$

$Q_3 = \frac{N \times 75}{100} = 13.5 = 57$

$IQR = Q_3 - Q_1 = 57 - 39 = 18$

Lower limit $= Q_1 - IQR \times 1.5$
$= 39 - 18 \times 1.5$
$= 12$

Upper limit $= Q_3 + IQR \times 1.5$
$= 57 + 18 \times 1.5$
$= 84$

No outlier

for fat:-

5.8, 7.5, 15.8, 23.9, 24.5, 28.8, 29.2, 30.2, 31.4, 31.4, 32.1, 33.2, 33.4336, 34.9, 35.7, 40.2, 49.5

$Q_1 = 4.5 = 5^{th} = 24.5$

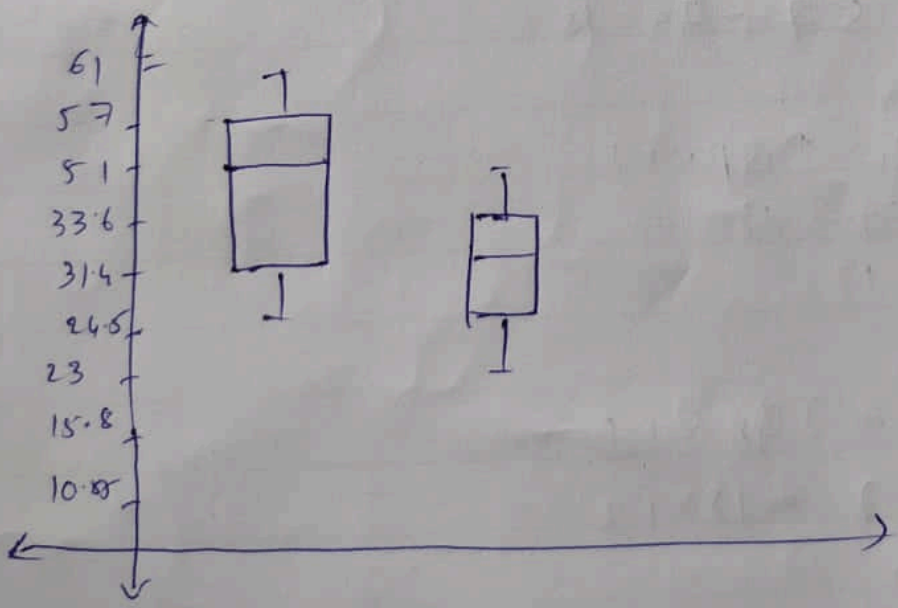
$Q_2 = \frac{9^{th} + 10^{th}}{2} = \frac{31.4 + 31.4}{2} = 31.4$

$Q_3 = 14^{th} = 33.6$

$IQR = 33.6 - 24.5$
$= 9.1$

Lower Limit $= 24.5 - (9.1 \times 1.5)$
$= 10.85$

upper limit $= 33.6 + (9.1 \times 1.5)$
$= 47.25$



4)

→

Age	Ward	Gender	Test	Health Progress	Fees
56	Ward A	F	P	Good	13674
76	Ward B	M	N	Better	12343
23	Ward B	M	N	Better	6542
47	Ward C	F	N	Best	3459

a) Nomical attribute :- Ward.

$$d(P, j) = \frac{P - M}{P}$$

$$d(2,1) = \frac{1-0}{1} = 1$$

$$d(3,1) = \frac{1-0}{1} = 1$$

$$d(3,2) = \frac{1-1}{1} = 0$$

$$d(4,1) = \frac{1-1}{1} = 0$$

$$d(4,2) = \frac{1-0}{1} = 1$$

$$d(4,3) = \frac{1-0}{1} = 1$$

$$d(i,j) = \begin{bmatrix} 0 & & & \\ 1 & 0 & 0 & \\ 1 & 0 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

11) Asymmetric Test

1 P → 1
2 N → 0
3 N → 0
4 N → 0

		obj 1	0
P	1	q	r
	0	s	t

$$d(i,j) = \frac{r+s}{q+r+s}$$

$$d(2,1) = \frac{1+0}{0+1+0} = 1$$

$$d(3,1) = \frac{0+1}{0+0+1} = 1$$

$$d(3,2) = \frac{0}{0} = 0$$

$$d(4,1) = \frac{0+1}{0+0+1} = 1$$

$$d(4,2) = \frac{0}{0} = 0$$

$$d(4,3) = \frac{0}{0} = 0$$

$$d(i,j) = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 0 & 0 & \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

→ more similarity.

Symmetric.

Obj i \ Obj j	1	0
1	q	r
0	s	t

F → 0
M → P.

gender
F → 0
M → 1
N → 1
P → 0.

$$d(2,1) = \frac{1+0}{0+1+0+0} = 1$$

$d(3,1) = 1$
$d(3,2) = 0$
$d(4,1) = 0$
$d(4,2) = 1$
$d(4,3) = 1$

$$d(i,j) = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 0 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Ordinal attribute

Health progress

Good → 3

Better → 2

Better → 2

Best → 1

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

$$Best = \frac{1-1}{3-1} = 0$$

$$Better = \frac{2-1}{3-1} = \frac{1}{2} = 0.5$$

$$Good = \frac{3-1}{3-1} = 1$$

$d(2,1) = 0.5$

$d(3,1) = 0.5$

$d(3,2) = 0$

$$d = \begin{bmatrix} 0 & & & \\ 0.5 & 0 & & \\ 0.5 & 0 & 0 & \\ 1 & 0.5 & 0.5 & 0 \end{bmatrix}$$

Numeric attributes.

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}|$$

$$d(2,1) = |12343 - 13674| + |76 - 50|$$
$$= 1331 + 20$$
$$= 1357.$$

$$d(3,1) = |6542 - 13674| + |23 - 566|$$

$$= 7132 + 33$$
$$= 7165$$

$$d(3,2) = |6542 - 12343| + |23 - 76|$$
$$= 7132 + 33$$
$$= 7165$$

$d(3,2) = 5854$

$d(4,1) = 10224$

$d(4,2) = 8913$

$d(4,3) = 3107$

$$d(i,j) = \begin{bmatrix} 0 & & & \\ 1357 & 0 & & \\ 7165 & 5854 & 0 & \\ 10224 & 8913 & 3107 & 0 \end{bmatrix}$$

Engineer.

5]

→

a)

Euclidean Distance = $\sqrt{(32-20)^2 + (10-0)^2 + (40-31)^2 + (20-5)^2}$

$= 27.037$.

b) Manhattan Distance = $|32-20| + |10-0| + |40-31| + |20-5|$

$= 46$ //.