Shubham Bhate 8318 TE Comps

DWM Assignment 2

1. P_1 (6,3) , P_2 (2,2), P_3 (3,4)

a) Manhattan distance

$$dist = \left[\sum_{k=1}^{n} |p_k - q_k|^r \right]^{1/r} \quad [\text{Put } r=1]$$

L1	P_1	P_2	P_3
P_1	0	5	4
P_2	5	0	3
P_3	4	3	0

b) Euclidean distance

Put $r = 2$ in formula

L2	P_1	P_2	P_3
P_1	0	4.123	3.162
P_2	4.123	0	2.236
P_3	3.162	2.236	0

c] Supremum distance

Put $r = \infty$ in formula

L_∞	P_1	P_2	P_3
P_1	0	4	3
P_2	4	0	2
P_3	3	2	0

2. Avg age = 26.6

Std. age = 11.9499

Avg income = 22600

Std income = 16697.305

Avg edu = 12.2

Std.income = 5.0695

Avg height = 158

Std. height = 19.235

Age-Avg	Income - Avg	Edu - avg	Height - avg
-16.6	-22600	-8.2	-28
-6.6	-7600	0.8	22
1.4	-2600	0.8	2
8.4	17400	5.8	-8
13.4	15400	0.8	12

corr (Age, income) = [(-16.6 * -22600) + (-6.6 * -7600) + (1.4 * -2600) + (8.4 * 17400) + (13.4 * 15400)] 14 * 11.94989 * 16697.305 = 0.97

correlation	Age	Income	Education	Height
Age	1	0.97	0.79	0.45
Income	0.97	1	0.86	0.39
Education	0.79	0.86	1	0.54
Height	0.45	0.39	0.54	1

3. D1 = 420201

D2 = 200022

D1 . D2 = 10

$||D_1|| = (4^2 + 2^2 + 1^2)^{0.5} = 4.58$

$||D_2|| = (2^2 + 2^2 + 2^2)^{0.5} = 3.46$

$$\cos(D_1, D_2) = \frac{(D_1 . D_2)}{|D_1| \times |D_2|} = \frac{10}{4.58 \times 3.46} = 0.63$$

4. p = 001101

q = 111101

$M_{01} = 2$ (p = 0, q = 1)

$M_{10} = 0$ (p = 1, q = 0)

$M_{00} = 1$ (p = 0, q = 0)

$M_{11} = 3$ (p = 1, q = 1)

$$smc = \frac{(M_{11} + M_{00})}{(M_{01} + M_{10} + M_{11} + M_{00})}$$

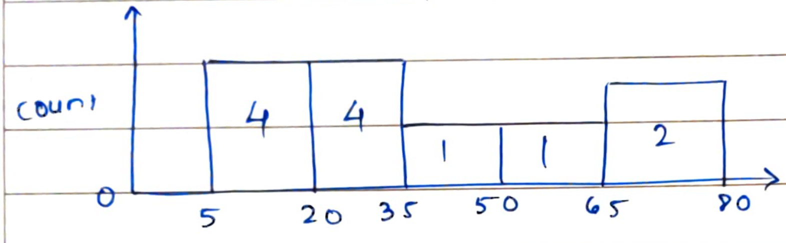
$$= \frac{(3+1)}{(2+0+3+1)} = 0.67$$

$$J = \frac{(M_{11})}{(M_{01} + M_{10} + M_{11})}$$

$$J = \frac{3}{2+3} = 0.6$$

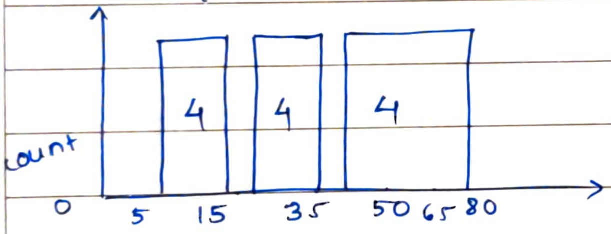
5. a] delta $= \dfrac{(\max - \min)}{x} = 15$

$C_1 = [5, 20)$, $C_2 = [20, 35)$, $C_3 = [35, 50)$
$C_4 = [50, 65)$, $C_5 = [65, 80)$



b) $F = N/K = 12/3 = 4$

$C_1 = \{5, 10, 10, 15\}$, $C_2 = \{20, 28, 30, 30\}$, $C_3 = \{35, 60, 70, 80\}$



6. Min-max normalization $= \dfrac{v - \min_p}{\max_p - \min_p}(\text{newmax}_p - \text{newmin}_p) + \text{new min}_p$

$v = 73600$, $\min_p = 12000$, $\max_p = 98000$, $\text{newmax}_p = 1$
$\text{newmin}_p = 0$

$\therefore$ MMN $= \dfrac{73600 - 12000}{98000 - 12000}(1-0) + 0 = 0.716$

7.

Attribute type	Description	Examples	Operation
① Nominal	values are just different names. Attribute provide enough info to distinguish one object from another	zipcodes, students, teacher, ID no, sex	mode entropy, contingency, corelation, x^2 test.

② Ordinal	the values provide enough info to order objects	grades, age, street no.	medians, percentage, rank corelation, run test, sign test
③ Interval	For interval attributes, the difference betn value are meaningful i.e a unit of measurement exist	class dimensions, calender dates.	mean, standard deviation, Pearson's correlation, t and F tests
④ Ratio	For ratio variables, both differences and ratios are meaningfull	age, height weight and monetary quanties	geometric mean, harmonic mean, percentage, variation

8. AGE

Ascending : 5, 10, 10, 15, 20, 28, 30, 30, 35, 60, 70, 80

$\frac{x}{4} = 3$, min = 5, max = 80

1st Quartile = 12.5 , 2nd Quartile = 29, 3rd Quartile = 47.5

Each quartile has 25% data

spread of 4 quarters are

1st : 12.5 − 5 = 7.5

2nd : 29 − 12.5 = 16.5

3rd : 47.5 − 29 = 18.5

4th : 80 − 47.5 = 32.5

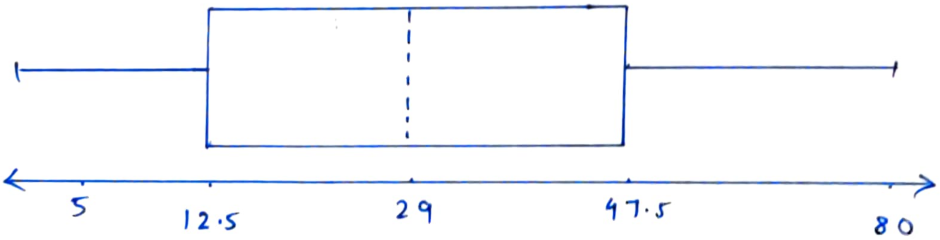
Range = max − min

= 80 − 5 = 75

inter quartile range = 47.5 − 12.5

= 35

Box plot :



9.

Li	I	like	the	Raj	loves	data	mining	than	DBMS
L1	1	1	1	0	0	1	1	1	1
L2	0	0	0	1	1	1	1	1	1

$d_1 = 111001111$

$d_2 = 000111111$

$\therefore d_1 \cdot d_2 = 4$

$|d_1| = 7^{0.5} = 2.646$

$|d_2| = 6^{0.5} = 2.449$

$$\cos(d_1, d_2) = \frac{d_1 . d_2}{|d_1| \times |d_2|} = 0.6172$$

10. We do not consider gender (symmetric attribute)

Let Y and P be 1 and N be 0

$$d\,(\text{Jack}, \text{Mary}) = \frac{0+1}{2+0+1} = 0.33$$

$$d(\text{Jack}, \text{Jim}) = \frac{1+1}{1+1+1} = 0.67$$

$$d\,(\text{Mary}, \text{Jim}) = \frac{1+2}{1+1+2} = 0.75$$

12. student < id, name, addr-id, major, status, univ>

course < id, name, dept>

Instr < id, name, dept>

sem < semid, name, yr >

address < addrid, street, city, country, zipcode >

13. a) setting $min = 0$, $max = 10$, $min_f = 200$, $max_f = 1000$

$$\text{min-max} = \frac{v - min_f}{max_f - min_f}\,(max - min) + min$$

$v = 200$, minmax $= 0$, $v = 400$, minmax $= 0.25$

$v = 300$, minmax $= 0.125$, $v = 500$, minmax $= 0.5$

$v = 1000$, minmax $= 1$

b) $\mu = 1/n \sum^{k} x_i = 500$, $\mu_{AP} = 1/n \sum^{n} (x_i - \mu) = 240$

$$z = \frac{v - \bar{A}}{\sigma A}$$

$v = 200$, $z = -1.25$, $v = 400$, $z = -0.417$, $v = 1000$, $z = 2.0$

$v = 300$, $z = -0.833$, $v = 600$, $z = 0.417$