DMW Assignment 1

Q-1) A) Scale of [0.0, 1.0]

data [200, 300, 400, 600, 1000]

i) Min-max Normalisation:

given data : [200, 300, 400, 600, 1000]

min: 200

max: 1000

v = V is respective value of attribute

$v_1 = 200$, $v_2 = 300$, $v_3 = 400$, $v_4 = 600$, $v_5 = 1000$

new max = 1, new min = 0

$$v' = \left\{ \frac{v - minA}{maxA - minA} \times (new\ maxA - new\ minA) \right\} + new\ minA$$

∴ for

$$min\text{-}max = \left\{ \frac{(v - min\ mark)(new\ max - new\ min)}{max - min} \right\} + new\ min$$

$$= \frac{200 - 200(1-0)}{1000 - 200} + 0$$

$$= 0$$

for 300:

$$min\ max = \frac{300 - 200(1-0)}{1000 - 200} + 0$$

$$= \frac{100}{800}$$

$$= 0.125$$

for 400:

$$\text{min-max} = \frac{400 - 200}{1000 - 200}(1-0) + 0$$

$$= \frac{200}{800}$$

$$= 0.25$$

for 600:

$$\text{min max} = \frac{600 - 200}{1000 - 200}(1-0) + 0$$

$$= \frac{400}{800}$$

$$= 0.5$$

for 1000:

$$\text{min max} = \frac{1000 - 200}{1000 - 200}(1-0)$$

$$= \frac{800}{800} = 1$$

original data	200	300	400	500	1000
(0, 1) normaliz^n	0	0.125	0.25	0.5	1

ii) Z score normalizatio!

given data = {200, 300, 400, 600, 1000}

$$\text{std. deviation} = \sqrt{\frac{\sum (\text{every indi} - (\text{mean data}))^2}{n}}$$

Now, mean value $= \frac{200+300+400+600+1000}{5}$

$= 2500/5 = 500$

$$S.D. = \sqrt{\frac{(200-500)^2+(300-500)^2+(400-500)^2+(600-500)^2+(1000-500)^2}{5}}$$

$$= \sqrt{\frac{(-300)^2+(-200)^2+(-100^2)+(100)^2+(500)^2}{5}}$$

$$= \sqrt{\frac{90000+40000+10000+10000+25000}{5}}$$

$$= \sqrt{\frac{400000}{5}}$$

$$= \sqrt{80000} = 282.8$$

2 score $= \frac{x-\mu}{\sigma} = \frac{200-500}{282.8} = -1.06$
200

2 score $= \frac{x-\mu}{\sigma} = \frac{300-500}{282.8} = -0.7$
300

2 score $= \frac{x-\mu}{\sigma} = \frac{400-500}{282.8} = -0.35$
400

2 score 600 = $\frac{x-\mu}{\sigma} = \frac{600-500}{282.8} = 0.35$

2 score 1000 = $\frac{x-\mu}{\sigma} = \frac{1000-500}{282.8} = 1.78$

original data	200	300	400	600	1000
2 scores	-1.06	-0.7	-0.35	0.35	1.78

Q-1) B) given data : {13, 15, 16, 16, 19, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 36, 40, 45, 46, 52, 70}

i. So soothing by mean in this each value of bin

1) Sort the data

Here data given is already started.

2) Partition data into depth given depth of bin = 3

bin 1 : 13, 15, 16
bin 2 : 16, 19, 20
bin 3 : 20, 21, 22
bin 4 : 22, 22, 25
bin 5 : 25, 25, 30
bin 6 : 33, 33, 35
bin 7 : 35, 35, 35
bin 8 : 36, 40, 45
bin 9 : 46, 52, 70

3) Calculated arithmatic mean of each bin bin 1 : 14, bin 2 : 18, bin 3 : 21, bin 4 : 24, bin 5 : 26, bin 6 : 33, bin 7 : 35, bin 8 : 40, bin 9 : 56.

4) Replace each value in bin by arithmatic mean calc. for bin.

bin 1 : 14, 14, 14, bin 2 : 18, 18, 18, bin 3 : 21, 21, 21, bin 4 : 24, 24, 24, bin 5 : 26, 26, 26, bin 6 : 33, 33, 33

bin 7: 35, 35, 35, bin 8: 40, 40, 40, bin 9: 56, 56, 56

ii. Smooth by Boundry

In this, min & max value of bin are identified as bin boundries. Each bin value is replaced by closet boundary value.

1) Sort the data

Here data is already sorted.

2) Part into bins

bin 1: 13, 15, 16, bin 2: 16, 19, 20, bin 3: 20, 21, 22
bin 4: 22, 25, 25, bin 5: 25, 25, 30, bin 6: 33, 33, 35
bin 7: 35, 35, bin 8: 36, 40, 45, bin 9: 46, 70, 52

3) Find bin boundaries:

bin 1: 13-16, bin 2: 16-20, bin 3: 20-22
bin 4: 22-25, bin 5: 25-30, bin 6: 33-35
bin 7: 35-35, bin 8: 36-45, bin 9: 46-70

4) Substitute boundaries in bin:

bin 1: 13, 16, 16, bin 2: 16, 20, 20, bin 3: 20, 22, 22
bin 4: 22, 25, 25, bin 5: 25, 25, 30, bin 6: 33, 33, 35
bin 7: 35, 35, 35, bin 8: 36, 36, 45, bin 9: 46, 46, 70

iii. Smooth by median

In this, each value in bin is replaced by median of bin.

1) Sort data

Here data is already sorted.

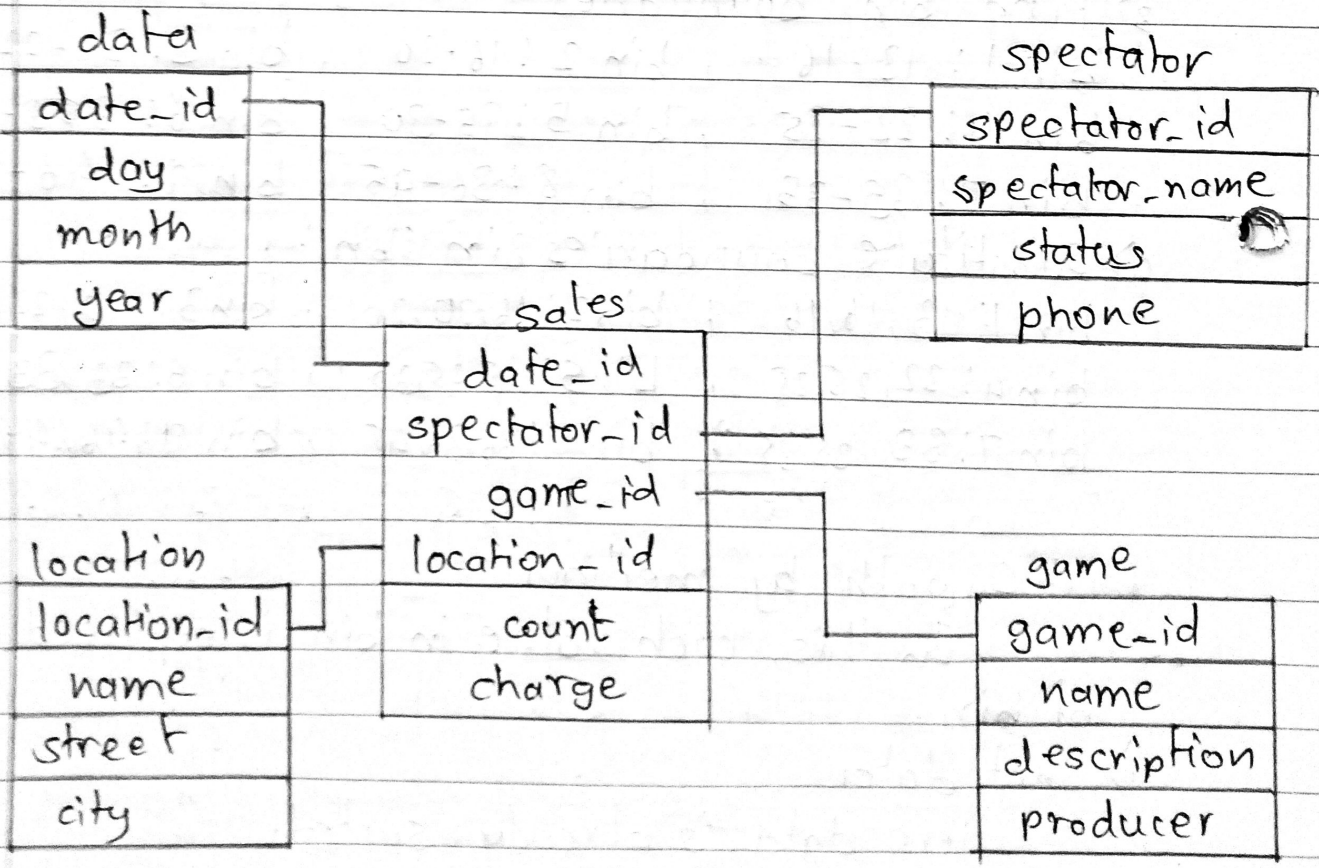
2) Distribute bins:

bin 1: 13, 15, 16 , bin 2: 16, 19, 20 , bin 3: 20, 21, 22
bin 4: 22, 25, 25 , bin 5: 25, 25, 30 , bin 6: 33, 33, 35
bin 7: 35, 35, 35 , bin 8: 36, 40, 45 , bin 9: 46, 52, 70

3) Calculate median of each bin-
bin 1: 15 , bin 2: 19 , bin 3: 21
bin 4: 25 , bin 5: 25 , bin 6: 33
bin 7: 35 , bin 8: 40 , bin 9: 52

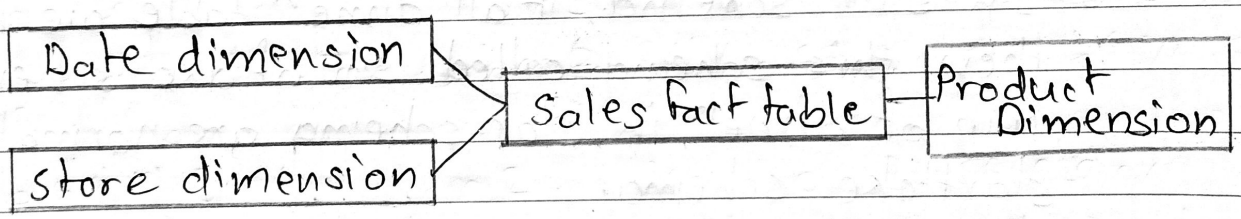
4) Substitute median value in bin
bin 1: 15, 15, 15 , bin 2: 19, 19, 19 , bin 3: 21, 21, 21
bin 4: 25, 25, 25 , bin 5: 25, 25, 25 , bin 6: 33, 33, 33
bin 7: 35, 35, 35 , bin 8: 40, 40, 40 , bin 9: 52, 52, 52

Q-2) A) 1. Draw a star schema diagram:



ii) base cuboid [data; spectator; location; game]
The specific OLAP operⁿ to be performed are.
a) Roll up on date from date_id to year.
b) Roll up on spectator for spectator_id to status.
c) Roll-up on location from location_id to name.
d) Roll_up on game from game_id to all.
e) Dice with status = 1 students location name = 1 Gilla Place, year = 2004.

Q-2) B) i) Star, snowflake, schema, fact constellation.
a) Star schema



1) Star schema is most popular schema design for data warehouse. Dimension stored in dimensions in dimensions table & every entry has its own unique identifier.
2) Every dimension table is related to one or more fact table. All unique identifier from the dimension table make up for composite day in fact table.
3) The fact table also contains facts. A combination of store_id, date_key & product_id giving amount of certain product sold on given day at given store.
4) Typer of facts in star schema:
i. Fully-additive

ii. Semi-additive.
iii. Non-additive.

b) Snowflake Schema:
1) A snowflake schema is used to recover low co-ordinality attribute having low distance value, terual attri. from dimension table & placing them secondary dimension table.
2) Its a normalization process carried out manage size of dimension table. But this may affect is performance as joins need to be performed.
3) In star schema, if all dimen^{n} table are normalized then this schema called snowflake schema & if only few of dimen^{n} in star-schema are normalized called starflake schema.

c) Fact Constellation Schema:
1) As its name implies, its shaped like constell^{n} of star.
2) This schema is more complex than on snowflake varities, which is due to fact that it contains multiple fact table.
3) This allow dimension table to shared amonget fact table.
4) Schema of this type should only be used for app^{ln} needing high level of sophistica^{n}.
5) In fact constellation schema, different fact table are explicitely assigned to dimen^{r} which are given given facts aggr^{n} must be considered.

ii) Enterprise warehouse, datamart:

1. Enterprise Warehouse -

a) Enterprise Warehous collect all of info. about subject open" entire organization. It provide larporate wide data intg", usually from one or more OS or external info. provide & is cross fun" in scope.

b) It typically contain detailed data as well as summarized data & can range in size from few gigabyte to hundred of gb, tb or beyond.

c) It require extensive buisness modelling & may take years to design.

2. Data mart:

a) Data mart contain subset of corporate wide data that it is of value to a specific group of user.

b) Scope is confirmed to specified selected subject.