62020
SEM: 06
Rebecca

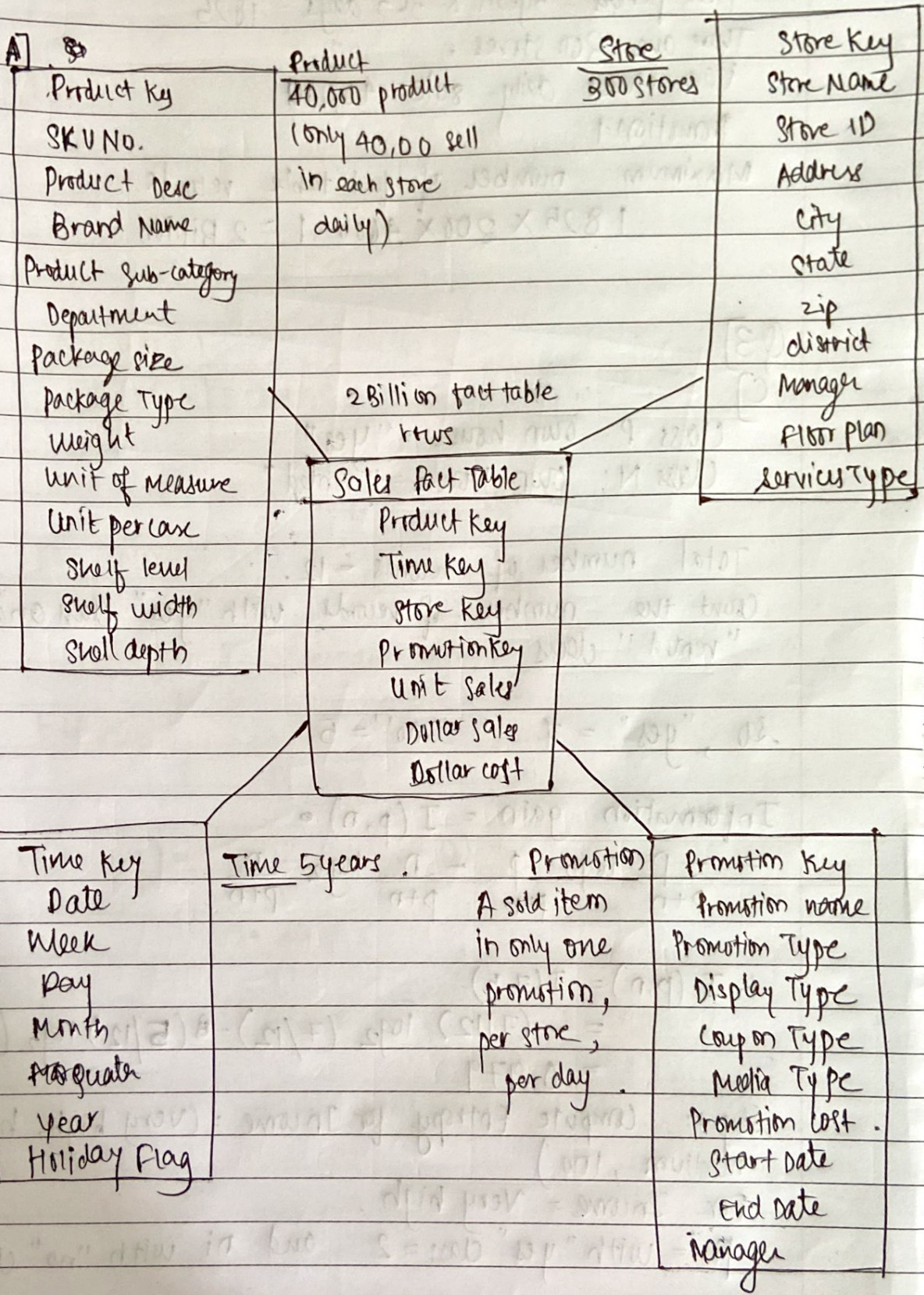
DWM

BARFI Page No. Date

5/9

Q:3] Star schema

A]



62020
SEM: 06 DWM

6/9

Time period = 5years × 365 days = 1825
There are 300 stores,
Each stores daily sale = 40,000
Promotion = 1
Maximum number of fact table records:
1825 × 300 × 4000 × 1 = 2 Billion

Q3]
C]
Class P: own house = "yes"
Class N: own House = "rented"

Total number of records = 12
Count the number of records with "yes" class and "rented" class

So, "yes" = 7 "no" = 5

Information gain = $I(p,n)$
$= \frac{-p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$

$\therefore I(p,n) = I(7,5)$
$= -(7/12)\log_2(7/12) - (5/12)\log_2(5/12)$
$= 0.979$

Step 1: Compute Entropy for Income: (very high, high, medium, low)
For Income = very high,
p_i = with "yes" class = 2 and n_i with "no" class = 0

Therefore, $I(pi, ni) = I(2,0) = 0$

For Income = High,

pi = yes = 4 , ni = no = 0

$\therefore I(pi, ni) = I(4,0) = 0$

For Income = Medium

pi = yes = 1 , ni = no = 2

$\therefore I(pi, ni) = I(1,2) = -(1/3)\log_2(1/3) - (2/3)\log_2(2/3)$

$= 0.918$

For Income = Low

Pi = yes = 0 , ni = no = 3

$\therefore I(pi, ni) = I(0,3) = 0$

$$E(A) = \sum_{i=1}^{v} \frac{pi+ni}{p+n} I(pi, ni)$$

$E(\text{Income}) = 2/12 \times I(2,0) + 4/12 \times I(4,0) + 3/12 \times I(0,3)$

$= 0.229$

Hence,

$\text{Gain}(S, \text{Income}) = I(p,n) - E(\text{Income})$

$= 0.979 - 0.229$

$= 0.75$

Step 2: Compute entropy for age: (Young, medium, old)

Similarly for different age ranges I(pi, ni) is calculated as

Age	pi	ni	I(pi, ni)
Young	3	1	0.811
Medium	3	2	0.971
old	1	2	0.918

Cal. entropy

$$E(Age) = 4/12 \times I(3,1) + 5/12 \times I(3,2) + 3/12 \times I(1,2)$$

$$= \frac{4}{12} \times 0.811 + \frac{5}{12} \times 0.971 + \frac{3}{12} \times 0.918$$

$$= 0.904$$

Hence, $Gain(S, age) = I(p,n) - E(age)$

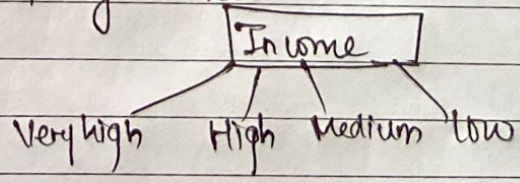
$= 0.979 - 0.904$

$= 0.075$

Income attribute has the highest gain, therefore it is used as the decision attribute in the root node

Step 3: Consider Income = "very high" and count the number of tuples from the original given set

$S_{very\ high} = 2$



9/9

Since both the tuples have class label = "yes"
so directly give "yes" as a class label below "very high"

Check for income = "high" and "low" having "yes" and "rented"

for income = "medium" we have class 'yes' is 1 and rented is 2

so put age label below income = "medium"

Final decision tree is :