Question Bank

Module 1: Data Warehouse and OLAP

(i) Generate Star schema for a Student's Performance chart. Prepare information Package Diagram for the same.

→ Subject: Student's Performance
Facts: Attendance, aggregate

Time dimension	Day	Week	Month	Semester	Year
Professor dimension	Name	professional qualification	Experience	Grade	no. of subjects
Subject dimension	Name	Semester	Theory/ practical		
Student dimension	Name	Grad/PG	family income	Division in XII	Grade
Course dimension	Name	Duration	Type	University	

(i) IPD for a student monitoring system

(ii) Dimension table for Student's Performance Chart

(a) Time Dimension

Time Key	Day	Week	Month	Semester	Year
T001	Monday	First	May	II	2008
T002	Friday	First	March	III	2008
T003	Monday	Second	March	III	2008

(b) Professor Dimension

Professor Key	Name	Professional Qualification	Experience	Grade	No. of subjects
P001	John	MTech	4 years	XII	2
P002	Neil	BTech	2 years	XII	1
P003	Pam	BTech	3 years	XII	1

(c) Subject Dimension

Subject Key	Name	Semester	Theory/ Practical
S001	DWM	III	Theory
S002	WC	III	Theory
S003	AIML	III	Practical

(d) Student Dimension

Student Key	Name	Grad/ PG	Family Income	Division in class XII	Grade
ST001	Brad	Grad	7,00,000	A	XII
ST002	Leo	Grad	9,00,000	B	XII
ST003	Charlie	PG	8,00,000	A	XII

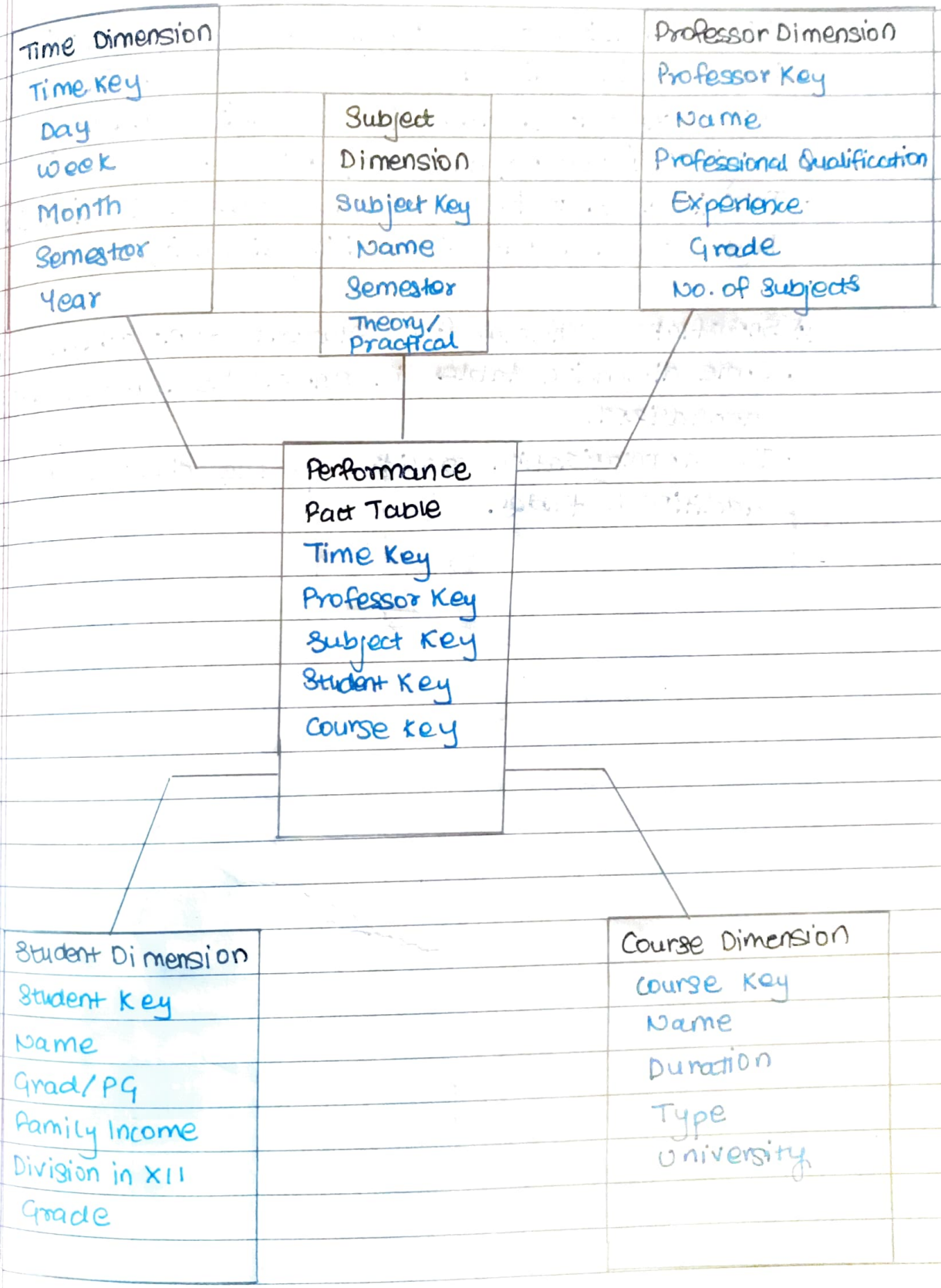
(e) Course Dimension

Course Key	Name	Duration	Type	University
C001	DSGT	7 months	online	MU
C002	DBMS	3 months	online	MU
C003	Stats	5 months	online	PU

(iii) Fact Table for Student's Performance chart

Time Key	Professor Key	Subject Key	Student Key	Course Key
T001	P001	S001	ST001	C001
T002	P002	S002	ST002	C002
T003	P003	S003	ST003	C003

(iv) Star Schema for Student's Performance Table



- Each dimension in the star schema is represented with only one-dimension table.
- This dimension table contains the set of attributes.
- The following diagram shows the student performance with respect to the five dimensions namely time, subject, professor, student, course.
- There is a fact table at the center. It contains the keys to each of the five dimensions.

(v) Snowflake schema for student performance chart
- Some dimension tables in snowflake schema are normalized.
- The normalization splits up the data into additional tables.

Module 2 : Introduction to Data Mining, Data Exploration and Data Preprocessing

Binning

(1) Partition the given data into 4 bins using equi-depth binning method and perform smoothing according to the following methods, smoothing by bin mean, smoothing by bin median, smoothing by bin boundaries.

Data: 11, 13, 13, 15, 15, 16, 19, 20, 20, 20, 21, 21, 22, 23, 24, 30, 40, 45, 45, 45, 71, 72, 73, 75

→ i) Partition into 4 equal depth bins

Bin 1 : 11, 13, 13, 15, 15, 16

Bin 2 : 19, 20, 20, 20, 21, 21

Bin 3 : 22, 23, 24, 30, 40, 45

Bin 4 : 45, 45, 71, 72, 73, 75

ii) Smoothing by bin means

Bin 1 : 13.83, 13.83, 13.83, 13.83, 13.83, 13.83

Bin 2 : 20.16, 20.16, 20.16, 20.16, 20.16, 20.16

Bin 3 : 30.66, 30.66, 30.66, 30.66, 30.66, 30.66

Bin 4 : 63.5, 63.5, 63.5, 63.5, 63.5, 63.5

iii) Smoothing by bin boundaries

Bin 1 : 11, 11, 11, 16, 16, 16

Bin 2 : 19, 19, 19, 19, 21, 21

Bin 3 : 22, 22, 22, 22, 45, 45

Bin 4 : 45, 45, 75, 75, 75, 75

(iv) Smoothing by bin median

Bin 1 : 14, 14, 14, 14, 14, 14

Bin 2 : 20, 20, 20, 20, 20, 20

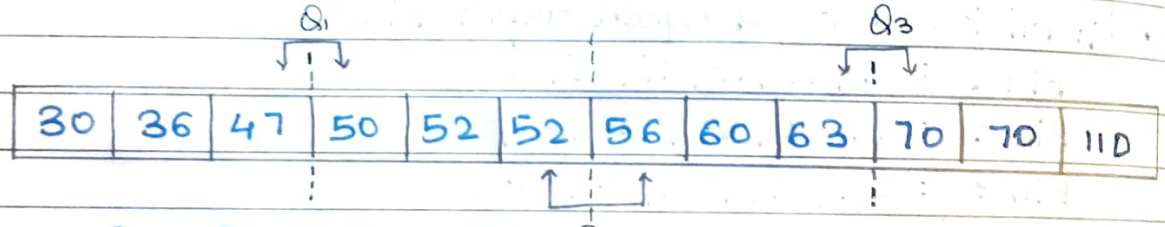
Bin 3 : 27, 27, 27, 27, 27, 27

Bin 4 : 71.5, 71.5, 71.5, 71.5, 71.5, 71.5

Box Plot

(4) Data for salary analysis include 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Compute 1st, 2nd and 3rd quartile for this data. visualize using Box Plot.

→



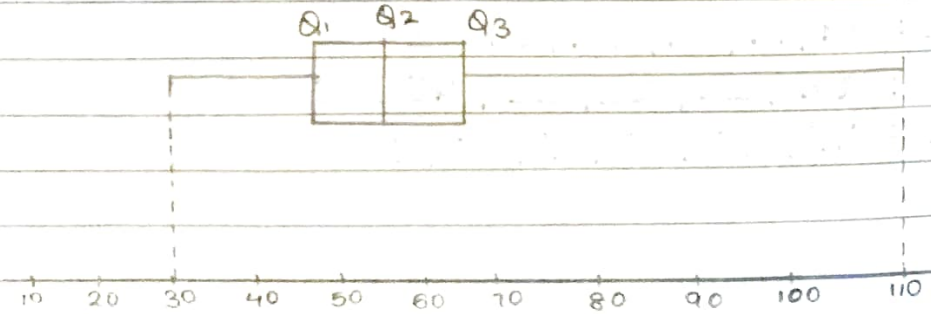
$Q_2 = \frac{52+56}{2} = 54$

$Q_1 = \frac{47+50}{2} = 48.5$

$Q_3 = \frac{63+70}{2} = 66.5$

minimum value = 30

maximum value = 110



Correlation coefficient

(1) Find the value of the correlation coefficient from the following table

subject	Age x	Glucose level y	xy	x^2	y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
	247	486	20485	11409	40022

$n = 6$

Substituting the values in the formula

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

$$= \frac{6(20485) - (247)(486)}{\sqrt{[6(11409) - (247)^2][6(40022) - (486)^2]}}$$

$$= \frac{2868}{\sqrt{7445[-212064]}} \qquad \frac{2868}{5413.27}$$

$$r = 0.5298$$

Normalization of data (z-score / min-max)

(i) Data for salary analysis include
1000, 2000, 3000, 5000, 9000
Apply min-max, z-score, decimal scaling to normalize the data.

→ (i) Min-max normalization : min = 1000
max = 9000

$$v = \frac{x - min}{max - min}$$

(a) $v = \frac{1000 - 1000}{9000 - 1000} = 0$

(b) $v = \frac{2000 - 1000}{9000 - 1000} = 0.125$

(c) $v = \frac{3000 - 1000}{9000 - 1000} = 0.25$

(d) $v = \frac{5000 - 1000}{9000 - 1000} = 0.5$

(e) $v = \frac{9000 - 1000}{9000 - 1000} = 1$

Data	Normalized Data (v)
1000	0
2000	0.125
3000	0.25
5000	0.5
9000	1

(ii) Z-score normalization

$$Z = \frac{x - \mu}{\sigma}$$

$\mu = (1000 + 2000 + 3000 + 5000 + 9000)/5$
$\mu = 4000$

$$\sigma = \sqrt{\frac{\Sigma (x_i - \mu)^2}{n - 1}}$$

$$\sigma = \sqrt{\frac{(1000-4000)^2+(2000-4000)^2+(3000-4000)^2+(5000-4000)^2+(9000-4000)^2}{5-1}}$$

$$= \sqrt{\frac{40000000}{4}}$$

$$= 3162.27766$$

$$= 3162.28$$

(a) $Z = \frac{1000-4000}{3162.28} = -0.9486$

(b) $Z = \frac{2000-4000}{3162.28} = -0.6324$

(c) $Z = \frac{3000-4000}{3162.28} = -0.3162$

(d) $Z = \frac{5000-4000}{3162.28} = 0.3162$

(e) $Z = \frac{9000-4000}{3162.28} = 1.5811$

Data	Normalized Data (v)
1000	-0.9486
2000	-0.6324
3000	-0.3162
5000	0.3162
9000	1.5811

(iii) decimal normalization

$$v_i = \frac{v_i}{10^j \longrightarrow \text{maximum no. of digits}}$$

(a) $v_i = \frac{1000}{10^4} = 0.1$

(b) $v_i = \frac{2000}{10^4} = 0.2$

(c) $v_i = \frac{3000}{10^4} = 0.3$

(d) $v_i = \frac{5000}{10^4} = 0.5$

(e) $v_i = \frac{9000}{10^4} = 0.9$

Module 3 : Classification

Linear Regression

(1)

x	y	x^2	y^2	xy
1	3	1	9	3
2	4	4	16	8
3	5	9	25	15
4	7	16	49	28
10	19	30	99	54

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$a = \frac{19(30) - 10(54)}{4(30) - (10)^2}$$

$$a = 1.5$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{4(54) - 10(19)}{4(30) - (10)^2}$$

$$b = 1.3$$

$y = 1.3x + 1.5$ ($\because y = Bx + A$)

for $x = 1$

$y = 1.3(1) + 1.5$

$y = 2.8$

for $x = 2$

$y = 1.3(2) + 1.5$

$y = 4.1$

For $X=3$

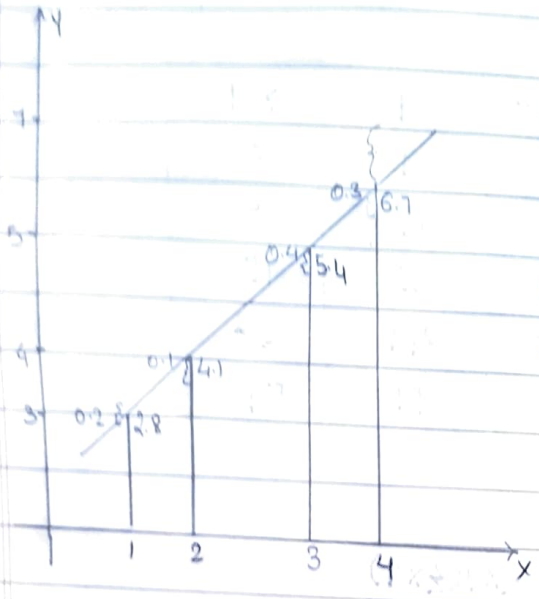
$Y = 1.3(3) + 1.5$

$Y = 5.4$

For $X=4$

$Y = 1.3(4) + 1.5$

$Y = 6.7$



Y	P	Error
		ABS(Y-P)
3	2.8	0.2
4	4.1	0.1
5	5.4	0.4
7	6.7	0.3

Confusion Matrix

(1) Total sample 165

		Actual values: Yes	Actual values: No	
Predicted values	Yes	50 [TP]	10 [FN]	60
	No	5 [FP]	100 [TN]	105
		55	110	165

$$\text{Accuracy} = \frac{TP + TN}{\text{total}}$$

$$= \frac{100 + 50}{165}$$

$$\text{Accuracy} = 0.91$$

Error rate = 1 - Accuracy
= 1 - 0.91
= 0.09

$\left[\frac{FP+FN}{total}\right]$

Precision = $\frac{TP}{TP+PP} = \frac{50}{50+55}$

= $\frac{TP}{\text{Predicted Yes}}$

Recall = $\frac{TP}{\text{actual Yes}} = \frac{100}{105} = 0.95$

$\left[\frac{TP}{TP+FN}\right]$

F-score -

$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

Error rate = $\frac{FP+FN}{total}$ OR 1 - Accuracy

Sensitivity = $\frac{TP}{TP+FN}$

Specificity = $\frac{TN}{TN+FP}$