Time: 03 Hours **Marks: 80**

Note: 1. Question 1 is compulsory
2. Answer any three out of remaining five questions.
3. Assume any suitable data wherever required and justify the same.

Q1 a) Why is data integration required in a data warehouse, more so than in an operational application? [5]

b) Describe the steps involved in Data Mining when viewed as a process of knowledge Discovery. [5]

c) A dimension table is wide, the fact table is deep. Explain [5]

d) Elucidate Market Basket Analysis with an example. [5]

Q2 a) Suppose that a data warehouse consists of the three dimensions time, doctor and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit. [10]

(i) Draw a star schema diagram for the above data warehouse.

(ii) Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2010?

(iii)To obtain the same list, write an SQL query assuming the data are stored in a relational database with the schema fee (day, month, year, doctor, hospital, patient, count, charge).

b) Develop a model to predict the salary of college graduates with 10 years of work experience using linear regression. [10]

Years of experience (x)	Salary in $100 (y)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

Q3 a) Suppose that the data for analysis includes the attribute salary. We have the following values for salary (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. [10]

(i) What are the *mean, median, mode* and *midrange* of the data?

(ii) Find the *first quartile* (Q1) and the *third quartile* (Q3) of the data.

(iii)Show a *boxplot* of the data.

b) Why is entity-relationship modeling technique not suitable for the data warehouse? How is dimensional modeling different? [10]

Q4 a) Why is tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning? [10]

b) Consider the transaction database given below, [10]

TID	Items
10	1, 3, 4
20	2, 3, 5
30	1, 2, 3, 5
40	2, 5
50	1, 3, 5

Use Apriori Algorithm with min-support count = 2 and min-confidence = 60% to find all frequent itemsets and strong association rules.

Q5 a) Show the dendrogram created by the complete link clustering algorithm for the given set of points. [10]

	A	B
P1	2	4
P2	8	2
P3	9	3
P4	1	5
P5	8.5	1

b) What is spatial data? Explain CLARANS Extension. [10]

Q6 a) Demonstrate Multidimensional and Multilevel Association Rule Mining with suitable examples. [10]

b) What is Web Structure Mining? List the, approaches used to structure the web pages to improve on the effectiveness of search engines and crawlers. Explain Page Rank technique in detail. [10]

***