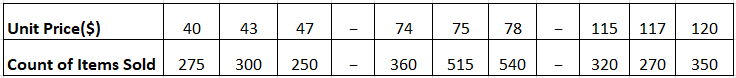
**QUESTION BANK**

**UNIT-I**

1. Compare and contrast Artificial Intelligence, machine learning and Data science [K2]
2. How to extract meaning patterns in data science? [K2]
3. Illustrate how to build representative models[K2]
4. Define Data Science? [K1]
5. Differentiate traditional program and machine learning[K2]
6. List Associated fields of Data Science that heavily relies on it[K2]
7. Outline Data object and attribute? List various types of attributes with examples[K2]
8. Justify the following statement[K4]

“Data science borrows computational techniques from the disciplines of statistics, machine learning, experimentation, and database theories”.

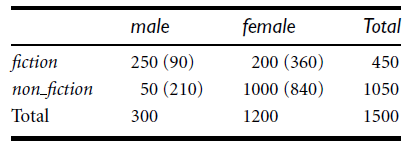
1. Write a python code for summarizing the data using Histogram for the following and plot the result[K2]



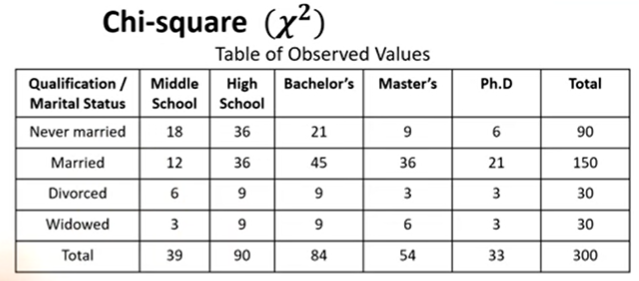
1. Compare and Contrast various Data Science tasks with examples[K2]
2. How to measure the Dispersion of data using range, Quartile, variance, Standard deviation and Interquartile range? Explain with an example[K2]
3. Briefly outline how to compute the dissimilarity between objects described by the following [K2]
   1. Nominal Attributes
   2. Binary Attributes
   3. Ordinal Attributes
   4. Numeric Attributes
4. Write a python code for measuring central tendency (Mean, Median and Mode) for the following values of salary (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110 [K2]
5. What is meant by outliers? [K1]
6. What is Inter Quartile Range(IQR)? [K1]
7. Explain about various attribute types. [K2]
8. Summarize data visualization techniques[K2]

**UNIT-II**

1. Describe Discrete Wavelet Transform[K2]
2. Define Data Integration and Data Transformation? [K2]
3. What is Entity Identification Problem[K2]
4. Elaborate the approaches for cleaning the data[K2]
5. Explain how to handle noisy data[K2]
6. Discussthe following [K2]
   1. Principal Component Analysis
   2. Data Transformation strategies
7. How the missing values can be filled as part of data cleaning? [K2]
8. Discuss different data reduction strategies[K2]
9. Discuss about attribute subset selection[K2]
10. Explain Discretization by binning[K2]
11. Describe Redundancy and Correlation Analysis in data integration[K2]
12. Suppose that a group of 1500 people was surveyed. The gender of each person was noted. Each person was polled as to whether his or her preferred type of reading material was fiction or nonfiction. Thus, we have two attributes, gender and preferred reading. The observed frequency (or count) of each possible joint event is summarized in the contingency table shown below. Calculate the correlation and expected frequencies on the data distribution for both attributes[K3]



1. Illustrate data discretization techniques[K2]
2. Illustrate the Data Transformation by Normalization[K2]
3. Discuss χ2 Correlation Test for Nominal Data. Test whether the two attributes are strongly correlated or not for the following data by assuming the significance level is 0.05 and degree of freedom is 21.03[K3]



16. Discuss greedy (heuristic) methods for attribute subset selection[K2]

17. Summarize the following[K2]

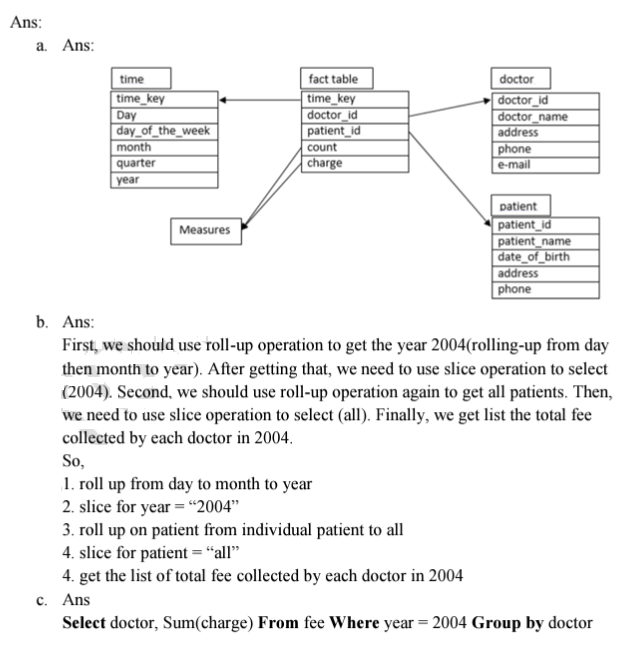
* + - Min-Max Normalization
    - Z-Score Normalization
    - Normalization by decimal scaling

**UNIT-III**

1. What is Data warehouse? [K2]
2. Illustrate the differences between operations database systems and datawarehouses[K2]
3. Define Data mart and enterprise warehouse[K2]
4. Write a short note on data cube[K2]
5. Differentiate OLAP and OLTP systems[K2]
6. Discuss three-tier data warehousing architecture with a neat sketch[K2]
7. Draw the concept hierarchy for location[K2]
8. Discuss about Extraction, Transformation and Loading[K2]
9. Compare and contrast ROLAP and MOLAP[K2]
10. Design a data warehouse for a regional weather bureau. The weather bureau has about 1000 probes, which are scattered throughout various land and ocean locations in the region to collect basic weather data, including air pressure, temperature, and precipitation at each hour. All data are sent to the central station, which has collected such data for more than 10 years. Your design should facilitate efficient querying and online analytical processing[K3]
11. Illustrate Data cube? List and explain star, snow flake and fact constellation schemas. [K2]
12. Discuss the typical OLAP operations[K2]
13. Explain in detail about Fact Constellations[K2]
14. 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. [K3]

* Draw a Snowflake schema for this data warehouse.
* 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 2004?
* 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)

****

15. State why, for the integration of multiple heterogeneous information sources, many companies in industry prefer the update-driven approach (which constructs and uses data warehouses), rather than the query-driven approach (which applies wrappers and integrators). Describe situations where the query-driven approach is preferable over the update-driven approach.

16. Suppose that a data warehouse for *Big-University* consists of the following four dimensions: *student, course, semester*, and *instructor*, and two measures *count* and *avg grade*. When at the lowest conceptuallevel (e.g., for a given student, course, semester, and instructor combination), the *avg grade* measurestores the actual course grade of the student. At higher conceptual levels, *avg grade* stores the averagegrade for the given combination.

(a) Draw a *snowflake schema* diagram for the data warehouse.

(b) Starting with the base cuboid [*student, course, semester, instructor*], what specific *OLAP operations* (e.g., roll-up from *semester* to *year*) should one perform in order to list the average gradeof *CS* courses for each *Big-University* student.

(c) If each dimension has five levels (including all), such as “*student < major < status < university <* all”, how many cuboids will this cube contain (including the base and apex cuboids)?

17. Suppose that a data warehouse consists of the four dimensions, date, spectator, location, and game, and the two measures, count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.

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

(b) Starting with the base cuboid [date, spectator, location, game], what specific OLAP operations should one perform in order to list the total charge paid by student spectators at GM Place in 2010?

18. A data warehouse can be modeled by either a star schema or a snowflake schema. Briefly describe the similarities and the differences of the two models, and then analyze their advantages and disadvantages with regard to one another. Give your opinion of which might be more empirically useful and state the reasons behind your answer.

19. A data cube, C, has n dimensions, and each dimension has exactly p distinct values in the base cuboid. Assume that there are no concept hierarchies associated with the dimensions

(a) What is the maximum number of cells possible in the base cuboid?

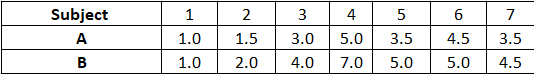
(b) What is the minimum number of cells possible in the base cuboid?

(c) What is the maximum number of cells possible (including both base cells and aggregate cells) in the data cube, C?

(d) What is the minimum number of cells possible in the data cube, C?

**UNIT-IV**

1. Write the formulas for support(A→B) and confidence (A→B) [K2]
2. Write a short note on clustering and classification? [K2]
3. Mention the importance of Association Rule Mining[K2]
4. How to generate association rules from the frequent item sets? Explain with an example. [K2]
5. Write the need for tree pruning in decision tree induction? [K4]
6. Write the algorithm for decision tree induction[K2]
7. Define frequent sets, confidence and support[K1]
8. Distinguish supervised learning from unsupervised learning[K4]
9. State K-means algorithm. Apply k-means algorithm with two iterations to form two clusters by taking the initial cluster centres as subjects 1 and 4 [K3]



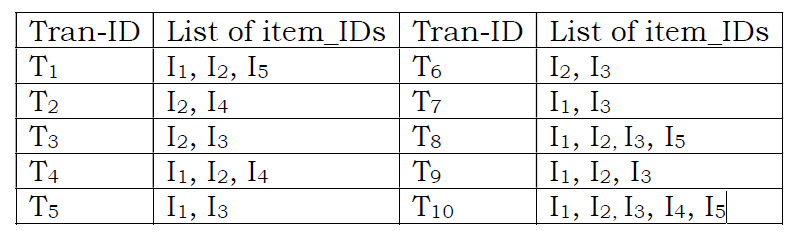
1. A database has five transactions. Let min sup = 60% and min conf = 80%.[K3]



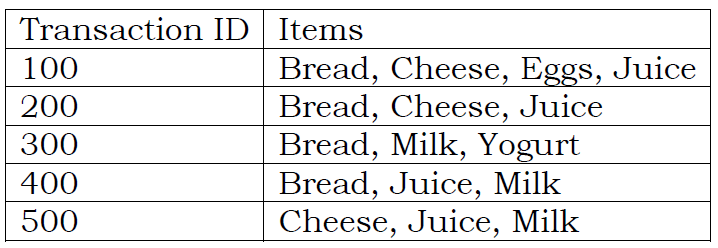
* 1. Find all frequent itemsets using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes.
  2. List all the strong association rules (with support s and confidence c) matching the following meta rule, where X is a variable representing customers, and itemi denotes variables representing items (e.g., “A,” “B,”):



1. What is prediction? Explain about Decision tree Induction classification technique. [K2]
2. Explain Apriori algorithm with an example for mining frequent item sets[K3]
3. Explain market basket analysis[K2]
4. Construct an FP-tree for the dataset given below[K3]



1. A database has five transactions. Find all frequent item sets using Apriori algorithm with minimum support of 20% and minimum confidence of 80%[K3]



1. Illustrate k-means clustering algorithm[K2]
2. Discuss attribute selection measures [K2]
3. Discuss about Gini index, Gain Ratio[K2]
4. List the requirements for cluster analysis[K1]

