|  |  |
| --- | --- |
| **SET- C Key** | |
| **1** | Which of the following process uses intelligent methods to extract data patterns?  **a. Data mining**  b. Text mining  c. Warehousing  d. Data selection |
| **2** | Rahul’s family drove through 7 states on summer vacation. The prices of Gasoline differ from state to state. Calculate the median of gasoline cost.  1.79, 1.61, 2.09, 1.84, 1.96, 2.11, 1.75   1. 1.61   **b. 1.84**   1. 1.75 2. 1.96 |
| **3** | Smoothing is one of the strategies of   1. Data Reduction 2. **Data Transformation** 3. Data Cleaning 4. Data integration |
| **4** | Ramu downloaded a cloud dataset that has no class labels and he wants to generate the class label for the group of data. Which model he will choose?   1. Classification 2. Regression 3. Logistic Regression   **d. Clustering** |
| **5** | The classification or mapping of a class using a predefined class or group is called:  a. Data Sub Structure  b. Data Set  **c. Data Discrimination**  d. Data Characterisation |
| **6** | Which graph/chart is used to visually examine the relationship between two quantitative variables.   1. Bar graph 2. **Scatterplot** 3. Line graph 4. Pie chart |
| **7** | The weight of 8 boys in kgs are 54, 49, 51, 58, 61, 52, 54, 60. Find the median weight.   1. 49 2. 51 3. **54**   d. 58 |
| **8** | Quartiles divide the entire set into   1. 2 equal parts 2. 3 equal parts 3. **4 equal parts**   5 equal parts |
| **9** | The analysis that measures how strongly one attribute implies the other, based on the available data is   1. **Correlation analysis** 2. Variance 3. Business analysis 4. Performance analysis |
| **10** | The process of reducing the number of random variables or attributes under consideration is   1. Numerosity reduction 2. **Dimentionality reduction** 3. Data Compression 4. None of the above |

**Part-B**

**11. Write Short notes on kinds of data that can be mined.**

As a general technology, data mining can be applied to any kind of data as long as the data are meaningful for a target application. The most basic forms of data for mining applications are database data, data warehouse data, and transactional data.

**Database Data**

* A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data.
* The software programs provide mechanisms for defining database structures and data storage; for specifying and managing concurrent, shared, or distributed data access; and for ensuring consistency and security of the information stored despite system crashes or attempts at unauthorized access.
* A relational database is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows).
* Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values.
* A semantic data model, such as an entity-relationship (ER) data model, is often constructed for relational databases. An ER data model represents the database as a set of entities and their relationships.

**Data Warehouses**

* A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.
* A data warehouse is usually modeled by a multidimensional data structure, called a data cube, in which each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure such as count or sum(sales amount). A data cube provides a multidimensional view of data and allows the precomputation and fast access of summarized data

**Transactional Data**

* In general, each record in a transactional database captures a transaction, such as a customer’s purchase, a flight booking, or a user’s clicks on a web page.
* A transaction typically includes a unique transaction identity number (trans ID) and a list of the items making up the transaction, such as the items purchased in the transaction.
* A transactional database may have additional tables, which contain other information related to the transactions, such as item description, information about the salesperson or the branch, and so on.

**12. Find the mean, median, mode and range for the given data:**

**90, 94, 53, 68, 79, 94, 53, 65, 87, 90, 70, 69, 65, 89, 85, 53, 47, 61, 27, 80**

**Mention the formula for all the types.**

Given,

90, 94, 53, 68, 79, 94, 53, 65, 87, 90, 70, 69, 65, 89, 85, 53, 47, 61, 27, 80

Number of observations = 20

**Mean** = (Sum of observations)/ Number of observations

= (90 + 94 + 53 + 68 + 79 + 94 + 53 + 65 + 87 + 90 + 70 + 69 + 65 + 89 + 85 + 53 + 47 + 61 + 27 + 80)/20

= 1419/20

= 70.95

Therefore, mean is 70.95.

**Median:**

The ascending order of given observations is:

27, 47, 53, 53, 53, 61, 65, 65, 68, 69, 70, 79, 80, 85, 87, 89, 90, 90, 94,94

Here, n = 20

Median = 1/2 [(n/2) + (n/2 + 1)]th observation

= 1/2 [10 + 11]th observation

= 1/2 (69 + 70)

= 139/2

= 69.5

Thus, the median is 69.5.

**Mode:**

The most frequently occurred value in the given data is 53.

Therefore, mode = 53

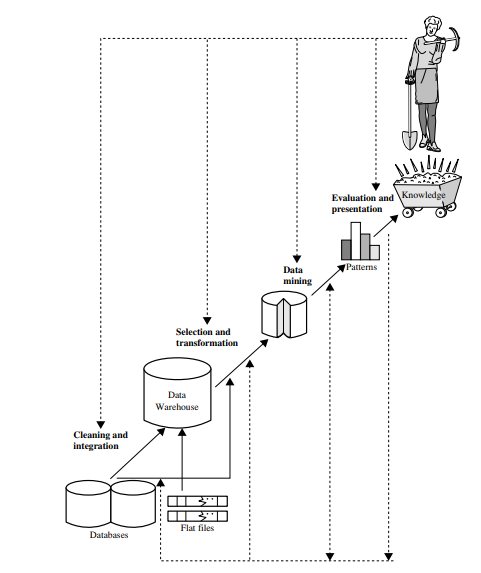
**Range** = Highest value – Lowest value

= 94 – 27

= 67

**Part- C**

**13. Explain the steps in the process of knowledge discovery in databases**



1. Data cleaning (to remove noise and inconsistent data)

2. Data integration (where multiple data sources may be combined)

3. Data selection (where data relevant to the analysis task are retrieved from the database)

4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)

5. Data mining (an essential process where intelligent methods are applied to extract data patterns)

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness measures)

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

**14. What is data cleaning? Explain the different methods of data cleaning.**

Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

1. Missing Values

a. Ignore the tuple

b. Fill in the missing value manually

c. Use a global constant to fill in the missing value

d. Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value

e. Use the attribute mean or median for all samples belonging to the same class as the given tuple

f. Use the most probable value to fill in the missing value

**2.** Noisy Data “

Noise is a random error or variance in a measured variable. Some basic statistical description techniques (e.g., boxplots and scatter plots), and methods of data visualization can be used to identify outliers, which may represent noise. The following are the data smoothing techniques

a. Binning

b. Regression

**c.** Outlier analysis

**3.** Data Cleaning as a Process

The first step in data cleaning as a process is discrepancy detection. Discrepancies can be caused by several factors, including poorly designed data entry forms that have many optional fields, human error in data entry, deliberate errors (e.g., respondents not wanting to divulge information about themselves), and data decay (e.g., outdated addresses). Discrepancies may also arise from inconsistent data representations and inconsistent use of codes. Other sources of discrepancies include errors in instrumentation devices that record data and system errors. Errors can also occur when the data are (inadequately) used for purposes other than originally intended. There may also be inconsistencies due to data integration (e.g., where a given attribute can have different names in different databases).