1. The steps involved in KDD are

Understanding the data set :

It is critical to know the properties, limitations, and rules of data or information

Data selection:

Data selected must be helpful in order to meet the goals of the analysis

Cleaning and preprocessing:

Data cleaning is done in order to achieve make the data more useful and meaningful. For example, removing the variables with missing values or elimination audio or image that is not critical for the analysis

Data transformation:

Transformation process such as dimensionality reduction, and converting the values from numeric to categorical is achieved in this step .

Select the data mining task:

Classification, Regression or grouping are the various data mining tasks that has to be chosen according to the goal of the project .

Selecting data mining Algorithm :

We choose the algorithm for example to search for the pattern and get the necessary information. Choose the algorithm that fits the data well. The algorithm should work well on the data.

Applying data mining algorithm:

The selected, cleaned and well-processed data will be used now by applying the algorithms on them. Tune the parameters and achieve the desired results. The parameters differ according to the method selected.

Evaluation and Interpretation:

After applying the algorithm to the data set, perform the evaluation of patterns and the performance . for example cross-validation technique divides the data as training and test

Interpretation:

Solve the problem by interpreting the derived results in this step. If the results need more value , begin repeating the process from analyzing the selected data to evaluation stage .

1. Answer in paper format

3)Google file system :

Google File System, a scalable distributed file system for large, distributed data-intensive applications. It provides fault tolerance while running on inexpensive commodity hardware, and it delivers high aggregate performance to many clients.

First, component failures are the norm rather than the exception.

•Second, files are huge by traditional standards. Multi-GB files are common.

•Third, most files are mutated by appending new data rather than overwriting existing data.

•Fourth, co-designing the applications and the file system API benefits the overall system by increasing our flexibility.

4) Answer in paper format .

The key will be a combination of demographics and category... Colour is the value. The end output is the show colour.So the output is the value. Based on demo and category the map reduce will fetch colours. So key is combination of demographics and category

5) answer in paper format

6).

<https://github.com/devisjsu/Midterm256/blob/main/SingleLinkMidterm.ipynb>

7) Long tail phenomena:

The long tail is a business strategy that allows companies to realize significant profits by selling low volumes of hard-to-find items to many customers, instead of only selling large volumes of a reduced number of popular items.

The long tail concept considers less popular goods that are in lower demand.

This theory is supported by the growing number of online marketplaces that alleviate the competition for shelf space and allow an unmeasurable number of products to be sold, specifically through the Internet.

* The strategy theorizes that consumers are shifting from mass-market buying to more niche or artisan buying.

## **Long Tail Probability and Profitability**

The long tail of distribution represents a period in time when sales for less common products can return a profit due to reduced marketing and distribution costs. Overall, long tail occurs when sales are made for goods not commonly sold. These goods can return a profit through reduced marketing and distribution costs.

8)

Distances from each point to centroid :

0.9775479

2.3739419

1.5085092

1.89620674

0.52497619

1.89620674

Radius = max(distances) = 2.37

Diameter = 4.24 [distance between (12,6) and (9,3)]

9)

<https://github.com/devisjsu/Midterm256/blob/main/MarketingSingleLink.ipynb>

10)

Information retrieval (I) is the study of helping users to find information that matches their information needs. Technically, IR studies the acquisition, organization, storage, retrieval, and distribution of information. Historically, IR is about document retrieval, emphasizing the document as the basic unit.

(architecture uploaded in the paper format)

11)

<https://github.com/devisjsu/Midterm256/blob/main/KmeansClustering.ipynb>

12)paper format