Unit 1 :- Distributed Database :

- A **distributed database system** allows applications to access data from local and remote databases. In a **homogenous distributed database system**, each database is an Oracle Database. In a **heterogeneous distributed database system**, at least one of the databases is not an Oracle Database. Distributed databases use a **client/server** architecture to process information requests.

- A *table space* is a storage structure containing tables, indexes, large objects, and long data. They are used to organize data in a database into logical storage groupings that relate to where data is stored on a system. Table spaces are stored in database partition groups.

- Range partitioning **maps data to partitions based on ranges of values of the partitioning key that you establish for each partition**. It is the most common type of partitioning and is often used with dates.

Unit 2:- OLAP with Oracle:-

. What is OLAP?

OLAP is an acronym for Online Analytical Processing. OLAP performs multidimensional analysis of business data and provides the capability for complex calculations, trend analysis, and sophisticated data modeling.

2. What are Cubes in OLAP?

A data cube stores data in a summarized version which helps in a faster analysis of data. The data is stored in such a way that it allows reporting easily. E.g. using a data cube a user may want to analyze weekly, monthly performance of an employee. Here, month and week could be considered as the dimensions of the cube.

3. Where is Data Sources in OLAP?

Data source is where the data comes from in data warehousing. The data collected from various sources and is cleaned. The data source can be internal or external. Efficient Analysis and cleansing of source data is the key success to data warehousing.

4. What are Fact Tables?

Data in a warehouse comes from the transactions. Fact table in a data warehouse consists of facts and/or measures. The nature of data in a fact table is usually numerical..

5. What are Database roles in OLAP?

Database level roles are used to manage the security of the database. The role can be either fixed or flexible. Fixed roles are predefined while flexible roles can be created.

What is roll up in OLAP operations?

The ROLLUP command **calculates totals for a hierarchy of values where each level of the hierarchy is an aggregation of the values in the level below it**. The ROLLUP command only performs simple sum aggregation.

What are OLAP cubes used for?

An OLAP cube is a data structure that **overcomes the limitations of relational databases by providing rapid analysis of data**. Cubes can display and sum large amounts of data while also providing users with searchable access to any data points.

What is the use of RANK and Dense_rank?

RANK and DENSE_RANK are used **to order values and assign them numbers depending on where they fall in relation to one another**.

Can we use dense rank without ORDER BY?

Data does not have an order. You can't expect it to be returned in the same order unless you specify an order by -- so **trying to implement a rank without order by doesn't make sense**.

PARTITION BY clause groups the data by the columns defined in the PARTITION BY clause and performs the OLAP function within each group.

Unit 6:- Preprocessing in R

What Is Data Preprocessing?

Data preprocessing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning.

Data Preprocessing Steps

There are a number of data anomalies and inherent problems to look out for in almost any data set, for example:

- **Mismatched data types:** When you collect data from many different sources, it may come to you in different formats. While the ultimate goal of this entire process is to reformat your data for machines, you still need to begin with similarly formatted data.
  .
- **Mixed data values:** Perhaps different sources use different descriptors for features – for example, *man* or *male*. These value descriptors should all be made uniform.

- **Data outliers:** Outliers can have a huge impact on data analysis results. For example if you're averaging test scores for a class, and one student didn't respond to any of the questions, their 0% could greatly skew the results.

- **Missing data:** Take a look for missing data fields, blank spaces in text, or unanswered survey questions. This could be due to human error or incomplete data. To take care of missing data, you'll have to perform data cleaning

2. Data cleaning

Data cleaning is the process of adding missing data and correcting, repairing, or removing incorrect or irrelevant data from a data set. Data cleaning will correct all of the inconsistent data you uncovered in your data quality assessment.

Missing data

There are a number of ways to correct for missing data, but the two most common are:

- **Ignore the tuples:** A tuple is an ordered list or sequence of numbers or entities. If multiple values are missing within tuples, you may simply discard the tuples with that missing information.

- **Manually fill in missing data:** This can be tedious, but is definitely necessary when working with smaller data sets.

Noisy data

Data cleaning also includes fixing “noisy” data. This is data that includes unnecessary data points, irrelevant data, and data that’s more difficult to group together.

- **Binning:** Binning sorts data of a wide data set into smaller groups of more similar data. It’s often used when analyzing demographics.
- **Regression:** Regression is used to decide which variables will actually apply to your analysis. Regression analysis is used to smooth large amounts of data. This will help you get a handle on your data, so you’re not overburdened with unnecessary data.
- **Clustering:** Clustering algorithms are used to properly group data, so that it can be analyzed with like data. They’re generally used in unsupervised learning, when not a lot is known about the relationships within your data.

3. Data transformation

With data cleaning, we’ve already begun to modify our data, but data transformation will begin the process of turning the data into the proper format(s) you’ll need for analysis and other downstream processes.

This generally happens in one or more of the below:

1. Aggregation
2. Normalization
3. Feature selection
4. Discreditization
5. Concept hierarchy generation

4. Data reduction

Data reduction not only makes the analysis easier and more accurate, but cuts down on data storage.

It will also help identify the most important features to the process at hand.

- **Attribute selection:** Similar to discreditization, attribute selection can fit your data into smaller pools. It, essentially, combines tags or features, so that tags like *male/female* and *professor* could be combined into *male professor/female professor*.

- **Numerosity reduction:** This will help with data storage and transmission. You can use a regression model.

- **mensionality reduction:** This, again, reduces the amount of data used to help facilitate analysis and downstream processes. Algorithms like *K-nearest neighbors* use pattern recognition to combine similar data and make it more manageable.

Unit 7:- Data Mining - Classification using R-Programming :

1. What is a Linear Regression?

In simple terms, linear regression is adopting a linear approach to modeling the relationship between a dependent variable (scalar response) and one or more independent variables

What package is ggplot in R?

ggplot2 is **a R package dedicated to data visualization**. It can greatly improve the quality and aesthetics of your graphics, and will make you much more efficient in creating them

What is dim() function?

The DIM function **returns the number of elements in a one-dimensional array or the number of elements in a specified dimension of a multidimensional array when the lower bound of the dimension is 1**.

What does the head () function do?

The head() method **returns a specified number of rows, string from the top**. The head() method returns the first 5 rows if a number is not specified.

- The predict() function in R is **used to predict the values based on the input data**. All the modeling aspects in the R program will make use of the predict() function in their own way
- Data Frames are **data displayed in a format as a table**. Data Frames can have different types of data inside it. While the first column can be character , the second and third can be numeric or logical . However, each column should have the same type of data.

- cor() computes the correlation coefficient.

- cor.test() test for association/correlation between paired samples. It returns both the correlation coefficient and the significance level(or p-value) of the correlation .

- The plot() function is **used to draw points (markers) in a diagram**. The function takes parameters for specifying points in the diagram. Parameter 1 specifies points on the x-axis. Parameter 2 specifies points on the y-axis.

- The lm() function is **used to fit linear models to data frames** in the R Language. It can be used to carry out regression, single stratum analysis of variance, and analysis of covariance to predict the value corresponding to data that is not in the data frame

- We use the function ggplot() **to produce the plots when using the package**. Therefore, ggplot() is the command, and the whole package is called ggplot2.

Unit 8:- Data Mining - Clustering and Association using R-Programming :

- What is the data() function in R?
  There are more than 100 datasets available in R, included in the datasets package. **To see the list of available datasets, use data() function**.

- The head() function in R is **used to display the first n rows present in the input data frame**

- The unique() function in R is **used to eliminate or delete the duplicate values or the rows present in the vector, data frame, or matrix as well**. The unique() function found its importance as it directly identifies and eliminates the duplicate values in the data.

- The pairs function is provided in R Language by default and it **produces a matrix of scatterplots**. The pairs() function takes the data frame as an argument and returns a matrix of scatter plots between each pair of variables in the data frame.

- What does the train function do in R?
  train can be used to **tune models by picking the complexity parameters that are associated with the optimal resampling statistics**.

- What does DF () do in R?
  A data frame is used for **storing data tables**. It is a list of vectors of equal length.

- What is the function of KNN?
  The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which **uses proximity to make classifications or predictions about the grouping of an individual data point**

Unit 9:- Implementation and analysis of Apriori Algorithm using Market Basket Analysis.

- Split() is a built-in R function that **divides a vector or data frame into groups according to the function's parameters**. It takes a vector or data frame as an argument and divides the information into groups.

- The head() function in R is **used to display the first n rows present in the input data frame**. In this section, we are going to get the first n rows using head() function.

- Apriori algorithm is used for finding frequent itemsets in a dataset for **association rule mining**. It is called Apriori because it uses prior knowledge of frequent itemset properties. We apply an iterative approach or level-wise search where k-frequent itemsets are used to find k+1 itemsets. To improve the efficiency of the level-wise generation of frequent itemsets an important property is used called Apriori property which helps by reducing the search space

- ***Apriori Property:*** *All non-empty subsets of a frequent itemset must be frequent. Apriori assumes that all subsets of a frequent itemset must be frequent (Apriori property). If an itemset is infrequent, all its supersets will be infrequent.*

- The inspect function **opens an interactive window that allows for the manipulation of a number of arguments**. It offers several views to analyze the series graphically. With each change, the adjustment process and the visualizations are recalculated.

Unit 10:-

- What is the kmeans function?
  k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.

- The function geom_point() **adds a layer of points to your plot, which creates a scatterplot**. ggplot2 comes with many geom functions that each add a different type of layer to a plot.