1. **List of some best tools that can be useful for data-analysis?**

[R and Python](https://www.edureka.co/blog/top-10-data-analytics-tools/#randpython) [Microsoft Excel](https://www.edureka.co/blog/top-10-data-analytics-tools/#excel) [Tableau](https://www.edureka.co/blog/top-10-data-analytics-tools/#tableau) [RapidMiner](https://www.edureka.co/blog/top-10-data-analytics-tools/#rapidminer) [Power BI](https://www.edureka.co/blog/top-10-data-analytics-tools/#powerbi) Weka

1. **Mention what is the responsibility of a Data analyst?**

Data Analysts **search and extract relevant data for their organizations**. After extracting data, they perform exploratory data analysis to understand it and convert it into structured form for Data Scientists to build Machine Learning models.

1. **List out some of the best practices for data cleaning?**

1. Develop a Data Quality Plan

2. Standardize Contact Data at the Point of Entry

3. Validate the Accuracy of Your Data

4. Identify Duplicates

5. Append Data

1. **What is data cleansing?**

Data cleansing or data cleaning is the process of detecting and identifying corrupt or inaccurate records from a record set, table, or database then replacing, modifying, or deleting the dirty or coarse data.

1. **Name data cleaning techniques?**

Remove Irrelevant Values.

Get Rid of Duplicate Values.

Avoid Typos (and similar errors)

Convert Data Types.

Take Care of Missing Values

1. Define K-means and hierarchical clustering?

**K-means** is an unsupervised Learning algorithm, which groups the unlabeled **dataset** into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process.

**Hierarchical clustering** - a subset of similar data is created in a tree-like structure in which the root node corresponds to entire data, and branches are created from the root node to form several clusters

1. Difference between K-means and KNN Algorithm?

K-NN is a Supervised machine learning while K-means is an unsupervised machine learning. K-NN is a classification or regression machine learning algorithm while K-means is a clustering machine learning algorithm. K-NN is a lazy learner while K-Means is an eager learner.

1. What is ETL?

ETL stands for **Extract, Transform and Load**. An ETL tool extracts the data from different RDBMS source systems, transforms the data like applying calculations, etc. and then load the data to Data-Warehouse system.

1. **Explain Association Rule**

Association rules are if-then statements that help to show the probability of relationships between data items within large data sets in various types of databases.

1. What is the Application of A-Priori algorithm?

Best known algorithm for frequent item set mining and association rules learning over relational databases. It uses prior knowledge of frequent itemset properties.

**Step1:** Find all item sets that have minimum support.

**Step2:** It Create Association rule with support and Confidence.

1. What is Market Basket Analysis? Explain with suitable example?

Market basket analysis is one of the key data mining-technique used by large retailers to uncover associations between items to increase sales by better understanding customer purchasing patterns. It works by looking for combinations of items that occur together frequently in transactions.

1. What is Euclidean distance? Explain with Suitable example?

The Euclidean distance between two points in Euclidean space is the length of a line segment between the two points. It can be calculated from the Cartesian coordinates of the points using the Pythagorean theorem, aka the Pythagorean distance

1. What is hamming distance? Explain with Suitable example?

The Hamming distance between two equal-length strings of symbols is the number of positions at which the corresponding symbols are different.

1. What is Chi Square Distance? Explain with Suitable example?

Chi-square distance calculation is a statistical method, generally measures similarity between 2 feature matrices. Applications like similar image retrieval, image texture, feature extractions etc.

1. **What is use of Tokenize operator?**

Tokenize is an operator for splitting the sentence in the document into a sequence of words. The purpose of this sub process is to separate words from a document, so this list of words can be used for the next sub process.

1. Why we use Filter token and Filter stop word?

Removes stop words from a token stream. The stop filter supports predefined stop word lists for several languages. You can also specify your own stop words as an array or file.

1. How to use Read Document operator?

Open RapidMiner and click "New Process". On the left-hand pane find a tab that says "Operators"- this is where you can search and find all of the operators for RapidMiner including read document operator.

1. How to use Filter Class operator?

SAME AS READ DOCUMENT OPERATOR. Example of filter class operators are “equals,” “not equal to,” “includes,” and “excludes.