Data Cleaning, Transformation and Preprocessing

1) **Below Question needs to be asked before Data Analysis.**
   - Do character variables have valid values?
   - Are numerical variables within range?
   - Are there any missing values?
   - Are there any duplicated values?
   - Are values unique for some variables (e.g. ID VARIABLE?)
   - Are the dates valid?
   - Do we need to combine multiple data files?
   - Do we need to calculate multiple columns?
2) **What is Data Cleaning?**
   Data cleaning is process of identifying and removing an errors, duplicate and junks from the collected data in order to create a reliable dataset. This improves the quality of the training data for analytics and enables accurate decision-making.
3) **What is Data pre-processing?**
   Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.
4) **Why we use Data pre-processing?**
   In real world data is often incomplete, inconsistent and lacking in certain behaviors or trends, and is likely to contain many errors. Data pre-processing is a proven method of resolving this type of data.
5) **What is missing value?**
   Missing value is a data point that is not stored for a variable in the observation of interest.
6) **What is an outlier?**
   An Outlier is a data point that differs significantly from other observations.
7) **What is feature Scaling?**
   Feature scaling is the method to limit the range of variables so that they can be compared on common grounds.
   - Normalization
   - Box-Cox transformation
   - Principal Component Analysis.
8) **What is Dummy Variables?**
   Dummy Variables is the one that takes the value 0 or 1 to indicate the absence or presence of some Categorical data.
9) **What is Data partition?**
   Data partition is a technique to split a dataset into two parts that is training and testing.
   So, you can build model on training and test model performance on testing.
10) **What is training data set?**
    The training set represents a majority of the available data (about 80%), and it trains the model.
11) **What is testing data set?**
    The testing set represents a small proportion of the data set (about 20%) and it is used to test the accuracy of the data.
12) **What is categorical Data?**

Categorical data is a collection of information that is divided into groups. E.g. Gender = male/female

13) **What is Numerical Data?**
    Numerical data is data that is measurable, such as time, height, weight, amount and so on.
14) **What is Cross Tabulation?**
    Cross tabulation is a method to quantitatively analyze the relationship between multiple variables. Also known as contingency table or cross tabs.
15) **What is Rank?**
    Rank is a relationship between a set of items.
16) **What is Interval?**
    An interval is a space between things, points, limits etc.
17) **What is Impute?**
    Imputation is a method to handle missing values from dataset.