## Guidelines for Data Scientist

It should follow orderly for better results, it's not textbook rule but based on the practice that's what these are called guidelines.

1. Identify the type of data for each column --- Numerical, categorical, ordinal
2. Check and remove all DUPLICATE RECORDS from the DATAFRAME record means check row.
3. Check and remove all DUPLICATE COLUMNS from the DATAFRAME
4. If your columns are numerical columns, perform the following steps:

   If the data is CONTINOUS, check with reference to domain whether the following parameters are valid or not:

   1. Negative Numbers are allowed or not.
   2. Positive Numbers are allowed or not.
   3. Decimals are allowed or just Integers are Expected.

   - If any of the above is not allowed, delete that specific column entry.

   - If the data is DISCRETE, check with reference to domain whether the following parameters are valid or not:

   1. Negative Numbers are allowed or not.
   2. Positive Numbers are allowed or not.
   3. Decimals are allowed or just Integers are Expected.
   4. Check whether the number falls in the specified Range defined by DOMAIN

   If any of the above is not allowed, delete that specific column entry.

5. If your columns are categorical columns, perform the following:

   1. Get the unique values of the columns.
   2. Handle the data which has Spelling Errors, Case Errors (lowercase, uppercase)
   3. Check whether the groups/categories shown in the unique values match the domain spec.

   If there exists an unusual category, delete that specific entry

6. If your columns are Ordinal Column, perform the following,

   1. Get the unique values of the columns.
   2. Handle the data which has Spelling Errors, Case Errors (lowercase, uppercase)
   3. Check whether the groups/categories shown in the unique values match the domain spec.

   If there exists an unusual category, delete that specific entry

   4. Check the mathematical weightage of each unique group. Ensure it matches the domain spec
   5. If no mathematical weightage is present in the column, ensure you define the same at metadata level.

7. Dealing with Date Column (Only for Time Series Analysis)

   1. Convert date column into datetime datatype
   2. Ensure your row index is replaced with date column

**All the above rules must be backed with DOMAIN SPECIFICATIONS. If required, add more rules considering the domain requirement..**