Dataset Pre-Processing

i. Null value handling

- **Delete rows with missing values**
  Missing values can be handled by deleting the rows or columns having null values. If columns have more than half of the rows as null then the entire column can be dropped. The rows which are having one or more columns values as null can also be dropped.
- **Impute missing values with mean / median value**
  Columns in the dataset which are having numeric continuous values can be replaced with the mean, median, or mode of remaining values in the column.
- **Imputation method for categorical columns**
  When missing values is from categorical columns (string or numerical) then the missing values can be replaced with the most frequent category.

ii. Feature Selection

- **Domain knowledge**
  A data scientist or analyst is expected to have domain knowledge about the problem statement, and the set of features for any data science case study. Having domain knowledge or intuition about the features will help the data scientist to do the feature engineering and select the best features. For example, for a car price prediction problem, some features like manufacture year, fancy license number are key factors deciding the price of the car.
- **Missing values**
  The real-world dataset often contains missing values, caused due to data corruption or failure to record. There are various techniques to impute the missing values, but imputing the missing value may not match the real data. Hence model trained on features having a lot of missing value may not be of great importance. The idea is to drop the columns or features, that have missing values greater than a decided threshold.
- **Forward feature selection**
  Forward feature selection technique is used to find the subset of best-performing features for the machine learning model. For a given dataset if there are n features, the features are selected based on the inference of previous results. The forward feature selection techniques follow:
  - ✓ Train the model using each of the n features, and evaluate the performance.
  - ✓ The feature or set of features with the best performance is/are finalized.
  - ✓ Repeat steps until one gets the desired number of features.