Chapter 4: Data Preprocessing

Handling Missing Data:

Dropping Missing Data:

- Pandas **dropna() method** allows the user to analyze and drop Rows/Columns with Null values in different ways.
- **axis:** axis takes int or string value for rows/columns. Input can be 0 or 1 for Integer and ‘index’ or ‘columns’ for String.
- **how:** how takes string value of two kinds only (‘any’ or ‘all’). ‘any’ drops the row/column if ANY value is Null and ‘all’ drops only if ALL values are null.
- **thresh:** thresh takes integer value which tells minimum amount of not Nan data required to keep the row.
- **subset:** It’s an array which limits the dropping process to passed rows/columns through list.
- **inplace:** It is a boolean which makes the changes in data frame itself if True.

Filling Missing Data:

- sklearn.impute.SimpleImputer:
  - It is used for completing missing values.
- strategy:
  - mean ➔ replace missing numeric values using mean value for that column
  - most_frequent ➔ replace string/numeric values using most frequent value along that column.

Handling Categorical Data:

Dummy Columns:

pandas.get_dummies()

It converts categorical data into dummy or indicator variables.

One Hot Coding:

It is used for representing categorical values as binary vectors.