NAME:SATHEESH M

REG NO:422221104036

COLLEGE CODE:4222

COLLEGE NAME:SURYA GROUP OF INSTITUTIONS

DATA PREPROCESSING

Data preprocessing is a predominant step in machine learning to yield highly accurate and insightful results. Greater the quality of data, greater is the reliance on the produced results. **Incomplete, noisy, and inconsistent data** are the properties of large real-world datasets. Data preprocessing helps in increasing the quality of data by filling in missing incomplete data, smoothing noise and resolving inconsistencies.

- **Incomplete data** can occur for a number of reasons. Attributes of interest may not always be available, such as customer information for sales transaction data. Relevant data may not be recorded due to a misunderstanding, or because of equipment malfunctions.

- There are many possible reasons for **noisy data** (having incorrect attribute values). The data collection instruments used may be faulty. There may have been human or computer errors occurring at data entry. Errors in data transmission can also occur. Incorrect data may also result from inconsistencies in naming conventions or data codes used, or inconsistent formats for input fields, such as date.

Import the required libraries

```
import warnings
warnings.filterwarnings('ignore')

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from operator import itemgetter
from sklearn.experimental import enable_i
terative_imputer
from sklearn.impute import IterativeImput
er
from sklearn.preprocessing import Ordinal
Encoder
from category_encoders.target_encoder imp
ort TargetEncoder
from sklearn.preprocessing import Standar
dScaler
```

```
from sklearn.ensemble import (GradientBoo
stingRegressor, GradientBoostingClassifie
r)
import xgboost
```

Load

the dataset for training and testing

```
linkcode
train = pd.read_csv('../input/house-price
s-advanced-regression-techniques/train.cs
v')
test = pd.read_csv('../input/house-prices
-advanced-regression-techniques/test.csv'
)
```

.