Name : Devkumar Biswas

Class: BE(AI&DS)

Div: B

Subject : DMV(CL-I Lab)

Roll no. : BEAD21267

**Assignment No. - 3**

**Problem Statement** : Data Wrangling

Problem Statement: Data Wrangling on Real Estate Market

Dataset: "RealEstate\_Prices.csv"

Description: The dataset contains information about housing prices in a specific real estate market. It includes various attributes such as property characteristics, location, sale prices, and other relevant features. The goal is to perform data wrangling to gain insights into the factors influencing housing prices and prepare the dataset for further analysis or modelling.

Tasks to Perform: 1. Import the "RealEstate\_Prices.csv" dataset. Clean column names by removing spaces, special characters, or renaming them for clarity. 2. Handle missing values in the dataset, deciding on an appropriate strategy (e.g., imputation or removal). 3. Perform data merging if additional datasets with relevant information are available (e.g., neighborhood demographics or nearby amenities). 4. Filter and subset the data based on specific criteria, such as a particular time period, property type, or location. 5. Handle categorical variables by encoding them appropriately (e.g., one-hot encoding or label encoding) for further analysis. 6. Aggregate the data to calculate summary statistics or derived metrics such as average sale prices by neighborhood or property type. 7. Identify and handle outliers or extreme values in the data that may affect the analysis or modeling process.

CODE:-

import pandas as pd

import numpy as np

from matplotlib import pyplot as plt

%matplotlib inline

import matplotlib

matplotlib.rcParams["figure.figsize"] = (20,10)

df1 = pd.read\_csv("Bengaluru\_House\_Data.csv")

df1.head()

df1.shape

df1.columns

df1['area\_type']

df1['area\_type'].unique()

df1['area\_type'].value\_counts()

df2 = df1.drop(['area\_type','society','balcony','availability'],axis='columns')

df2.shape

df2.isnull().sum()

df2.shape

df3 = df2.dropna()

df3.isnull().sum()

df3.shape

df3['size'].unique()

df3['bhk'] = df3['size'].apply(lambda x: int(x.split(' ')[0]))

df3['bhk'] = df3['size'].apply(lambda x: int(x.split(' ')[0]))

df3.head()

df3.bhk.unique()

df3[df3.bhk>20]

df3.total\_sqft.unique()

OUTPUT:-



 



