In [ ]:
```
NAME: HEKARE SAURABH KIRAN
COURSE: CL I ROLL NO.23
CLASS: BE AI&DS
```

In [ ]:
```
# 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
#characteristics, location, sale prices,and other relevant features. The goal is to
#housing prices and prepare the dataset for further analysis or modeling.
# Tasks to Perform:
# 1. Import the "RealEstate_Prices.csv" dataset. Clean column names by removing spa
# 2. Handle missing values in the dataset, deciding on an appropriate strategy (e.g
# 3. Perform data merging if additional datasets with relevant information are avai
# 4. Filter and subset the data based on specific criteria, such as a particular ti
# 5. Handle categorical variables by encoding them appropriately (e.g., one-hot enc
# 6. Aggregate the data to calculate summary statistics or derived metrics such as
# 7. Identify and handle outliers or extreme values in the data that may affect the
```

Imports

In [1]:
```
import pandas as pd
import numpy as np
from sklearn.preprocessing import LabelEncoder
from scipy import stats
```

In [2]:
```
data = pd.read_csv("Mumbai_Property.csv")
```

Data Preprocessing

In [3]:
```
data.head()
```

Out[3]:

	Property_Name	Location	Region	Property_Age	Availability	Area_Tpye	Area_SqFt
0	Omkar Alta Monte	W E Highway Malad East Mumbai	Malad Mumbai	0 to 1 Year	Ready To Move	Super Built Up Area	2900.0
1	T Bhimjyani Neelkanth Woods	Manpada Thane Mumbai	Manpada Thane	1 to 5 Year	Ready To Move	Super Built Up Area	1900.0
2	Legend 1 Pramila Nagar	Dahisar West Mumbai	Dahisar Mumbai	10+ Year	Ready To Move	Super Built Up Area	595.0
3	Unnamed Property	Vidyavihar West Vidyavihar West Central Mumbai...	Central Mumbai	5 to 10 Year	Ready To Move	Built Up Area	1450.0
4	Unnamed Property	176 Cst Road Kalina Mumbai 400098 Santacruz Ea...	Santacruz Mumbai	5 to 10 Year	Ready To Move	Carpet Area	876.0

In [4]:
```
data.tail()
```

Out[4]:

	Property_Name	Location	Region	Property_Age	Availability	Area_Tpye	Area_SqFt
2575	Shagun White Woods	Sector 23 Ulwe Navi Mumbai Mumbai	Ulwe Navi-Mumbai	1 to 5 Year	Ready To Move	Built Up Area	1180.0
2576	Guru Anant	Sector 2 Ulwe Navi Mumbai Mumbai	Ulwe Navi-Mumbai	0 to 1 Year	Ready To Move	Built Up Area	1090.0
2577	Balaji Mayuresh Delta	Ulwe Navi Mumbai Mumbai	Ulwe Navi-Mumbai	1 to 5 Year	Ready To Move	Built Up Area	1295.0
2578	Balaji Mayuresh Delta	Ulwe Navi Mumbai Mumbai	Ulwe Navi-Mumbai	1 to 5 Year	Ready To Move	Built Up Area	1850.0
2579	Gurukrupa Tulsi Heights	Ulwe Navi Mumbai Mumbai	Ulwe Navi-Mumbai	0 to 1 Year	Ready To Move	Built Up Area	1100.0

In [5]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2580 entries, 0 to 2579
Data columns (total 12 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Property_Name  2580 non-null   object 
 1   Location       2580 non-null   object 
 2   Region         2580 non-null   object 
 3   Property_Age   2580 non-null   object 
 4   Availability   2580 non-null   object 
 5   Area_Tpye      2580 non-null   object 
 6   Area_SqFt      2580 non-null   float64
 7   Rate_SqFt      2580 non-null   int64  
 8   Floor_No       2580 non-null   int64  
 9   Bedroom        2580 non-null   int64  
 10  Bathroom       2580 non-null   int64  
 11  Price_Lakh     2580 non-null   float64
dtypes: float64(2), int64(4), object(6)
memory usage: 242.0+ KB
```

In [6]:
```
data.columns
```

Out[6]:
```
Index(['Property_Name', 'Location', 'Region', 'Property_Age', 'Availability',
       'Area_Tpye', 'Area_SqFt', 'Rate_SqFt', 'Floor_No', 'Bedroom',
       'Bathroom', 'Price_Lakh'],
      dtype='object')
```

In [7]:
```
data.describe()
```

Out[7]:

	Area_SqFt	Rate_SqFt	Floor_No	Bedroom	Bathroom	Price_Lakh
count	2580.000000	2.580000e+03	2580.000000	2580.000000	2580.000000	2580.000000
mean	1026.105058	1.911185e+04	8.839535	1.962016	2.066667	174.389806
std	2287.126278	4.076088e+04	8.100081	0.844726	0.749960	369.484393
min	33.570000	8.400000e+01	-1.000000	1.000000	1.000000	13.000000
25%	630.750000	8.791750e+03	3.000000	1.000000	2.000000	67.000000
50%	850.000000	1.378500e+04	6.000000	2.000000	2.000000	111.500000
75%	1156.000000	2.265000e+04	12.000000	2.000000	2.000000	200.000000
max	100000.000000	1.650000e+06	59.000000	6.000000	7.000000	16500.000000

Clean column names by removing spaces, special characters, or renaming them for clarity.

In [8]:
```
data.columns = data.columns.str.replace(' ', '_')
data.columns = data.columns.str.replace('[^a-zA-Z0-9_]', '')
```

In [9]:
```
data = data.rename(columns={'Area_Tpye': 'Area_Type'})
```

In [10]:
```
print(data.Area_Type.unique())
print(data.Availability.unique())
print(data.Property_Age.unique())
```

```
['Super Built Up Area' 'Built Up Area' 'Carpet Area' 'Plot Area']
['Ready To Move' 'Under Construction']
['0 to 1 Year' '1 to 5 Year' '10+ Year' '5 to 10 Year'
 'Under Construction']
```

Handle missing values in the dataset, deciding on an appropriate strategy

In [11]:
```
data.isna().sum()
```

Out[11]:
```
Property_Name    0
Location         0
Region           0
Property_Age     0
Availability     0
Area_Type        0
Area_SqFt        0
Rate_SqFt        0
Floor_No         0
Bedroom          0
Bathroom         0
Price_Lakh       0
dtype: int64
```

No missing values

Filter and subset the data based on specific criteria, such as a particular time period, property type, or location

In [12]:
```
filtered_data_1 = data[(data['Price_Lakh'] >= 2000) & (data['Availability'] == 'Rea
```

In [13]:
```
filtered_data_1
```

Out[13]:

	Property_Name	Location	Region	Property_Age	Availability	Area_Type	Area_SqF
1416	Unnamed Property	Juhu Mumbai South West Mumbai	Juhu Mumbai	10+ Year	Ready To Move	Super Built Up Area	5700.(
2065	White City	005 Kandivali East Mumbai	Kandivali Mumbai	0 to 1 Year	Ready To Move	Super Built Up Area	1000.(

In [14]:
```
filtered_data_2 = data[(data['Area_Type'] == 'Plot Area') & (data['Bedroom'] >2)]
filtered_data_2
```

Out[14]:

	Property_Name	Location	Region	Property_Age	Availability	Area_Type
97	Unnamed Property	Ramdev Park Ramdev Park Mira Road And Beyond M...	Mira Road	10+ Year	Ready To Move	Plot Area
104	Unnamed Property	New Panvel Navi Mumbai Mumbai	Panvel Navi-Mumbai	0 to 1 Year	Ready To Move	Plot Area
183	Unnamed Property	Wada Mumbai Beyond Thane Mumbai	Wada Mumbai	5 to 10 Year	Ready To Move	Plot Area
237	Unnamed Property	O 13 Sector 9 Belapur Navi Mumbai Mumbai	Belapur Navi-Mumbai	10+ Year	Ready To Move	Plot Area
281	Unnamed Property	Sector 2 Airoli Navi Mumbai Mumbai	Airoli Navi-Mumbai	10+ Year	Ready To Move	Plot Area
345	Unnamed Property	Sector 9 Belapur Navi Mumbai Mumbai	Belapur Navi-Mumbai	10+ Year	Ready To Move	Plot Area
521	Unnamed Property	Sector 1 Koparkhairane Sector 1 Koparkhairane ...	Koparkhairane Navi-Mumbai	5 to 10 Year	Ready To Move	Plot Area
567	Unnamed Property	101 Manpada Thane Mumbai	Manpada Thane	5 to 10 Year	Ready To Move	Plot Area
1201	Ravi Gaurav Greens	Mira Road East Mira Road And Beyond Mumbai	Mira Road	5 to 10 Year	Ready To Move	Plot Area
1984	Unnamed Property	Khardi Mumbai Beyond Thane Mumbai	Mumbai Thane	10+ Year	Ready To Move	Plot Area

Handle categorical variables by encoding them appropriately (e.g., one-hot encoding or label encoding) for further analysis.

In [15]:
```
data = pd.get_dummies(data, columns=['Property_Age'], prefix=['Property_Age'])
```

In [16]:
```
label_encoder = LabelEncoder()
```

In [17]:
```
column_to_encode=['Area_Type','Availability']
label_encoder=LabelEncoder()
for col in column_to_encode:
    data[col] = label_encoder.fit_transform(data[col])
```

In [18]:
```
data.head()
```

Out[18]:

	Property_Name	Location	Region	Availability	Area_Type	Area_SqFt	Rate_SqFt	Flo
0	Omkar Alta Monte	W E Highway Malad East Mumbai	Malad Mumbai	0	3	2900.0	17241	
1	T Bhimjyani Neelkanth Woods	Manpada Thane Mumbai	Manpada Thane	0	3	1900.0	12631	
2	Legend 1 Pramila Nagar	Dahisar West Mumbai	Dahisar Mumbai	0	3	595.0	15966	
3	Unnamed Property	Vidyavihar West Vidyavihar West Central Mumbai...	Central Mumbai	0	0	1450.0	25862	
4	Unnamed Property	176 Cst Road Kalina Mumbai 400098 Santacruz Ea...	Santacruz Mumbai	0	1	876.0	39954	

Aggregate the data to calculate summary statistics or derived metrics such as average sale prices by neighborhood or property type.

In [19]:
```
neighborhood_avg_prices = data.groupby('Area_Type')['Price_Lakh'].mean()
```

In [20]:
```
neighborhood_avg_prices

'''
0='Super Built Up Area'
```

```
1='Built Up Area'
2='Carpet Area'
3='Plot Area'
'''
```

Out[20]:
```
"\n0='Super Built Up Area' \n1='Built Up Area' \n2='Carpet Area' \n3='Plot Are
a'\n"
```

In [21]:
```
Region_avg_prices = data.groupby('Region')['Price_Lakh'].mean()
```

In [22]:
```
Region_avg_prices
```

Out[22]:
```
Region
Adaigaon Navi-Mumbai      26.500000
Adharwadi Mumbai          63.680000
Airoli Navi-Mumbai        87.625000
Ambernath Mumbai          36.857143
Ambika Nagar Mumbai       42.000000
                             ...
Village Navi-Mumbai       99.500000
Wada Mumbai               55.000000
Walkeshwar Mumbai       1062.500000
Wayle Nagar               47.000000
Yagna Nagar              221.333333
Name: Price_Lakh, Length: 145, dtype: float64
```

Identify and handle outliers or extreme values in the data that may affect the analysis or modeling process

In [23]:
```
z_scores = np.abs(stats.zscore(data['Price_Lakh']))
z_scores
```

Out[23]:
```
0       0.881426
1       0.177607
2       0.214908
3       0.543052
4       0.475377
          ...
2575    0.141819
2576    0.233857
2577    0.101214
2578    0.009176
2579    0.214908
Name: Price_Lakh, Length: 2580, dtype: float64
```

In [24]:
```
data_outliers = data[(z_scores > 3)]
```

In [25]:
```
data_outliers
```

Out[25]:

	Property_Name	Location	Region	Availability	Area_Type	Area_SqFt	Rate_Sq
39	Swan Lake Apartment	1 101 Khar West Mumbai South West Mumbai	South Mumbai	0	1	2715.0	662
203	Sagar Mahal	Opposite Gopi Birla School And Sheetal Baug Wa...	Walkeshwar Mumbai	0	0	2450.0	673
329	Jolly Maker Apartment	Cuffe Parade South Mumbai Mumbai	South Mumbai	0	0	2135.0	749
605	Hiranandani Gardens Richmond Tower	Hiranandani Gardens Powai Hiranandani Gardens ...	Central Mumbai	0	3	5000.0	330
634	Kalpataru Solitaire	Juhu Mumbai South West Mumbai	Juhu Mumbai	0	3	3000.0	460
635	Kalpataru Solitaire	Juhu Mumbai South West Mumbai	Juhu Mumbai	0	3	2800.0	464
1064	Unnamed Property	Juhu Mumbai South West Mumbai	Juhu Mumbai	0	0	4363.0	366
1067	Unnamed Property	Juhu Mumbai South West Mumbai	Juhu Mumbai	0	3	4200.0	452
1416	Unnamed Property	Juhu Mumbai South West Mumbai	Juhu Mumbai	0	3	5700.0	421
1675	Piramal Aranya	Byculla East Byculla East Mumbai Harbour Mumbai	Mumbai Harbour	0	1	2800.0	491

	Property_Name	Location	Region	Availability	Area_Type	Area_SqFt	Rate_Sq
2065	White City	005 Kandivali East Mumbai	Kandivali Mumbai	0	3	1000.0	16500

11 records had outlier values in Price_Lakh column

To store clean values in the new dataset we will take the values less than the threshold

In [26]:
```
cleaned_data = data[(z_scores < 3)]
```

In [27]:
```
cleaned_data.to_csv('Cleaned_Mumbai_RealEstate_Data.csv', index=False)
```