01.cleaning

August 20, 2024

0.1 1. Import Necessary Libraries

[ ]:
```
import re
import string
import pandas as pd

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

0.1.1 2. Import the Dataset

[ ]:
```
raw_data = pd.read_csv("Raw_Property.csv")
raw_data
```

[ ]:
```
                     Property_Name  \
0                Omkar Alta Monte
1     T Bhimjyani Neelkanth Woods
2          Legend 1 Pramila Nagar
3                  Unnamed Property
4                  Unnamed Property
…                              …
2576           Shagun White Woods
2577                   Guru Anant
2578         Balaji Mayuresh Delta
2579         Balaji Mayuresh Delta
2580       Gurukrupa Tulsi Heights

                                                 Location       Price Rate_SqFt  \
0                      W.E.Highway, Malad East,  Mumbai     5 Crore    17,241
1                                Manpada, Thane, Mumbai   2.4 Crore    12,631
2                                 Dahisar West,  Mumbai      95 Lac    15,966
3     Vidyavihar West, Vidyavihar West, Central Mumb…  3.75 Crore    25,862
4     176 Cst Road, Kalina, Mumbai 400098, Santacruz…   3.5 Crore    39,954
…                                                   …         …       …
2576                Sector 23 Ulwe, Navi-Mumbai, Mumbai  1.22 Crore    10,338
2577                 Sector 2 Ulwe, Navi-Mumbai, Mumbai      88 Lac     8,073
2578                          Ulwe, Navi-Mumbai, Mumbai  1.37 Crore    10,579
2579                          Ulwe, Navi-Mumbai, Mumbai  1.71 Crore     9,243
```

```
2580                          Ulwe, Navi-Mumbai, Mumbai      95 Lac     8,636

                                                 Area_Tpye  Bedroom  Bathroom  \
0     Super Built up area 2900(269.42 sq.m.)Built Up…         3         4
1     Super Built up area 1900(176.52 sq.m.)Built Up…         3         3
2                  Super Built up area 595(55.28 sq.m.)        1         2
3                    Built Up area: 1450 (134.71 sq.m.)        3         3
4                        Carpet area: 876 (81.38 sq.m.)        2         2
…                                                  …       …        …
2576                 Built Up area: 1180 (109.63 sq.m.)        2         2
2577                 Built Up area: 1090 (101.26 sq.m.)        2         2
2578                 Built Up area: 1295 (120.31 sq.m.)        2         2
2579                 Built Up area: 1850 (171.87 sq.m.)        3         3
2580                 Built Up area: 1100 (102.19 sq.m.)        2         2

     Floor_No      Property_Age             Availability
0       14th     0 to 1 Year Old            Ready to move
1        8th     1 to 5 Year Old   Ready to move Property
2        3rd        10+ Year Old            Ready to move
3        1st    5 to 10 Year Old            Ready to move
4        5th    5 to 10 Year Old            Ready to move
…        …                 …                        …
2576     2nd     1 to 5 Year Old            Ready to move
2577    11st     0 to 1 Year Old            Ready to move
2578     6th     1 to 5 Year Old            Ready to move
2579     6th     1 to 5 Year Old            Ready to move
2580     4th     0 to 1 Year Old            Ready to move

[2581 rows x 10 columns]
```

0.1.2 3. Data Preprocessing

3.1 Remove the Unwantend Symbols and Text

[ ]:
```
raw_data["Property_Age"] = raw_data["Property_Age"].str.replace(' Old','')
```

[ ]:
```
raw_data["Rate_SqFt"] = raw_data["Rate_SqFt"].str.replace(',','')
```

[ ]:
```
raw_data["Availability"] = raw_data["Availability"].str.title()
raw_data["Availability"] = raw_data["Availability"].str.replace(' Property','')
raw_data["Availability"] = [i.lstrip() for i in raw_data["Availability"]]
```

[ ]:
```
raw_data.Availability.unique()
```

[ ]:
```
array(['Ready To Move', 'Under Construction'], dtype=object)
```

3.2 Set Colum Area in SqFt

[ ]:
```
area=[]
for i in range(2581):
    clean_sqft = re.sub('[^0-9.]', " ", raw_data["Area_Tpye"][i])
    area.append(clean_sqft.split()[0])
raw_data['Area_SqFt']=area
```

3.3 Set Colum Type of Carpet Area

[ ]:
```
carpet=[]
for i in range(2581):
    clean_carpet = re.sub('[^a-zA-Z]', " ", raw_data['Area_Tpye'][i])
    carpet.append(clean_carpet.split()[0]+' '+clean_carpet.split()[1])
raw_data['Area_Tpye']=carpet
```

[ ]:
```
raw_data['Area_Tpye'] = raw_data.Area_Tpye.str.title()
```

[ ]:
```
raw_data['Area_Tpye'] = raw_data.Area_Tpye.str.replace('Super Built','Super␣
↪Built Up')
raw_data['Area_Tpye'] = raw_data.Area_Tpye.str.replace('Built Up','Built Up␣
↪Area')
raw_data['Area_Tpye'] = raw_data.Area_Tpye.str.replace('Carpet Area ','Carpet␣
↪Area')
```

[ ]:
```
raw_data.Area_Tpye.unique()
```

[ ]:
```
array(['Super Built Up Area', 'Built Up Area', 'Carpet Area', 'Plot Area'],
      dtype=object)
```

3.4 Remove Unwanted Text from Floor No

[ ]:
```
raw_data['Floor_No'] = raw_data.Floor_No.str.replace('Ground','0')
raw_data['Floor_No'] = raw_data.Floor_No.str.replace('Basement','-1')
```

[ ]:
```
floor=[]
for i in range(2581):
    clean_sqft = re.sub('[^0-9-]', "", raw_data["Floor_No"][i])
    floor.append(clean_sqft)
raw_data['Floor_No']=floor
```

3.5 Set Colum Region

[ ]:
```
raw_data["Location"] = [i.lstrip() for i in raw_data["Location"]]
```

[ ]:
```
location=[]
for i in range(2581):
    clean_location = re.sub('[^a-zA-Z-]', " ", raw_data["Location"][i])
    location.append(clean_location)
raw_data['Region']=location
```

[ ]:
```
raw_data['Region'] = raw_data.Region.str.title()
words = ['[0-9]','East','West','South','Suburbs','Sector','Beyond','And␣
 ↪Beyond', 'Scheme']
raw_data["Region"] = raw_data["Region"].str.replace('|'.join(words), '',␣
 ↪regex=True).str.strip()
```

[ ]:
```
raw_data
```

[ ]:
```
                    Property_Name  \
0                Omkar Alta Monte
1     T Bhimjyani Neelkanth Woods
2          Legend 1 Pramila Nagar
3                 Unnamed Property
4                 Unnamed Property
…                              …
2576           Shagun White Woods
2577                  Guru Anant
2578        Balaji Mayuresh Delta
2579        Balaji Mayuresh Delta
2580      Gurukrupa Tulsi Heights

                                                Location       Price Rate_SqFt  \
0                      W.E.Highway, Malad East,  Mumbai     5 Crore     17241
1                                Manpada, Thane, Mumbai   2.4 Crore     12631
2                                 Dahisar West,  Mumbai      95 Lac     15966
3     Vidyavihar West, Vidyavihar West, Central Mumb…  3.75 Crore     25862
4     176 Cst Road, Kalina, Mumbai 400098, Santacruz…   3.5 Crore     39954
…                                                   …         …       …
2576               Sector 23 Ulwe, Navi-Mumbai, Mumbai  1.22 Crore     10338
2577                Sector 2 Ulwe, Navi-Mumbai, Mumbai      88 Lac      8073
2578                         Ulwe, Navi-Mumbai, Mumbai  1.37 Crore     10579
2579                         Ulwe, Navi-Mumbai, Mumbai  1.71 Crore      9243
2580                         Ulwe, Navi-Mumbai, Mumbai      95 Lac      8636

                Area_Tpye  Bedroom  Bathroom Floor_No  Property_Age  \
0     Super Built Up Area        3         4       14   0 to 1 Year
1     Super Built Up Area        3         3        8   1 to 5 Year
2     Super Built Up Area        1         2        3      10+ Year
3           Built Up Area        3         3        1  5 to 10 Year
4             Carpet Area        2         2        5  5 to 10 Year
…                      …      …         …       …           …
2576        Built Up Area        2         2        2   1 to 5 Year
2577        Built Up Area        2         2       11   0 to 1 Year
2578        Built Up Area        2         2        6   1 to 5 Year
2579        Built Up Area        3         3        6   1 to 5 Year
2580        Built Up Area        2         2        4   0 to 1 Year
```

```
       Availability Area_SqFt  \
0     Ready To Move      2900
1     Ready To Move      1900
2     Ready To Move       595
3     Ready To Move      1450
4     Ready To Move       876
…               …       …
2576  Ready To Move      1180
2577  Ready To Move      1090
2578  Ready To Move      1295
2579  Ready To Move      1850
2580  Ready To Move      1100

                                                 Region
0                           W E Highway  Malad    Mumbai
1                                Manpada  Thane  Mumbai
2                                     Dahisar    Mumbai
3     Vidyavihar   Vidyavihar   Central Mumbai   Mumbai
4     Cst Road  Kalina  Mumbai          Santacruz   M…
…                                                  …
2576                           Ulwe  Navi-Mumbai  Mumbai
2577                           Ulwe  Navi-Mumbai  Mumbai
2578                           Ulwe  Navi-Mumbai  Mumbai
2579                           Ulwe  Navi-Mumbai  Mumbai
2580                           Ulwe  Navi-Mumbai  Mumbai

[2581 rows x 12 columns]
```

[ ]:
```
location=[]
for i in range(2581):
    try:
        location.append(raw_data['Region'][i].split()[-3]+' '
 ↪'+raw_data['Region'][i].split()[-2])
    except:
        location.append(raw_data['Region'][i].split()[-2]+' '
 ↪'+raw_data['Region'][i].split()[-1])
raw_data['Region']=location
```

[ ]:
```
raw_data.Region.value_counts().head(30)
```

[ ]:
```
Region
Central Mumbai               226
Mira Road                    201
Kharghar Navi-Mumbai         196
Ulwe Navi-Mumbai             174
Mumbai Thane                 166
Mumbai Harbour               104
```

```
Dombivli Thane                82
Hiranandani-Estate Thane      79
Ghansoli Navi-Mumbai          76
Kamothe Navi-Mumbai           64
Panvel Navi-Mumbai            61
Road Thane                    47
Kandivali Mumbai              44
Manpada Thane                 42
Thane Thane                   41
Koparkhairane Navi-Mumbai     39
Parel Mumbai                  36
Malad Mumbai                  31
Taloja Navi-Mumbai            31
Naupada Thane                 30
Vadavali Thane                29
Balkum Thane                  29
Nagar Thane                   25
Andheri Mumbai                24
Dahisar Mumbai                24
Borivali Mumbai               23
Goregaon Mumbai               22
Thakur Village                21
Kalher Thane                  19
Roadpali Navi-Mumbai          18
Name: count, dtype: int64
```

[ ]:
```
add=[]
for i in range(2581):
    clean_add = re.sub('[^a-zA-Z0-9]', " ", raw_data["Location"][i])
    add.append(clean_add)
raw_data['Location']=add
```

[ ]:
```
raw_data["Location"] = raw_data["Location"].str.replace('    ',' ')
raw_data["Location"] = raw_data["Location"].str.replace('  ',' ')
```

[ ]:
```
raw_data
```

[ ]:
```
                     Property_Name  \
0                Omkar Alta Monte
1     T Bhimjyani Neelkanth Woods
2          Legend 1 Pramila Nagar
3                 Unnamed Property
4                 Unnamed Property
…                              …
2576           Shagun White Woods
2577                  Guru Anant
2578        Balaji Mayuresh Delta
```

```
2579        Balaji Mayuresh Delta
2580      Gurukrupa Tulsi Heights

                                                  Location       Price Rate_SqFt  \
0                           W E Highway Malad East Mumbai     5 Crore     17241
1                                    Manpada Thane Mumbai   2.4 Crore     12631
2                                     Dahisar West Mumbai      95 Lac     15966
3     Vidyavihar West Vidyavihar West Central Mumbai…  3.75 Crore     25862
4     176 Cst Road Kalina Mumbai 400098 Santacruz Ea…   3.5 Crore     39954
…                                                    …         …       …
2576                  Sector 23 Ulwe Navi Mumbai Mumbai  1.22 Crore     10338
2577                   Sector 2 Ulwe Navi Mumbai Mumbai      88 Lac      8073
2578                            Ulwe Navi Mumbai Mumbai  1.37 Crore     10579
2579                            Ulwe Navi Mumbai Mumbai  1.71 Crore      9243
2580                            Ulwe Navi Mumbai Mumbai      95 Lac      8636

                Area_Tpye  Bedroom  Bathroom Floor_No  Property_Age  \
0     Super Built Up Area        3         4       14   0 to 1 Year
1     Super Built Up Area        3         3        8   1 to 5 Year
2     Super Built Up Area        1         2        3      10+ Year
3           Built Up Area        3         3        1  5 to 10 Year
4             Carpet Area        2         2        5  5 to 10 Year
…                    …      …       …      …           …
2576        Built Up Area        2         2        2   1 to 5 Year
2577        Built Up Area        2         2       11   0 to 1 Year
2578        Built Up Area        2         2        6   1 to 5 Year
2579        Built Up Area        3         3        6   1 to 5 Year
2580        Built Up Area        2         2        4   0 to 1 Year

       Availability Area_SqFt            Region
0     Ready To Move      2900     Highway Malad
1     Ready To Move      1900     Manpada Thane
2     Ready To Move       595    Dahisar Mumbai
3     Ready To Move      1450    Central Mumbai
4     Ready To Move       876  Santacruz Mumbai
…               …       …               …
2576  Ready To Move      1180  Ulwe Navi-Mumbai
2577  Ready To Move      1090  Ulwe Navi-Mumbai
2578  Ready To Move      1295  Ulwe Navi-Mumbai
2579  Ready To Move      1850  Ulwe Navi-Mumbai
2580  Ready To Move      1100  Ulwe Navi-Mumbai

[2581 rows x 12 columns]
```

3.6 Replace the all Values in Lac's from Price Column

[ ]:
```
def converter(x):
    if 'Lac' in x:
        return f"{(float(x.strip('Lac'))*1):,.1f}"
    elif 'Crore' in x:
        return f"{(float(x.strip('Crore'))*100):,.1f}"

raw_data['Price_Lakh'] = raw_data['Price'].apply(converter)
raw_data["Price_Lakh"] = raw_data["Price_Lakh"].str.replace(',','')
```

[ ]:
```
raw_data.head()
```

[ ]:
```
                  Property_Name  \
0             Omkar Alta Monte
1  T Bhimjyani Neelkanth Woods
2       Legend 1 Pramila Nagar
3               Unnamed Property
4               Unnamed Property

                                          Location       Price Rate_SqFt  \
0                    W E Highway Malad East Mumbai     5 Crore     17241
1                              Manpada Thane Mumbai   2.4 Crore     12631
2                               Dahisar West Mumbai      95 Lac     15966
3  Vidyavihar West Vidyavihar West Central Mumbai…  3.75 Crore     25862
4  176 Cst Road Kalina Mumbai 400098 Santacruz Ea…   3.5 Crore     39954

             Area_Tpye  Bedroom  Bathroom Floor_No  Property_Age  \
0  Super Built Up Area        3         4       14   0 to 1 Year
1  Super Built Up Area        3         3        8   1 to 5 Year
2  Super Built Up Area        1         2        3      10+ Year
3       Built Up Area        3         3        1  5 to 10 Year
4         Carpet Area        2         2        5  5 to 10 Year

    Availability Area_SqFt           Region Price_Lakh
0  Ready To Move      2900     Highway Malad      500.0
1  Ready To Move      1900     Manpada Thane      240.0
2  Ready To Move       595    Dahisar Mumbai       95.0
3  Ready To Move      1450    Central Mumbai      375.0
4  Ready To Move       876  Santacruz Mumbai      350.0
```

3.7 Check The Null Values and Remove

[ ]:
```
raw_data.isna().sum()
```

[ ]:
```
Property_Name    0
Location         0
Price            0
Rate_SqFt        0
Area_Tpye        0
```

```
Bedroom          0
Bathroom         0
Floor_No         0
Property_Age     0
Availability     0
Area_SqFt        0
Region           0
Price_Lakh       1
dtype: int64
```

[ ]:
```
raw_data.dropna(inplace=True)
raw_data.reset_index(drop=True, inplace=True)
```

[ ]:
```
raw_data = raw_data.to_csv('Property_Location.csv', index=False)
```

3.8 Sort the all Columns

[ ]:
```
raw_data = pd.read_csv('Property_Location.csv')
```

[ ]:
```
raw_data =␣
 ↪raw_data[['Property_Name','Location','Region','Property_Age','Availability','Area_Tpye','Ar
raw_data
```

[ ]:
```
                     Property_Name  \
0                Omkar Alta Monte
1     T Bhimjyani Neelkanth Woods
2          Legend 1 Pramila Nagar
3                  Unnamed Property
4                  Unnamed Property
…                              …
2575           Shagun White Woods
2576                   Guru Anant
2577         Balaji Mayuresh Delta
2578         Balaji Mayuresh Delta
2579      Gurukrupa Tulsi Heights

                                                Location             Region  \
0                          W E Highway Malad East Mumbai      Highway Malad
1                                   Manpada Thane Mumbai      Manpada Thane
2                                    Dahisar West Mumbai     Dahisar Mumbai
3     Vidyavihar West Vidyavihar West Central Mumbai…     Central Mumbai
4     176 Cst Road Kalina Mumbai 400098 Santacruz Ea…   Santacruz Mumbai
…                                                    …                …
2575                  Sector 23 Ulwe Navi Mumbai Mumbai   Ulwe Navi-Mumbai
2576                   Sector 2 Ulwe Navi Mumbai Mumbai   Ulwe Navi-Mumbai
2577                            Ulwe Navi Mumbai Mumbai   Ulwe Navi-Mumbai
2578                            Ulwe Navi Mumbai Mumbai   Ulwe Navi-Mumbai
2579                            Ulwe Navi Mumbai Mumbai   Ulwe Navi-Mumbai
```

```
      Property_Age   Availability            Area_Tpye  Area_SqFt  Rate_SqFt  \
0      0 to 1 Year  Ready To Move  Super Built Up Area     2900.0      17241
1      1 to 5 Year  Ready To Move  Super Built Up Area     1900.0      12631
2         10+ Year  Ready To Move  Super Built Up Area      595.0      15966
3     5 to 10 Year  Ready To Move        Built Up Area     1450.0      25862
4     5 to 10 Year  Ready To Move          Carpet Area      876.0      39954
…              …            …                    …        …        …
2575   1 to 5 Year  Ready To Move        Built Up Area     1180.0      10338
2576   0 to 1 Year  Ready To Move        Built Up Area     1090.0       8073
2577   1 to 5 Year  Ready To Move        Built Up Area     1295.0      10579
2578   1 to 5 Year  Ready To Move        Built Up Area     1850.0       9243
2579   0 to 1 Year  Ready To Move        Built Up Area     1100.0       8636

      Floor_No  Bedroom  Bathroom  Price_Lakh
0           14        3         4       500.0
1            8        3         3       240.0
2            3        1         2        95.0
3            1        3         3       375.0
4            5        2         2       350.0
…          …      …         …          …
2575         2        2         2       122.0
2576        11        2         2        88.0
2577         6        2         2       137.0
2578         6        3         3       171.0
2579         4        2         2        95.0

[2580 rows x 12 columns]
```

3.9 Create a Final Clean CSV File

[ ]:
```
property_mumbai = raw_data.to_csv('Mumbai_Property.csv', index=False)
```

[ ]:
```
property_mumbai = pd.read_csv('Mumbai_Property.csv')
property_mumbai
```

[ ]:
```
                    Property_Name  \
0                Omkar Alta Monte
1     T Bhimjyani Neelkanth Woods
2          Legend 1 Pramila Nagar
3                 Unnamed Property
4                 Unnamed Property
…                              …
2575           Shagun White Woods
2576                  Guru Anant
2577        Balaji Mayuresh Delta
2578        Balaji Mayuresh Delta
2579      Gurukrupa Tulsi Heights
```

```
                                                 Location            Region  \
0                          W E Highway Malad East Mumbai     Highway Malad
1                                    Manpada Thane Mumbai     Manpada Thane
2                                     Dahisar West Mumbai    Dahisar Mumbai
3     Vidyavihar West Vidyavihar West Central Mumbai…    Central Mumbai
4     176 Cst Road Kalina Mumbai 400098 Santacruz Ea…  Santacruz Mumbai
…                                                    …               …
2575                  Sector 23 Ulwe Navi Mumbai Mumbai  Ulwe Navi-Mumbai
2576                   Sector 2 Ulwe Navi Mumbai Mumbai  Ulwe Navi-Mumbai
2577                            Ulwe Navi Mumbai Mumbai  Ulwe Navi-Mumbai
2578                            Ulwe Navi Mumbai Mumbai  Ulwe Navi-Mumbai
2579                            Ulwe Navi Mumbai Mumbai  Ulwe Navi-Mumbai

      Property_Age   Availability            Area_Tpye  Area_SqFt  Rate_SqFt  \
0      0 to 1 Year  Ready To Move  Super Built Up Area     2900.0      17241
1      1 to 5 Year  Ready To Move  Super Built Up Area     1900.0      12631
2         10+ Year  Ready To Move  Super Built Up Area      595.0      15966
3     5 to 10 Year  Ready To Move        Built Up Area     1450.0      25862
4     5 to 10 Year  Ready To Move          Carpet Area      876.0      39954
…              …             …                    …        …        …
2575   1 to 5 Year  Ready To Move        Built Up Area     1180.0      10338
2576   0 to 1 Year  Ready To Move        Built Up Area     1090.0       8073
2577   1 to 5 Year  Ready To Move        Built Up Area     1295.0      10579
2578   1 to 5 Year  Ready To Move        Built Up Area     1850.0       9243
2579   0 to 1 Year  Ready To Move        Built Up Area     1100.0       8636

      Floor_No  Bedroom  Bathroom  Price_Lakh
0           14        3         4       500.0
1            8        3         3       240.0
2            3        1         2        95.0
3            1        3         3       375.0
4            5        2         2       350.0
…          …      …       …         …
2575         2        2         2       122.0
2576        11        2         2        88.0
2577         6        2         2       137.0
2578         6        3         3       171.0
2579         4        2         2        95.0

[2580 rows x 12 columns]
```