house-rent-eda

February 3, 2024

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:
```
df = pd.read_csv('House_Rent_Dataset.csv')
```

[3]:
```
df.head()
```

[3]:

	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
0	2022-05-18	2	10000	1100	Ground out of 2	Super Area	Bandel	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner
1	2022-05-13	2	20000	800	1 out of 3	Super Area	Phool Bagan, Kankurgachi	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
2	2022-05-16	2	17000	1000	1 out of 3	Super Area	Salt Lake City Sector 2	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
3	2022-07-04	2	10000	800	1 out of 2	Super Area	Dumdum Park	Kolkata	Unfurnished	Bachelors/Family	1	Contact Owner
4	2022-05-09	2	7500	850	1 out of 2	Carpet Area	South Dum Dum	Kolkata	Unfurnished	Bachelors	1	Contact Owner

[4]:
```
df.columns
```

[4]:
```
Index(['Posted On', 'BHK', 'Rent', 'Size', 'Floor', 'Area Type',
       'Area Locality', 'City', 'Furnishing Status', 'Tenant Preferred',
       'Bathroom', 'Point of Contact'],
      dtype='object')
```

[4]:
```
df.isnull().sum()
```

[4]:
```
Posted On            0
BHK                  0
Rent                 0
Size                 0
Floor                0
Area Type            0
Area Locality        0
City                 0
Furnishing Status    0
Tenant Preferred     0
Bathroom             0
Point of Contact     0
dtype: int64
```

[5]:
```
df.describe()
```

[5]:

	BHK	Rent	Size	Bathroom
count	4746.000000	4.746000e+03	4746.000000	4746.000000
mean	2.083860	3.499345e+04	967.490729	1.965866
std	0.832256	7.810641e+04	634.202328	0.884532
min	1.000000	1.200000e+03	10.000000	1.000000
25%	2.000000	1.000000e+04	550.000000	1.000000
50%	2.000000	1.600000e+04	850.000000	2.000000
75%	3.000000	3.300000e+04	1200.000000	2.000000
max	6.000000	3.500000e+06	8000.000000	10.000000

[6]:
```
df.corr()
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_7396\1134722465.py:1: FutureWarning:
The default value of numeric_only in DataFrame.corr is deprecated. In a future
version, it will default to False. Select only valid columns or specify the
value of numeric_only to silence this warning.
  df.corr()
```

[6]:

	BHK	Rent	Size	Bathroom
BHK	1.000000	0.369718	0.716145	0.794885
Rent	0.369718	1.000000	0.413551	0.441215
Size	0.716145	0.413551	1.000000	0.740703
Bathroom	0.794885	0.441215	0.740703	1.000000

[7]:
```
df.rename(columns={'Posted On':'Date'}, inplace=True)
```

[8]:
```
df.head(1)
```

[8]:
```
         Date  BHK   Rent  Size            Floor   Area Type Area Locality  \
0  2022-05-18    2  10000  1100  Ground out of 2  Super Area        Bandel

      City Furnishing Status  Tenant Preferred  Bathroom Point of Contact
0  Kolkata       Unfurnished  Bachelors/Family         2    Contact Owner
```

[9]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4746 entries, 0 to 4745
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Date               4746 non-null   object
 1   BHK                4746 non-null   int64
 2   Rent               4746 non-null   int64
 3   Size               4746 non-null   int64
 4   Floor              4746 non-null   object
 5   Area Type          4746 non-null   object
 6   Area Locality      4746 non-null   object
 7   City               4746 non-null   object
 8   Furnishing Status  4746 non-null   object
 9   Tenant Preferred   4746 non-null   object
 10  Bathroom           4746 non-null   int64
 11  Point of Contact   4746 non-null   object
dtypes: int64(4), object(8)
memory usage: 445.1+ KB
```

[10]:
```
df['Date'] = pd.to_datetime(df['Date'])
```

[11]:
```
df['Date'].dtype
```

[11]:
```
dtype('<M8[ns]')
```

[12]:
```
# create a separate column of day and month
df['Day'] = df['Date'].dt.day
df['Month'] = df['Date'].dt.month
```

[13]:
```
df.head(3)
```

[13]:
```
        Date  BHK   Rent  Size            Floor   Area Type  \
0 2022-05-18    2  10000  1100  Ground out of 2  Super Area
1 2022-05-13    2  20000   800       1 out of 3  Super Area
2 2022-05-16    2  17000  1000       1 out of 3  Super Area

              Area Locality     City Furnishing Status  Tenant Preferred  \
0                    Bandel  Kolkata       Unfurnished  Bachelors/Family
```

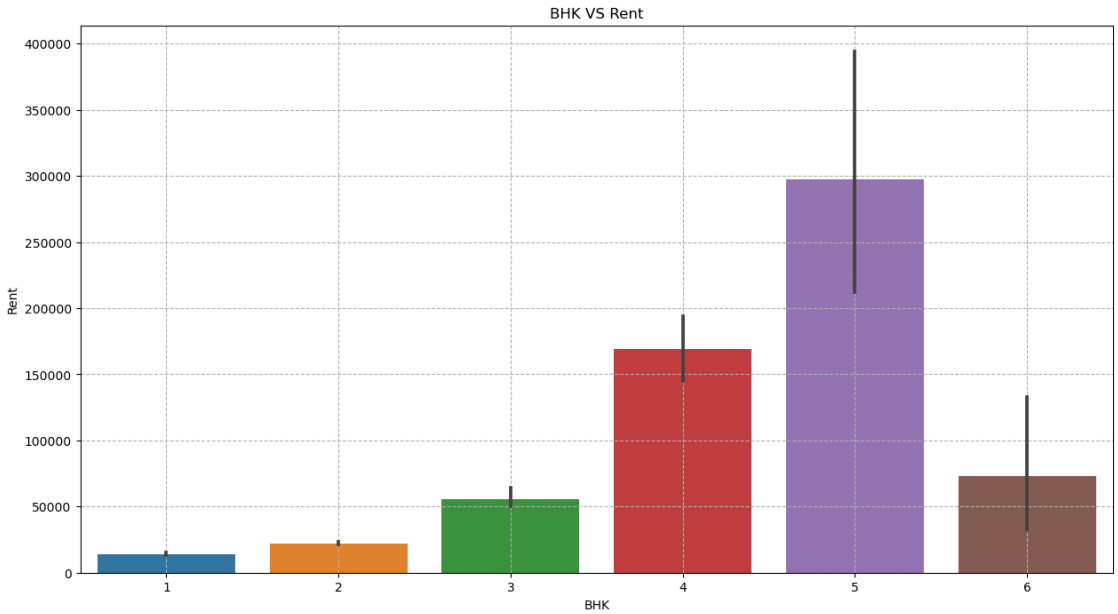
```
1  Phool Bagan, Kankurgachi  Kolkata    Semi-Furnished  Bachelors/Family
2   Salt Lake City Sector 2  Kolkata    Semi-Furnished  Bachelors/Family

   Bathroom Point of Contact  Day  Month
0         2    Contact Owner   18      5
1         1    Contact Owner   13      5
2         1    Contact Owner   16      5
```

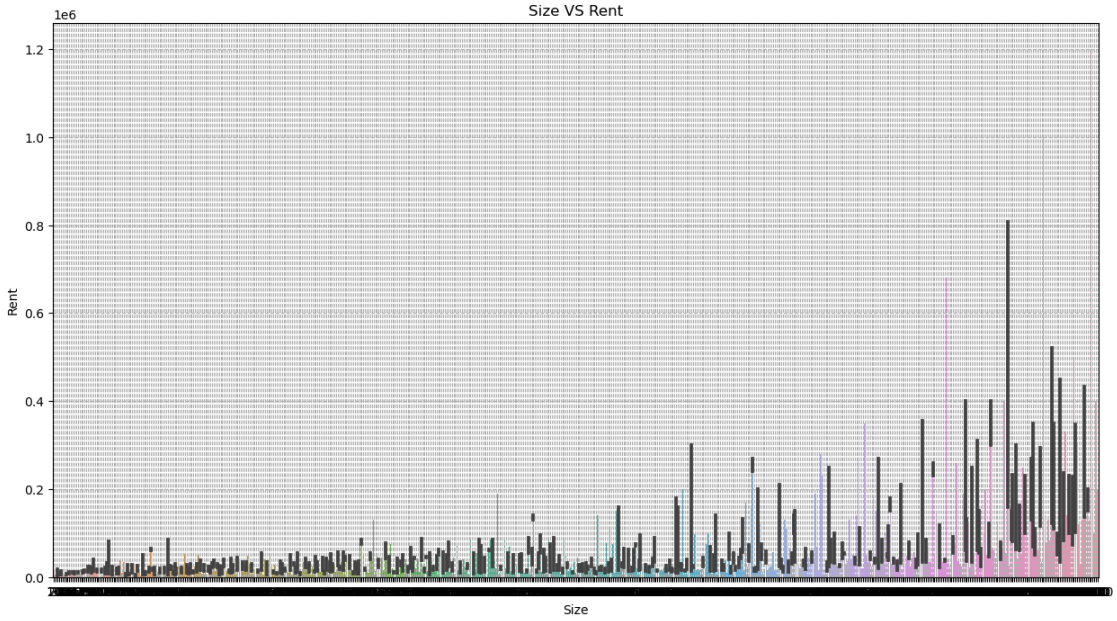
[14]:
```
df.drop('Date', axis=1, inplace=True)
```

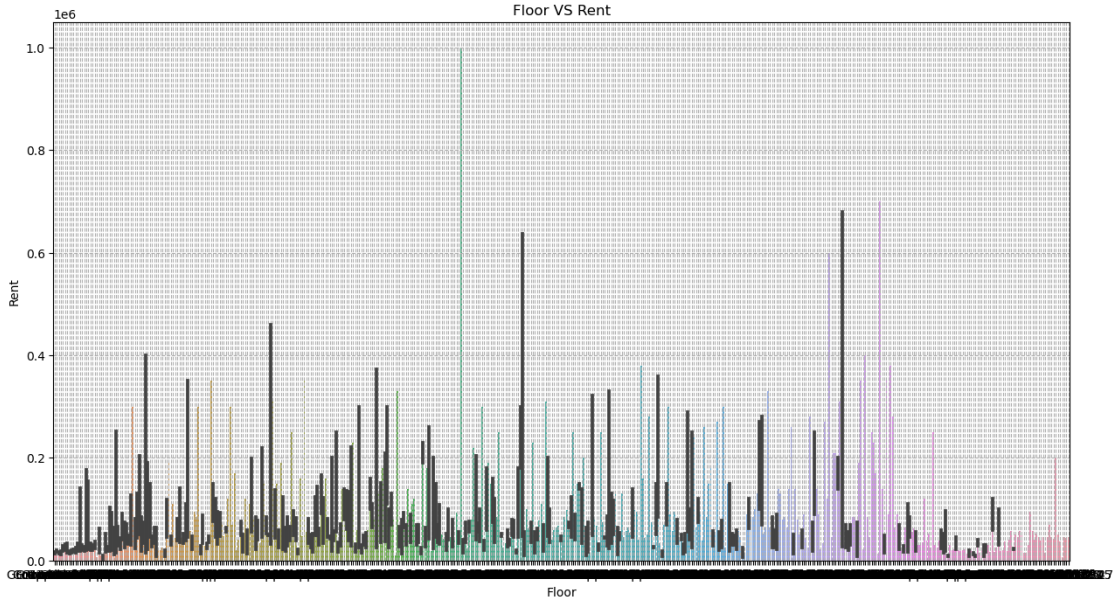
[15]:
```
df.head(1)
```

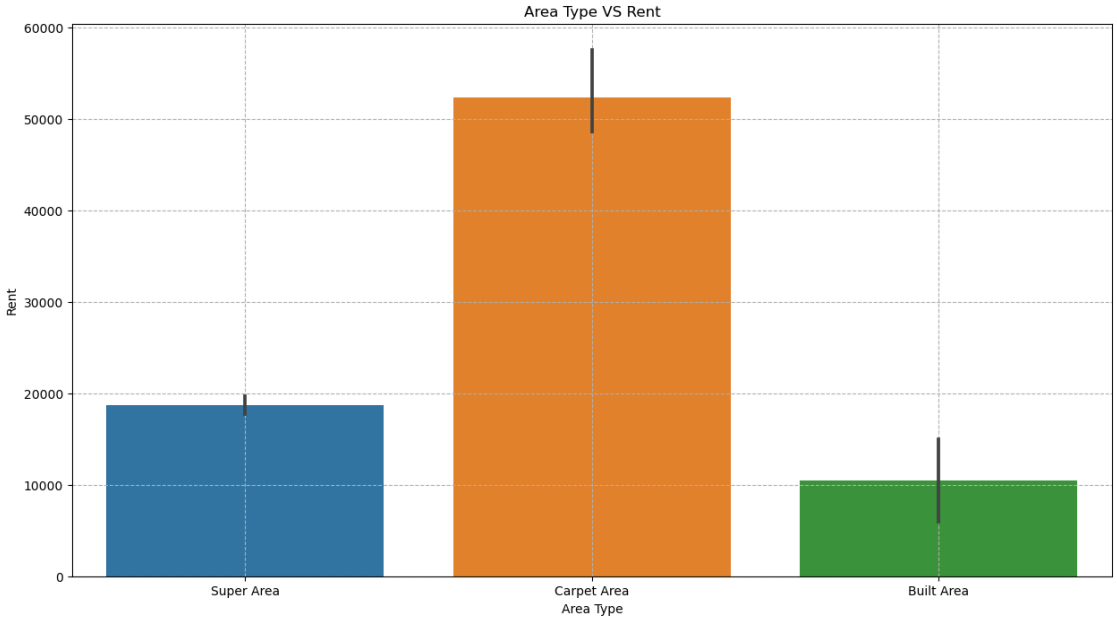
[15]:
```
   BHK   Rent  Size           Floor   Area Type Area Locality     City  \
0    2  10000  1100  Ground out of 2  Super Area        Bandel  Kolkata

  Furnishing Status  Tenant Preferred  Bathroom Point of Contact  Day  Month
0       Unfurnished  Bachelors/Family         2    Contact Owner   18      5
```

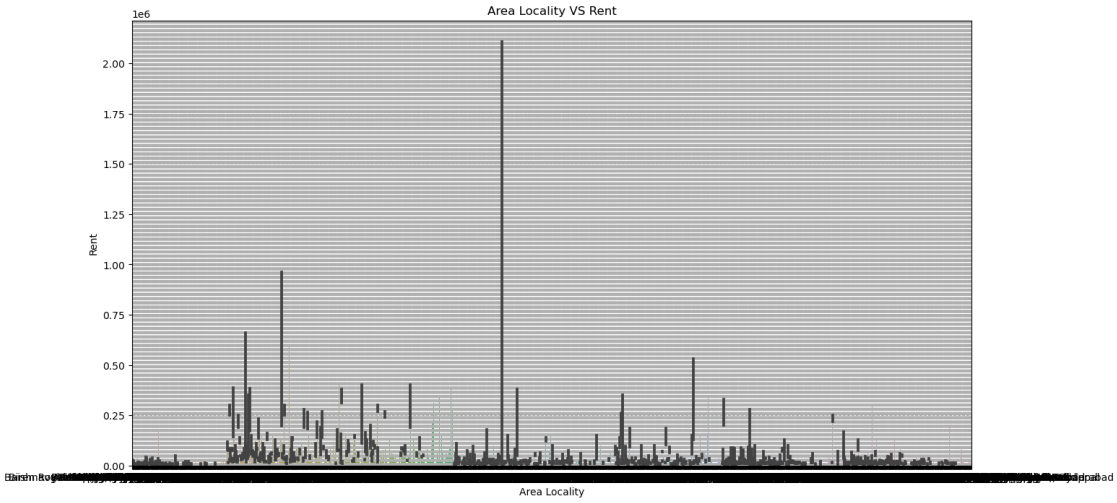
[19]:
```
target_column = 'Rent'
sub_columns = [p for p in df.columns if p != target_column ]
for sub_column in sub_columns:
    plt.figure(figsize=(15,8))
    sns.barplot(x = sub_column, y = target_column, data= df)
    plt.title(f'{sub_column} VS {target_column}')
    plt.grid(linestyle='--')
    plt.show()
```

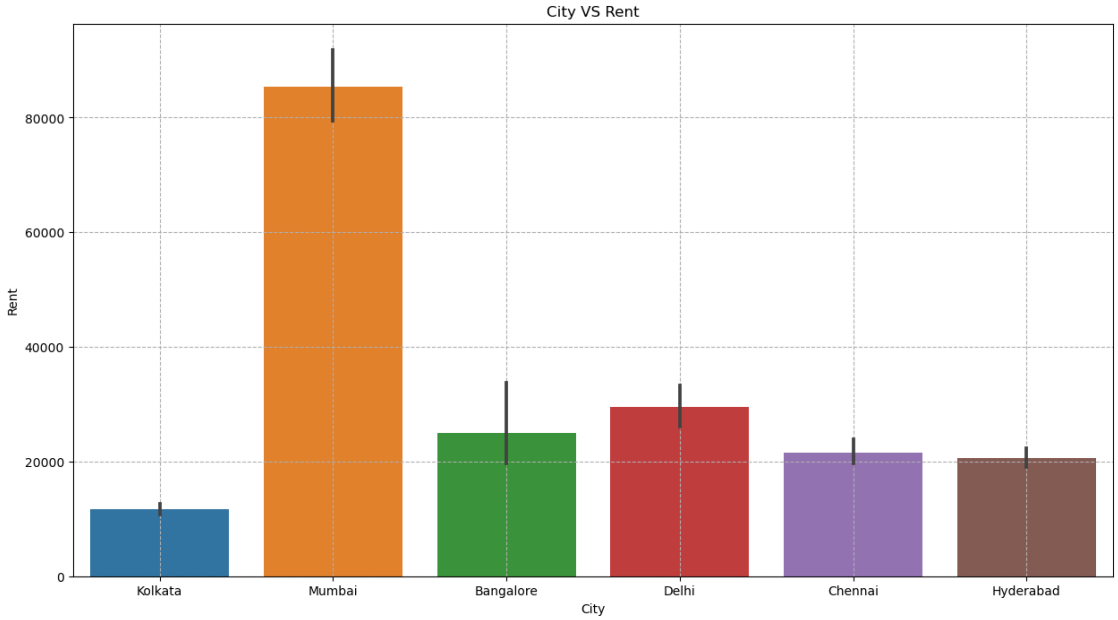


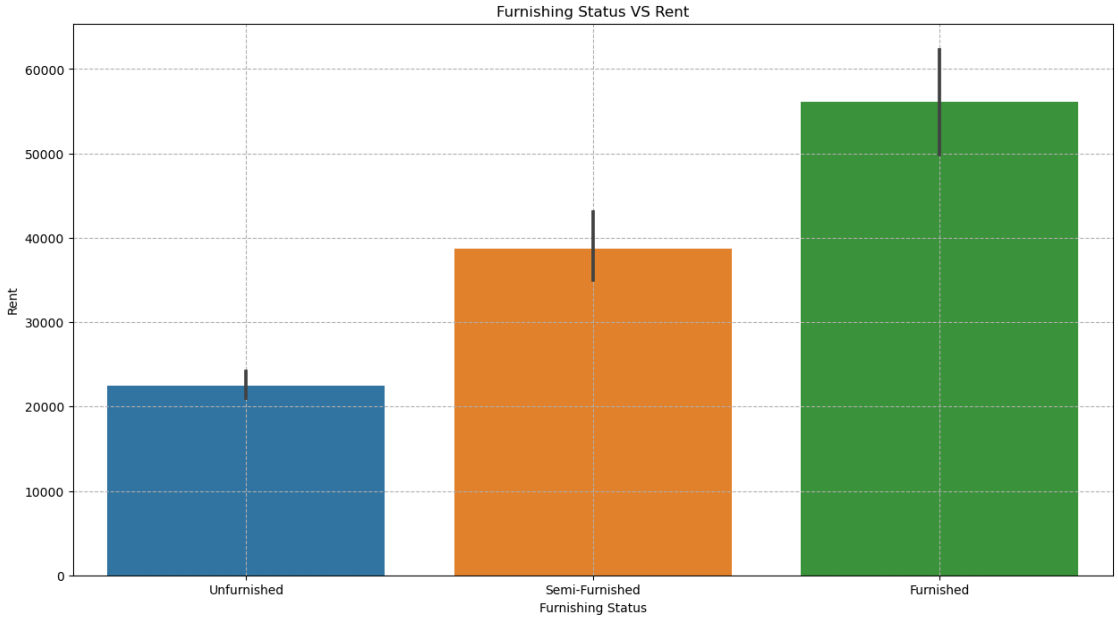

Size VS Rent
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Rent
Size

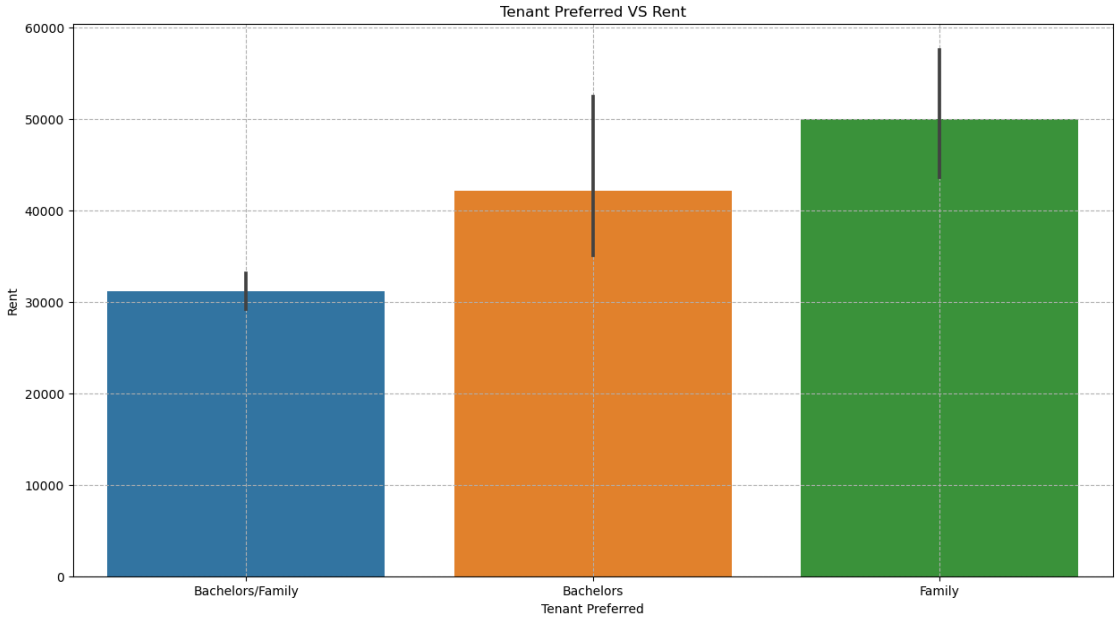

Floor VS Rent
1e6
1.0
0.8
0.6
0.4
0.2
0.0
Rent
Floor

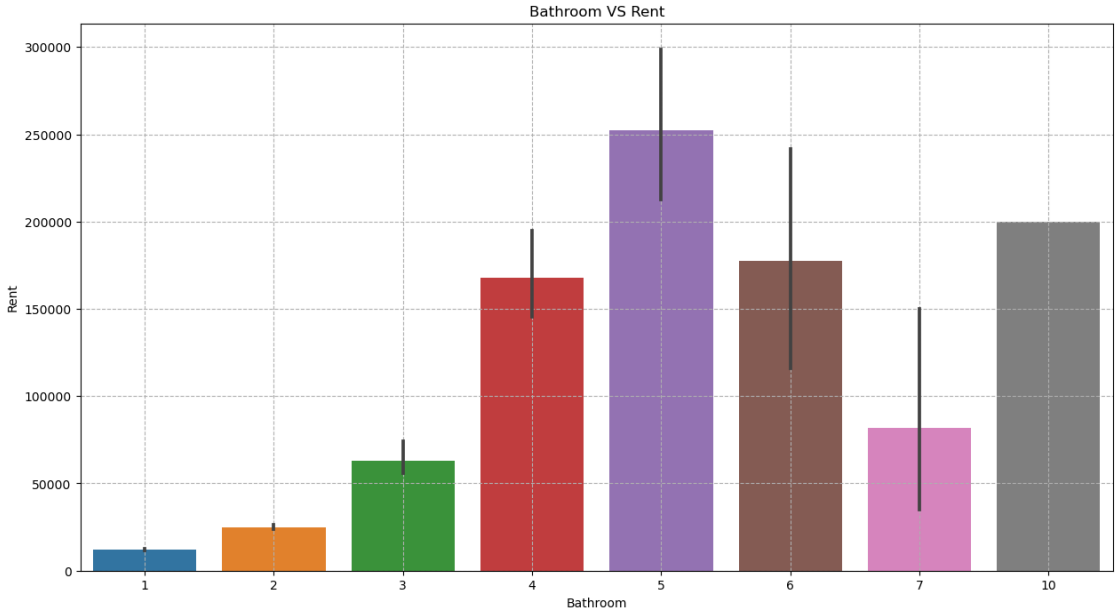

Area Type VS Rent
60000
50000
40000
30000
20000
10000
0
Rent
Super Area
Carpet Area
Built Area
Area Type

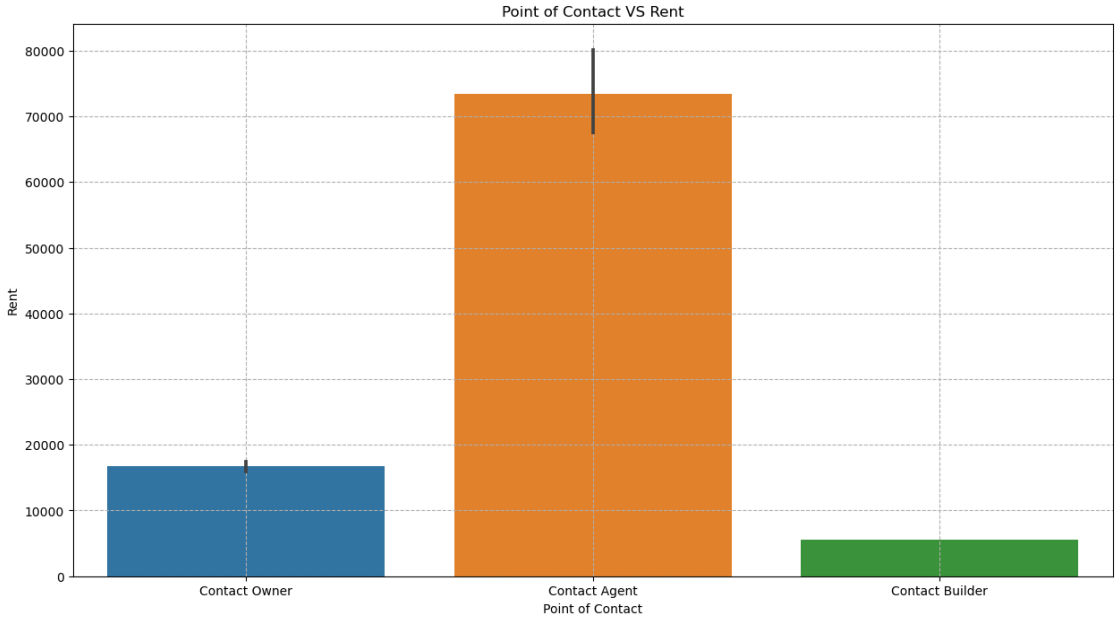

Area Locality VS Rent
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Rent
Area Locality

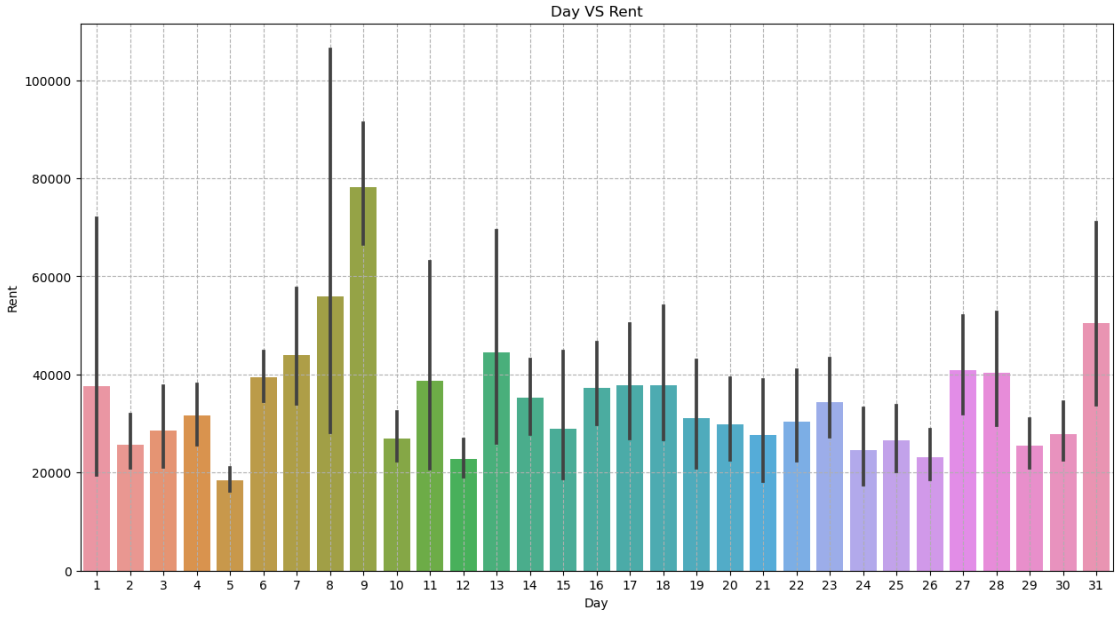

City VS Rent
Rent
0
20000
40000
60000
80000
Kolkata
Mumbai
Bangalore
Delhi
Chennai
Hyderabad
City

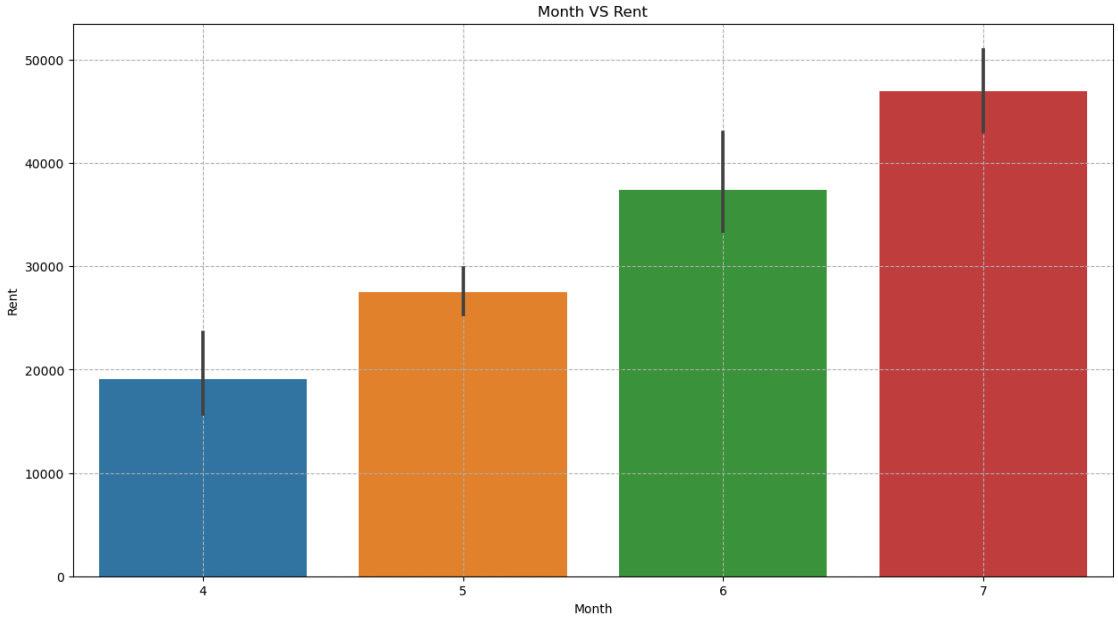

Furnishing Status VS Rent
Rent
0
10000
20000
30000
40000
50000
60000
Unfurnished
Semi-Furnished
Furnished
Furnishing Status


Tenant Preferred VS Rent
Rent
60000
50000
40000
30000
20000
10000
0
Bachelors/Family
Bachelors
Family
Tenant Preferred

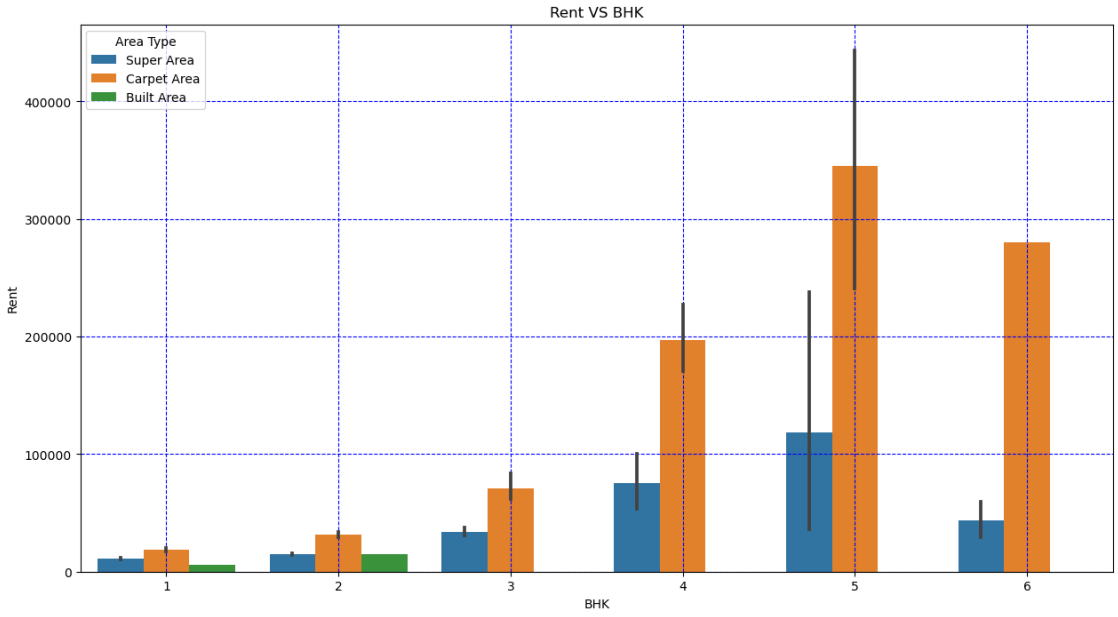

Bathroom VS Rent
Rent
300000
250000
200000
150000
100000
50000
0
1
2
3
4
5
6
7
10
Bathroom

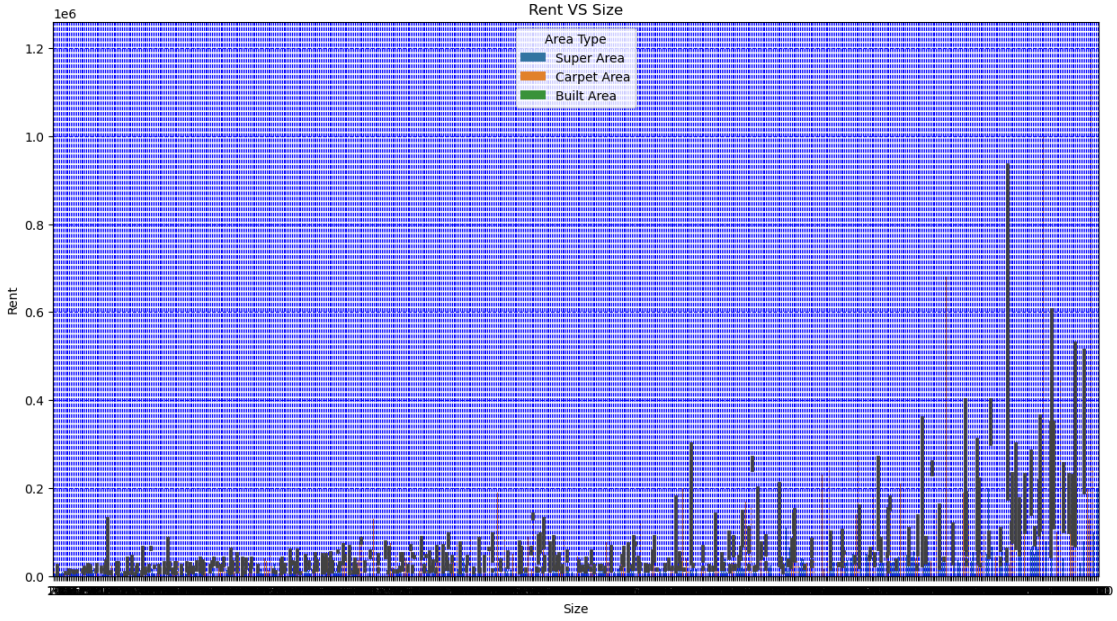

Point of Contact VS Rent
Rent
80000
70000
60000
50000
40000
30000
20000
10000
0
Contact Owner
Contact Agent
Contact Builder
Point of Contact

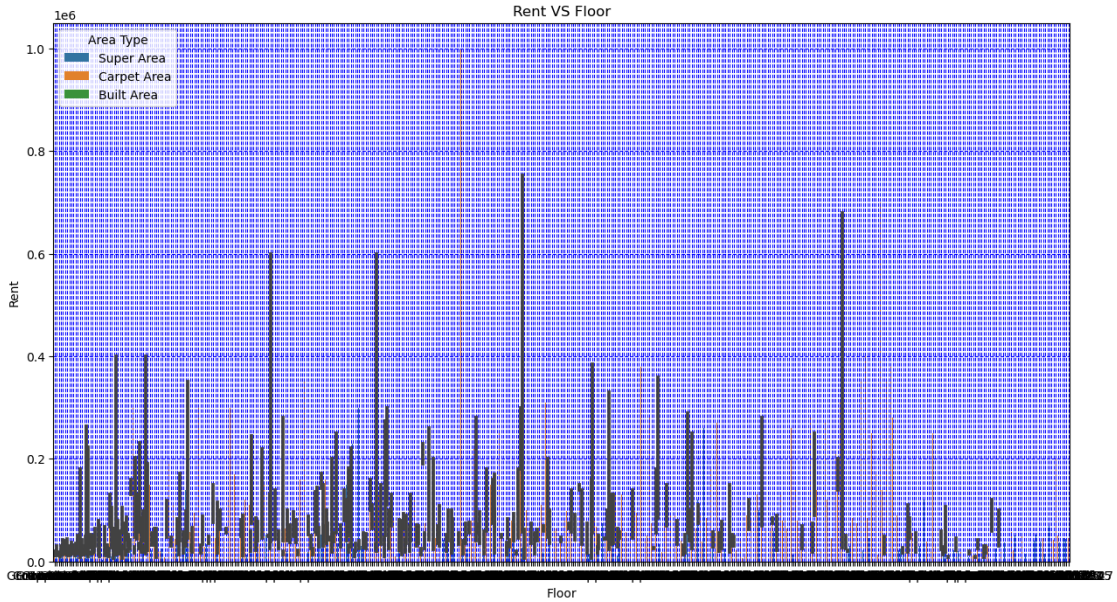

Day VS Rent
Rent
100000
80000
60000
40000
20000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
Day

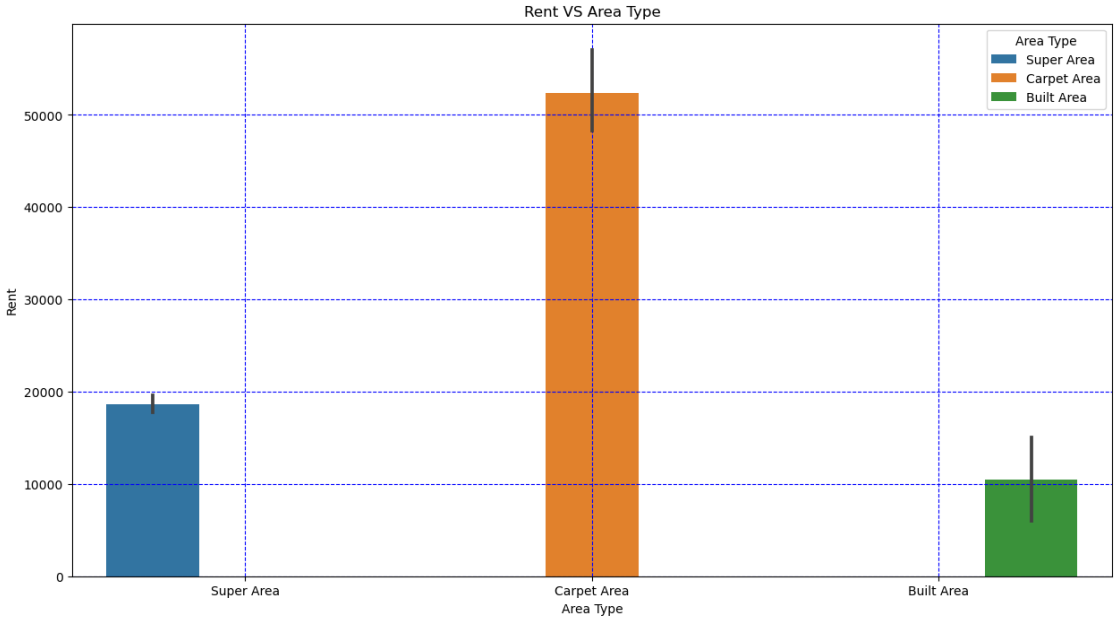


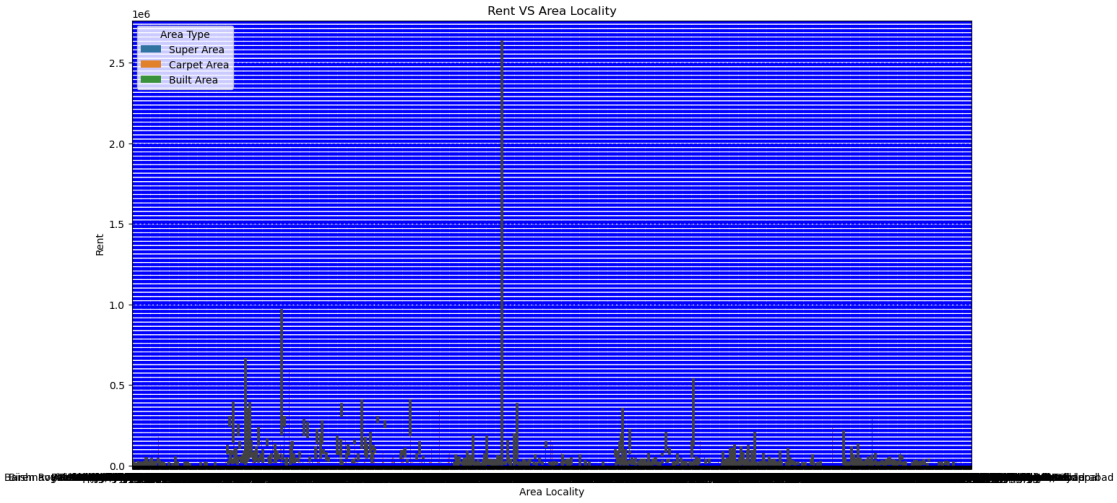
[24]:
```
target_column = 'Rent'
sub_ccolumns = [z for z in df.columns if z != target_column]
for sub_column in sub_ccolumns:
    plt.figure(figsize=(15,8))
    sns.barplot(x = sub_column, y=target_column, data=df, hue = 'Area Type')
    plt.title(f'{target_column} VS {sub_column}')
    plt.grid(True, linestyle='--', color='blue')
    plt.show()
```

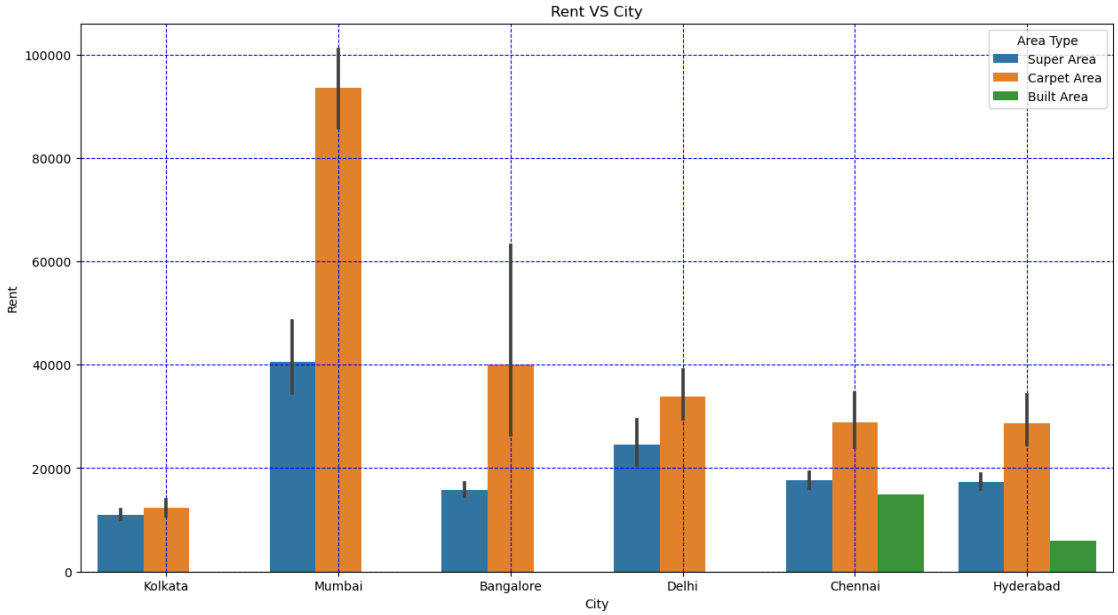


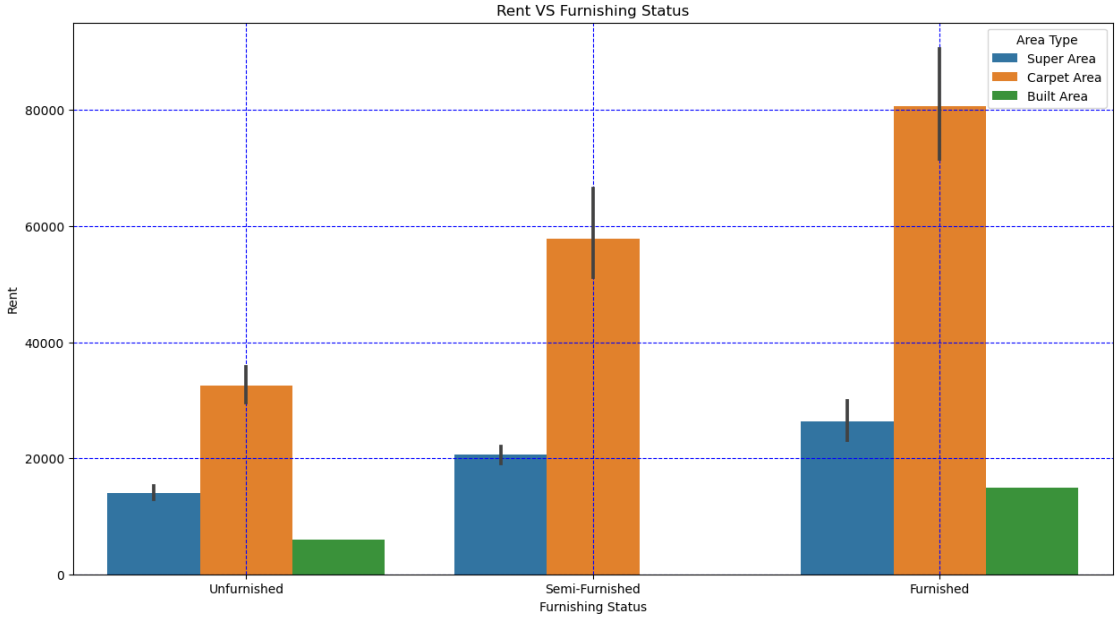

Rent VS Size
1e6
Area Type
Super Area
Carpet Area
Built Area
Rent
Size

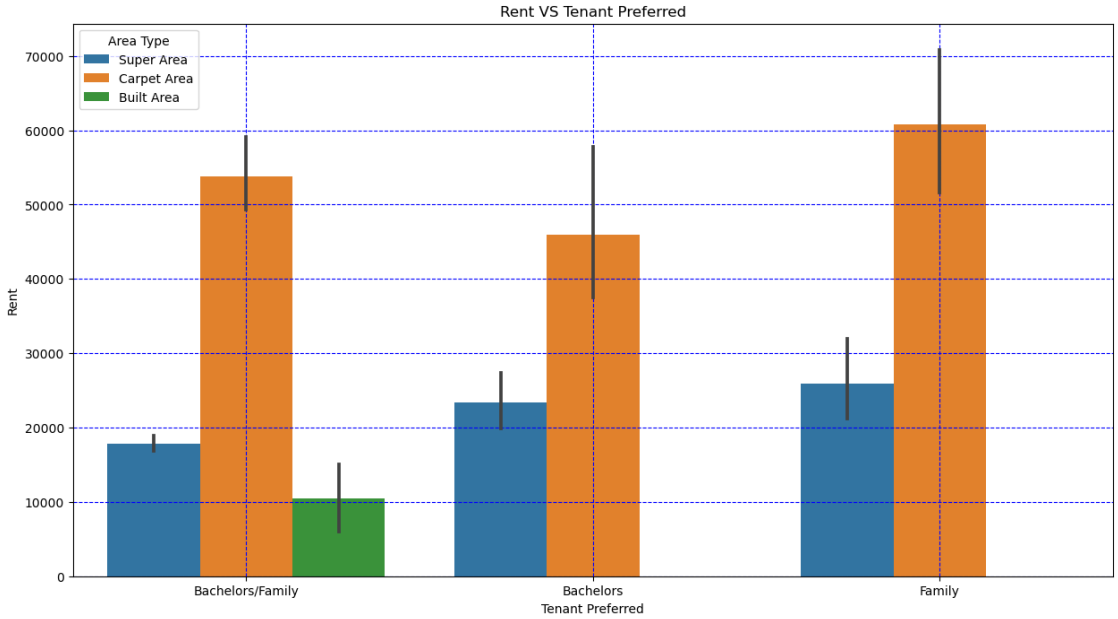

Rent VS Floor
1e6
Area Type
Super Area
Carpet Area
Built Area
Rent
Floor

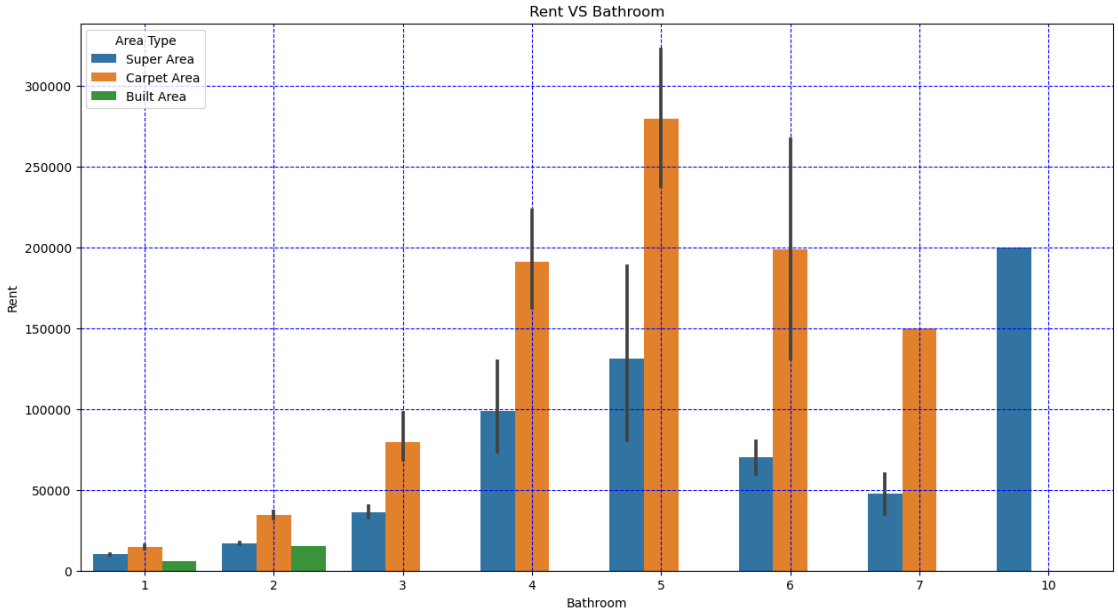

Rent VS Area Type
Area Type
Super Area
Carpet Area
Built Area
Rent
0
10000
20000
30000
40000
50000
Super Area
Carpet Area
Built Area
Area Type

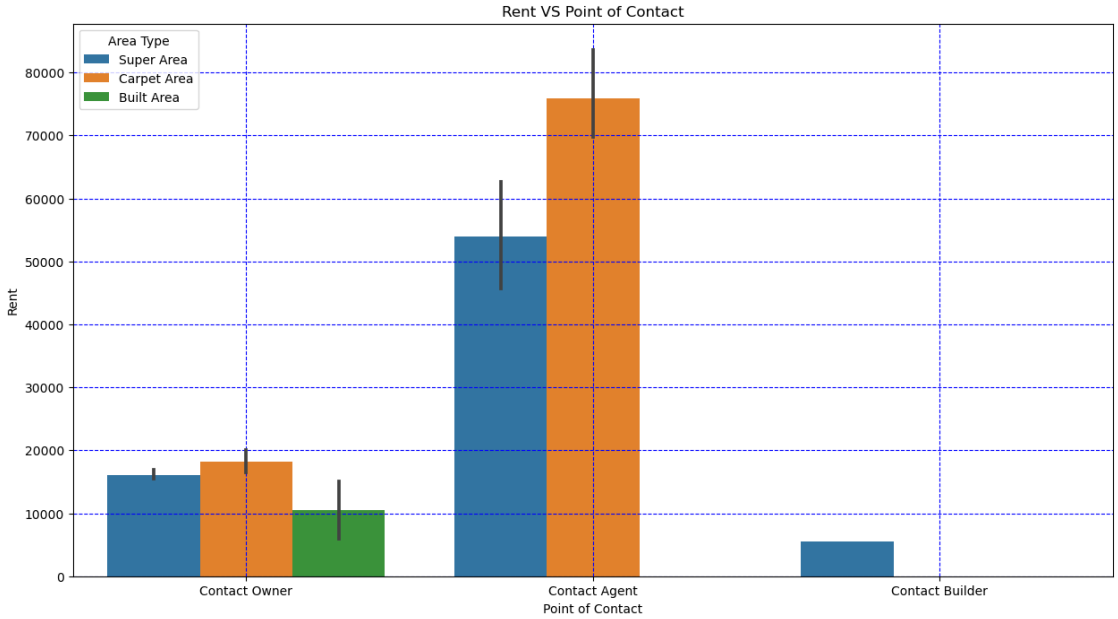

Rent VS Area Locality
1e6
Area Type
Super Area
Carpet Area
Built Area
Rent
0.0
0.5
1.0
1.5
2.0
2.5
Area Locality

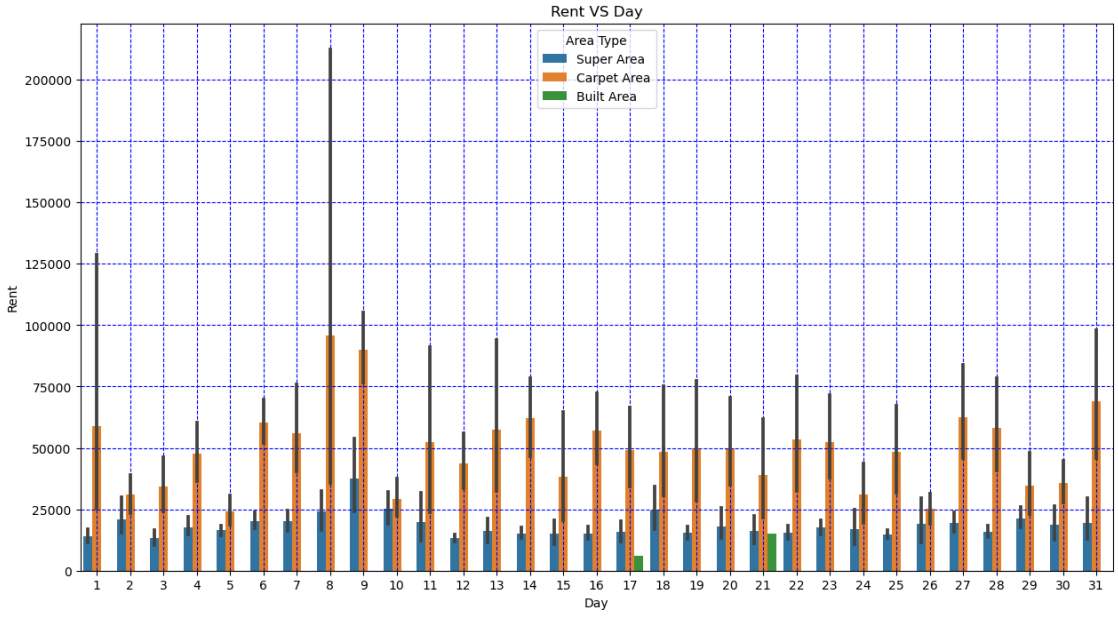

Rent VS City
Area Type
Super Area
Carpet Area
Built Area
Rent
0
20000
40000
60000
80000
100000
Kolkata
Mumbai
Bangalore
Delhi
Chennai
Hyderabad
City

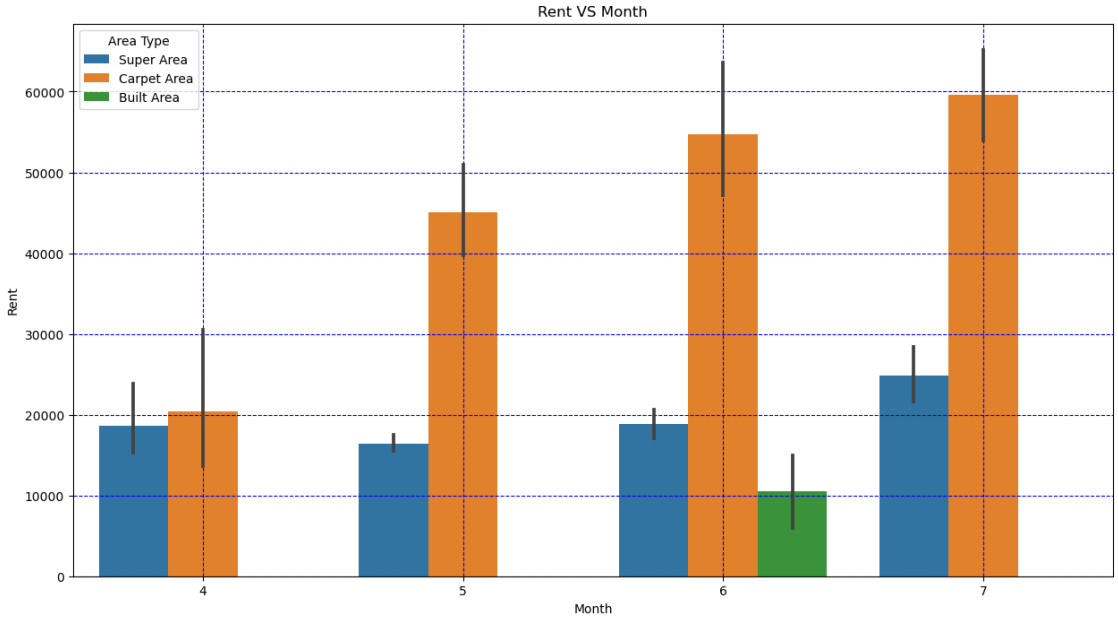

Rent VS Furnishing Status
Area Type
Super Area
Carpet Area
Built Area
Rent
0
20000
40000
60000
80000
Unfurnished
Semi-Furnished
Furnished
Furnishing Status


Rent VS Tenant Preferred
Area Type
Super Area
Carpet Area
Built Area
Rent
0
10000
20000
30000
40000
50000
60000
70000
Bachelors/Family
Bachelors
Family
Tenant Preferred


Rent VS Bathroom
Area Type
Super Area
Carpet Area
Built Area
Rent
0
50000
100000
150000
200000
250000
300000
1
2
3
4
5
6
7
10
Bathroom

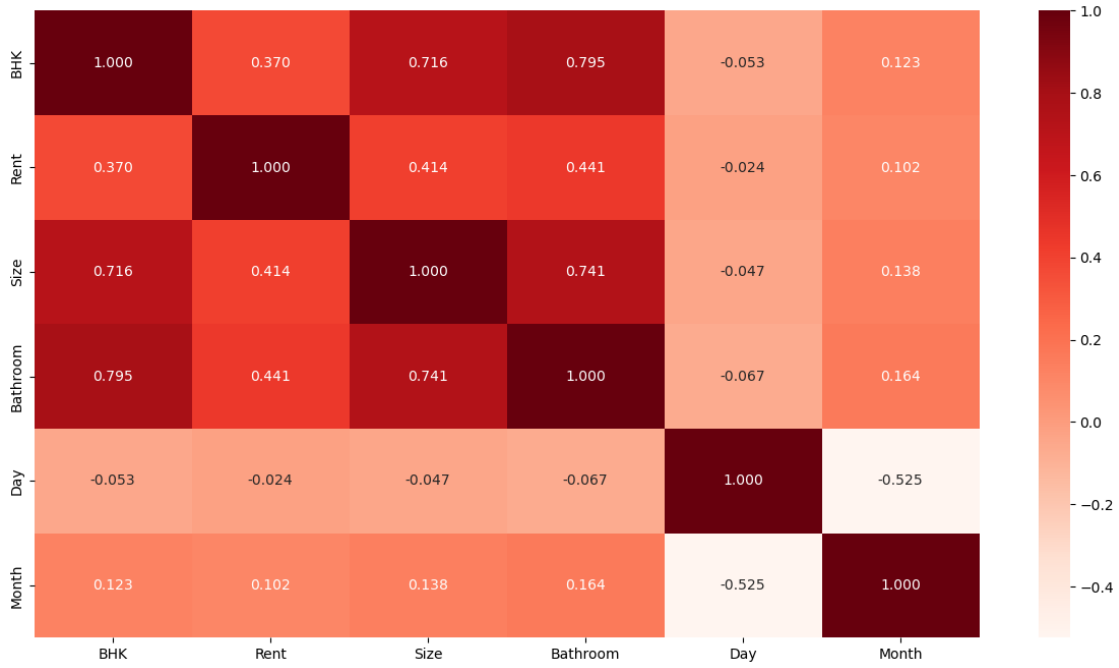

Rent VS Point of Contact
Area Type
Super Area
Carpet Area
Built Area
Rent
0
10000
20000
30000
40000
50000
60000
70000
80000
Contact Owner
Contact Agent
Contact Builder
Point of Contact


Rent VS Day
Area Type
Super Area
Carpet Area
Built Area
Rent
0
25000
50000
75000
100000
125000
150000
175000
200000
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
Day



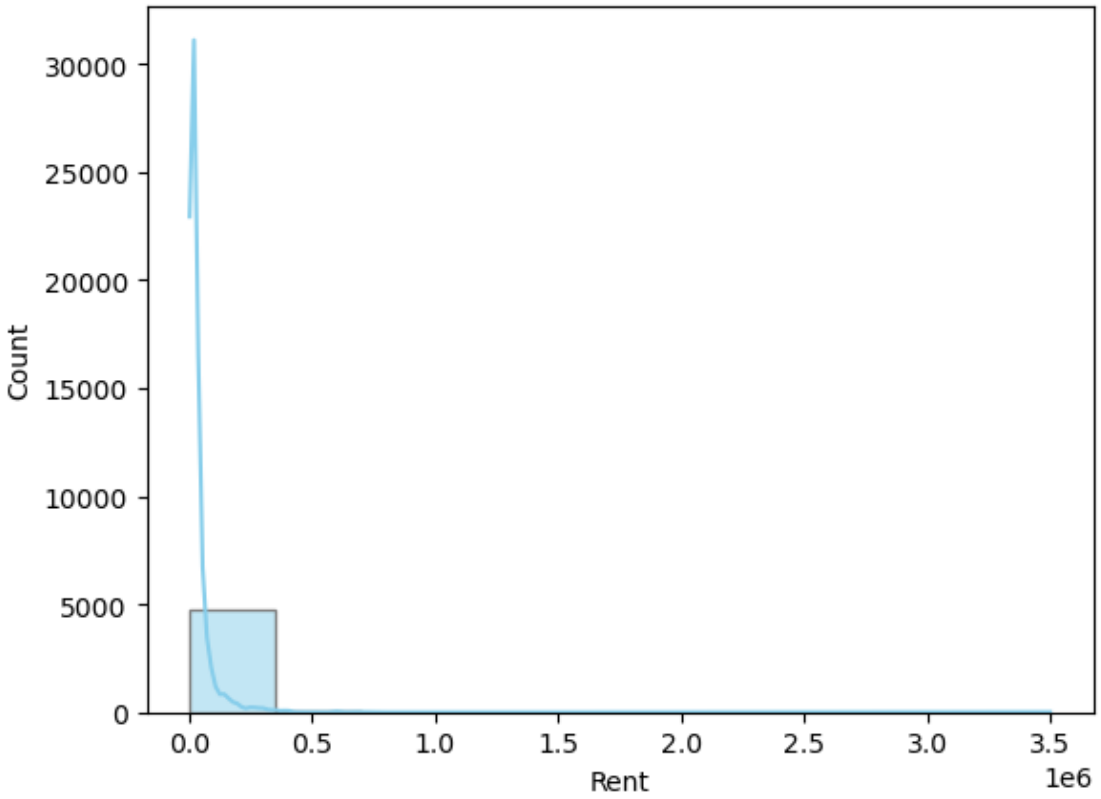
[28]:
```
plt.figure(figsize=(15,8))
sns.heatmap(df.corr(), annot=True, cmap='Reds', fmt='.3f')
plt.show()
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_7396\2229025032.py:2: FutureWarning:
The default value of numeric_only in DataFrame.corr is deprecated. In a future
version, it will default to False. Select only valid columns or specify the
value of numeric_only to silence this warning.
  sns.heatmap(df.corr(), annot=True, cmap='Reds', fmt='.3f')
```

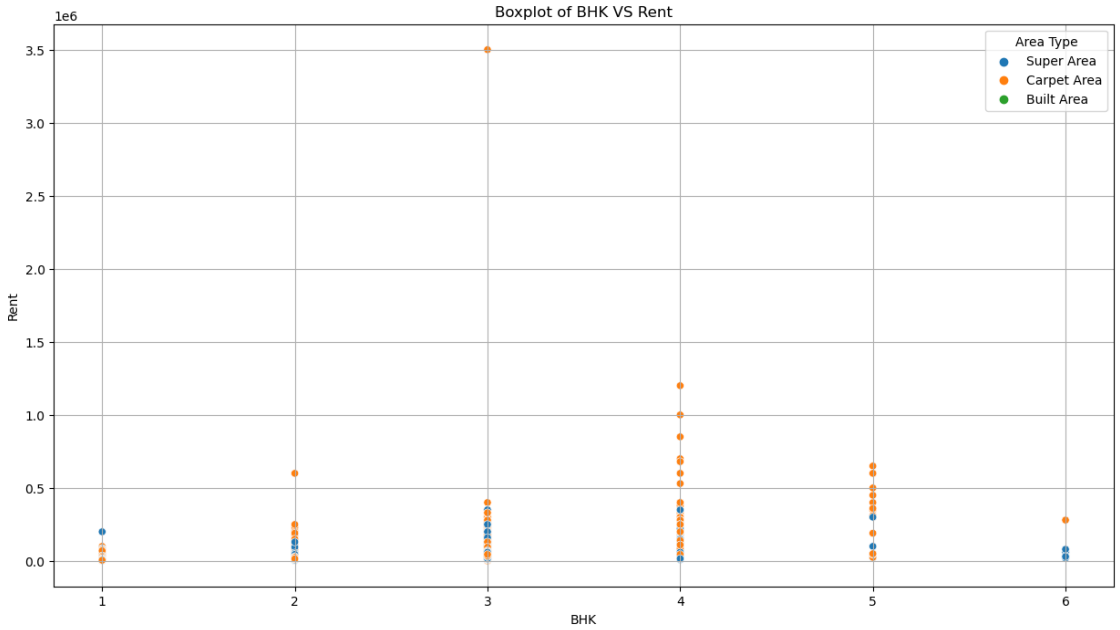


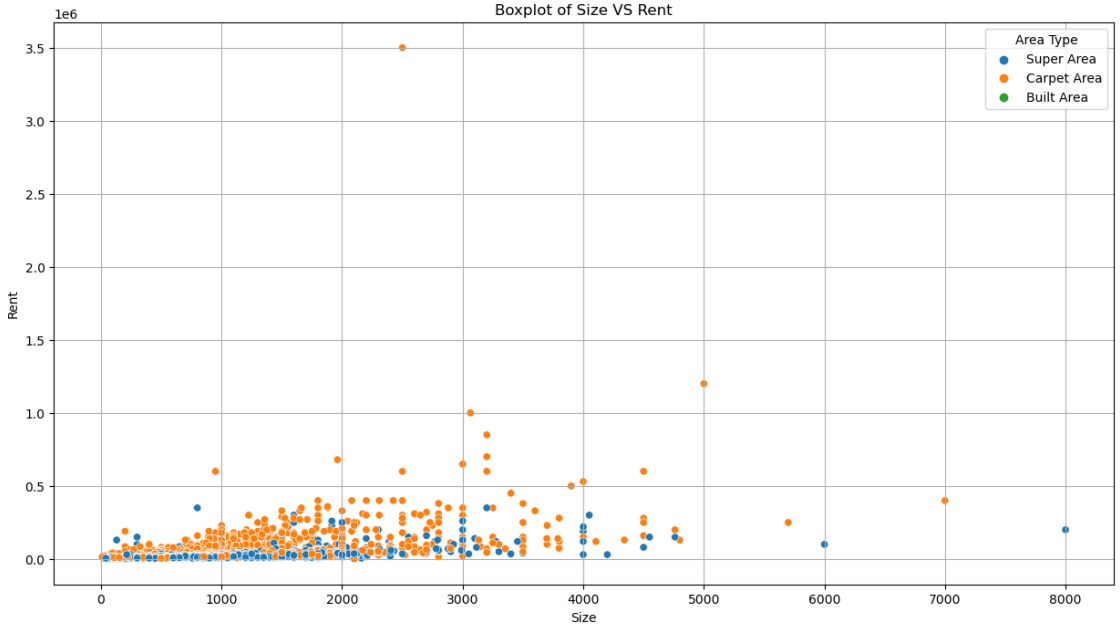
[30]:
```
sns.histplot(df['Rent'], kde=True, bins=10, color='skyblue', edgecolor='grey')
```

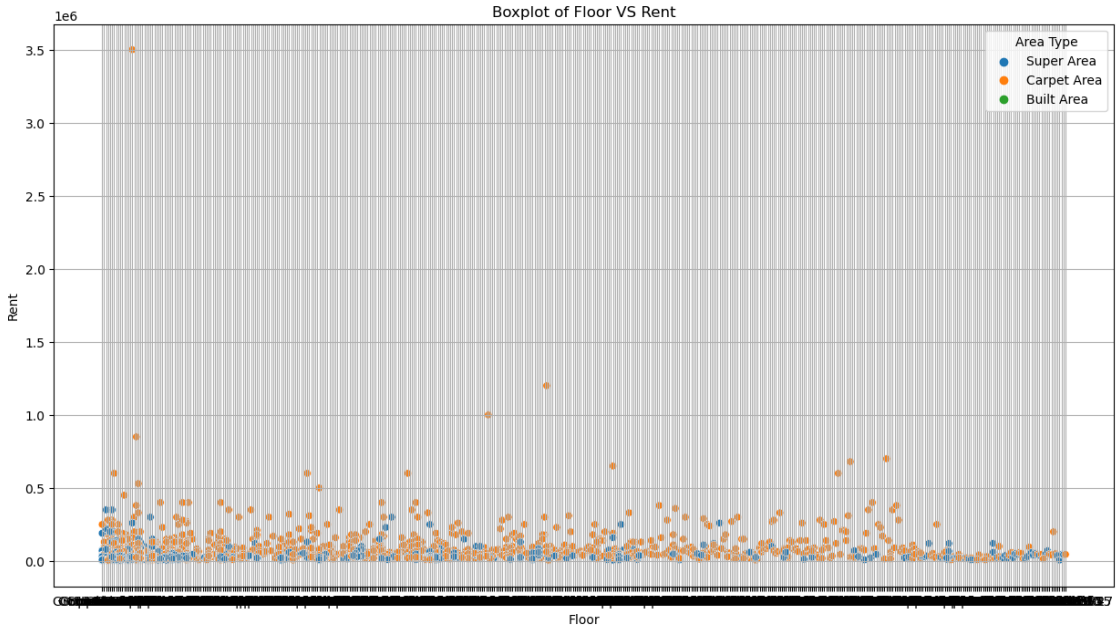
[30]: <Axes: xlabel='Rent', ylabel='Count'>

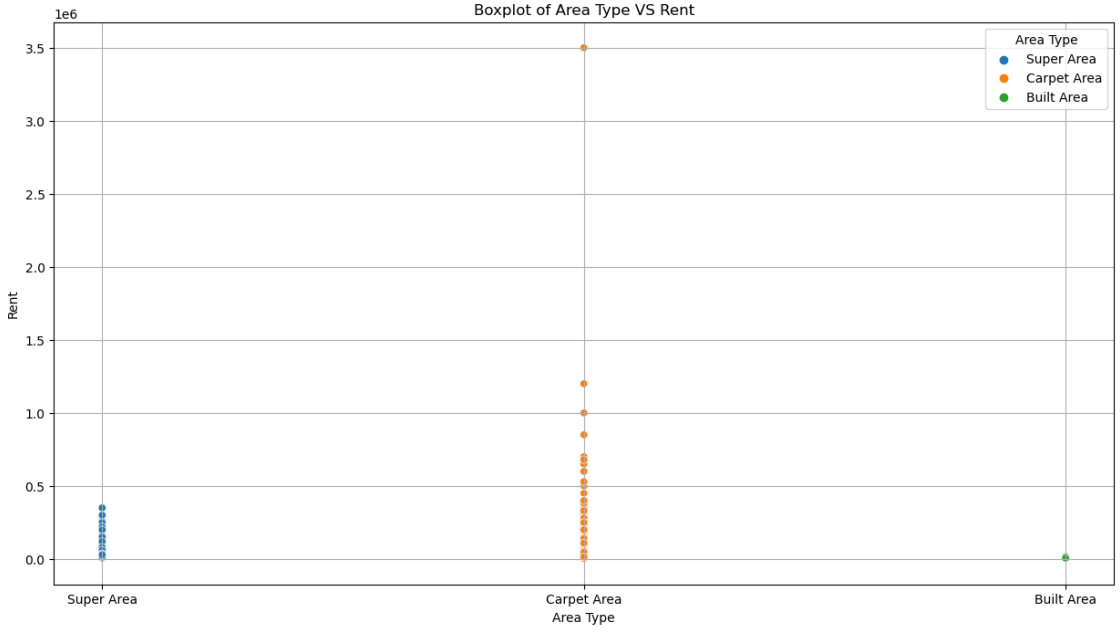


[42]:
```
target_column = 'Rent'
sub_columns = [m for m in df.columns if m != target_column]
for sub_column in sub_columns:
    plt.figure(figsize = (15,8))
    sns.scatterplot(x=sub_column, y = 'Rent', data=df, hue='Area Type')
    plt.title(f'Boxplot of {sub_column} VS {target_column}')
    plt.grid()
    plt.show()
```

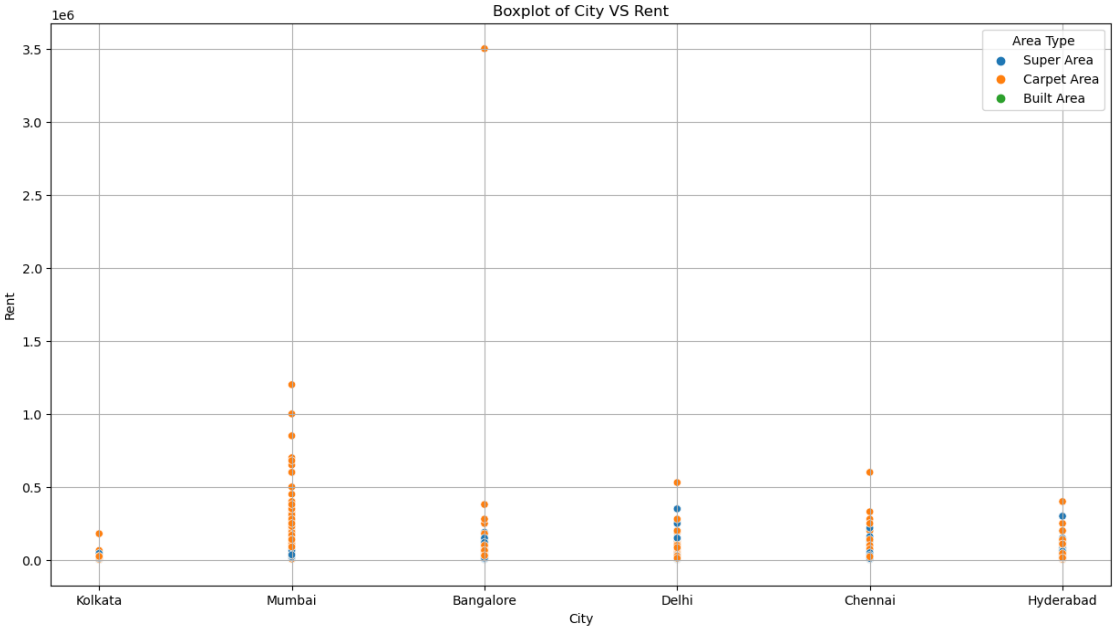

Boxplot of BHK VS Rent
1e6
Area Type
Super Area
Carpet Area
Built Area
Rent
BHK

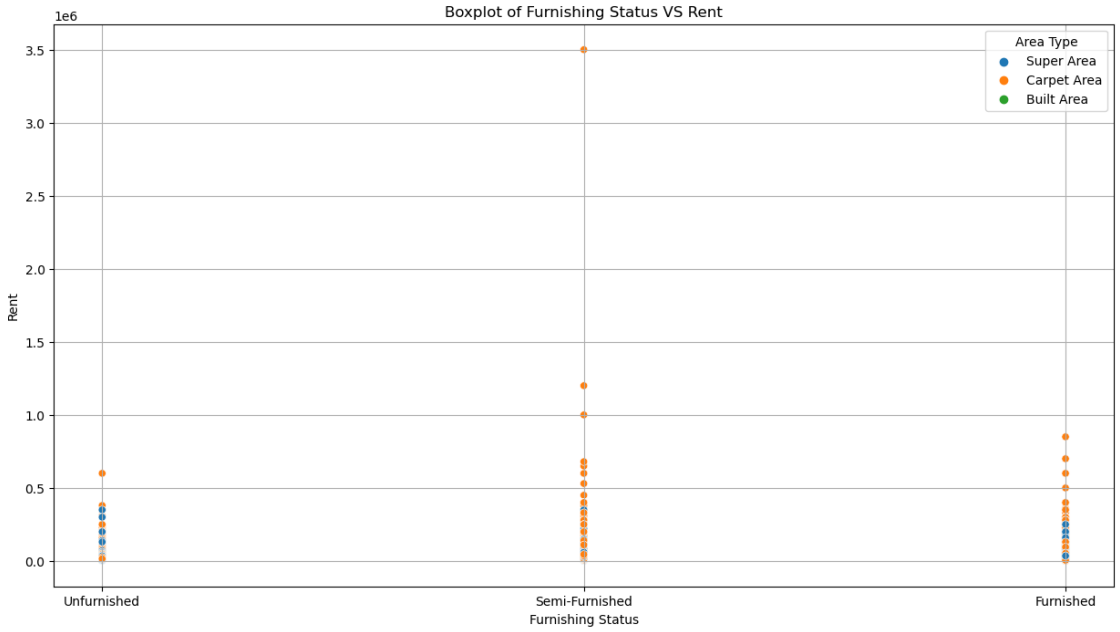

Boxplot of Size VS Rent
1e6
Area Type
Super Area
Carpet Area
Built Area
Rent
Size

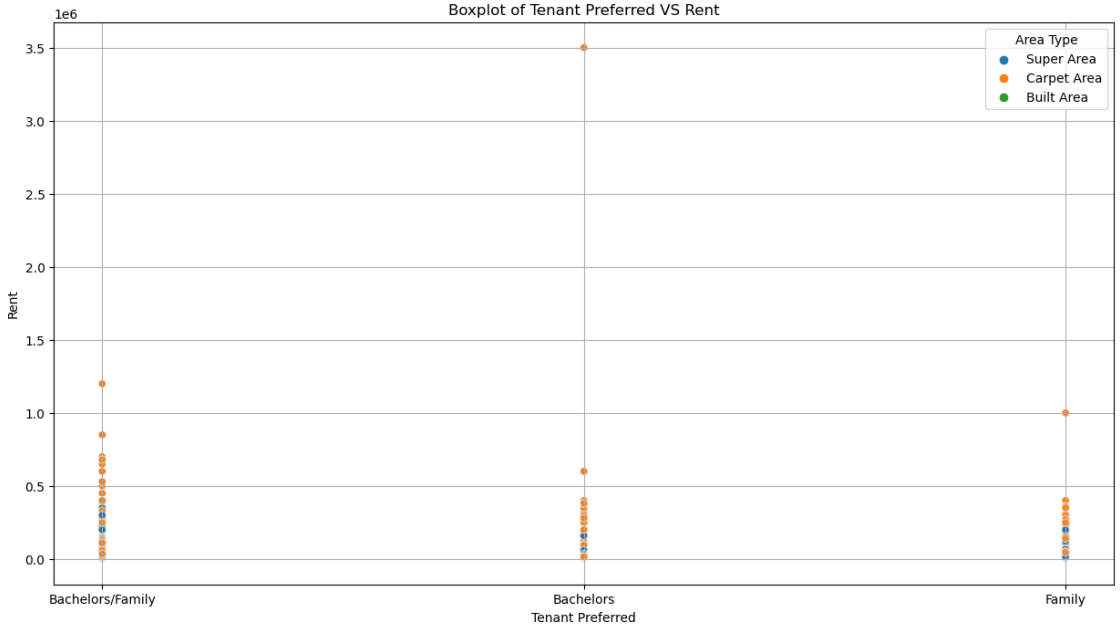

Boxplot of Floor VS Rent
1e6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Rent
Floor
Area Type
Super Area
Carpet Area
Built Area

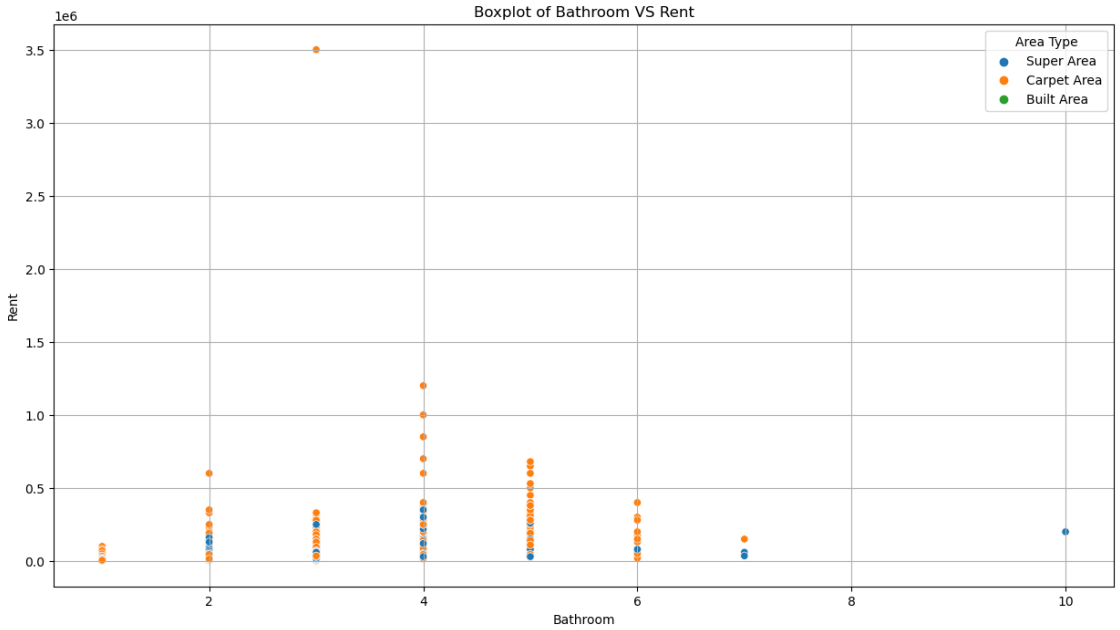

Boxplot of Area Type VS Rent
1e6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Rent
Super Area
Carpet Area
Built Area
Area Type
Area Type
Super Area
Carpet Area
Built Area

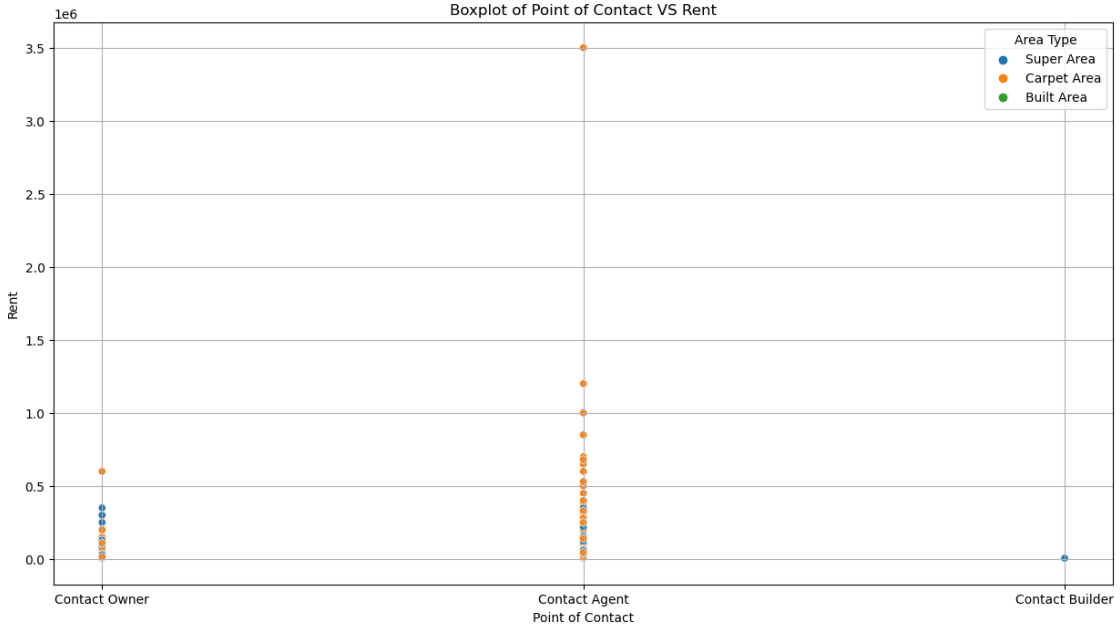

Boxplot of Area Locality VS Rent
1e6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Rent
Area Type
Super Area
Carpet Area
Built Area
Area Locality

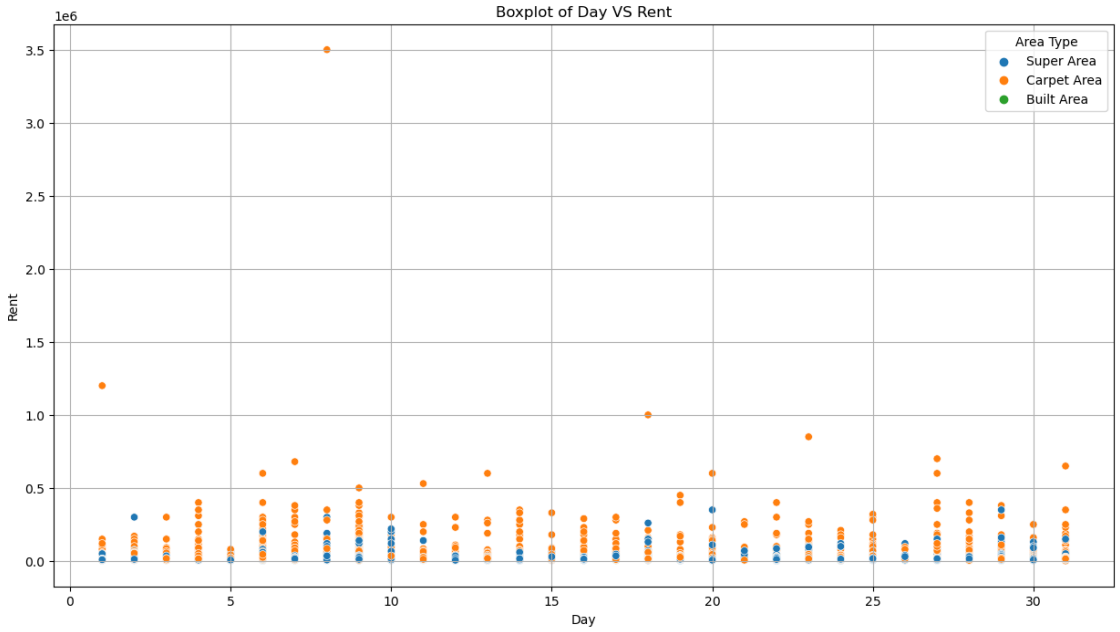

Boxplot of City VS Rent
1e6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Rent
Area Type
Super Area
Carpet Area
Built Area
Kolkata
Mumbai
Bangalore
Delhi
Chennai
Hyderabad
City

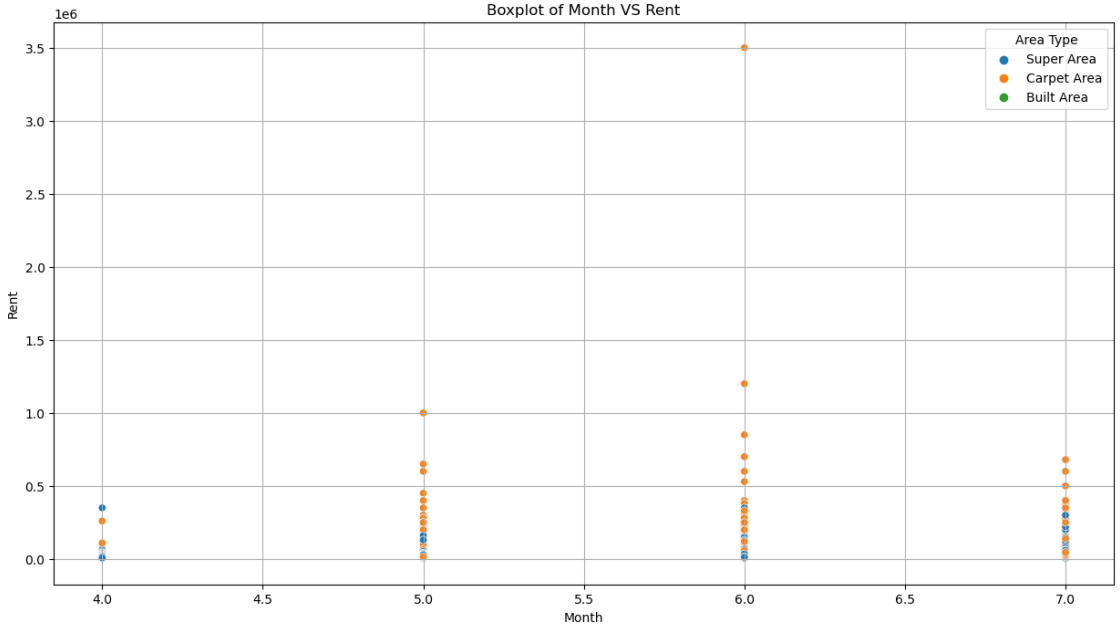

Boxplot of Furnishing Status VS Rent
1e6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Rent
Area Type
Super Area
Carpet Area
Built Area
Unfurnished
Semi-Furnished
Furnished
Furnishing Status


Boxplot of Tenant Preferred VS Rent
1e6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Rent
Area Type
Super Area
Carpet Area
Built Area
Bachelors/Family
Bachelors
Family
Tenant Preferred


Boxplot of Bathroom VS Rent
1e6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Rent
2
4
6
8
10
Bathroom
Area Type
Super Area
Carpet Area
Built Area

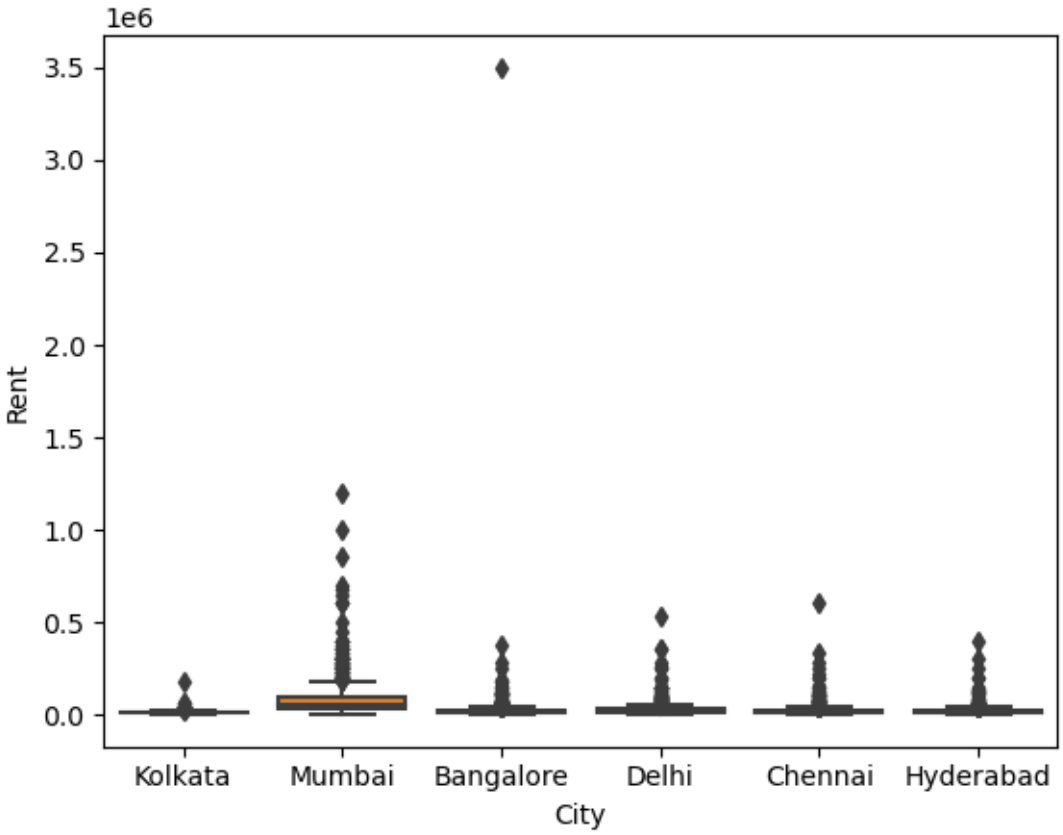

Boxplot of Point of Contact VS Rent
1e6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Rent
Contact Owner
Contact Agent
Contact Builder
Point of Contact
Area Type
Super Area
Carpet Area
Built Area





[37]:
```
sns.boxplot(x='City', y = 'Rent', data=df)
```

[37]: <Axes: xlabel='City', ylabel='Rent'>



[40]:
```
sns.pairplot(df)
```

[40]:
```
<seaborn.axisgrid.PairGrid at 0x1a06cd61490>
```

[43]:
```
df.head()
```

[43]:

	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
0	2	10000	1100	Ground out of 2	Super Area	Bandel	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner
1	2	20000	800	1 out of 3	Super Area	Phool Bagan, Kankurgachi	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
2	2	17000	1000	1 out of 3	Super Area	Salt Lake City Sector 2	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
3	2	10000	800	1 out of 2	Super Area	Dumdum Park	Kolkata	Unfurnished	Bachelors/Family	1	Contact Owner
4	2	7500	850	1 out of 2	Carpet Area	South Dum Dum	Kolkata	Unfurnished	Bachelors/Family	1	Contact Owner

```
3  Kolkata       Unfurnished  Bachelors/Family         1    Contact Owner
4  Kolkata       Unfurnished         Bachelors         1    Contact Owner

   Day  Month
0   18      5
1   13      5
2   16      5
3    4      7
4    9      5
```

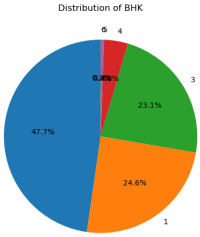
[65]:
```
column_plot = ['BHK','Rent','Size','Floor','Area Type','Area Locality',
               'City','Furnishing Status','Tenant Preferred','Bathroom','Point␣
 ↪of Contact','Day','Month']
column_plot
```

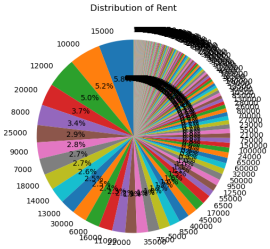
[65]:
```
['BHK',
 'Rent',
 'Size',
 'Floor',
 'Area Type',
 'Area Locality',
 'City',
 'Furnishing Status',
 'Tenant Preferred',
 'Bathroom',
 'Point of Contact',
 'Day',
 'Month']
```

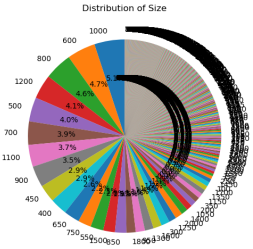
[76]:
```
#create number of columns for subplots
num_columns=2
num_rows = (len(column_plot)//num_columns) + (len(column_plot) % num_columns >␣
 ↪0)

#create subplots
fig, axes = plt.subplots(num_rows, num_columns, figsize = (15,5 * num_rows))
fig.suptitle('Pie Charts of column Distributions',y=1.02)

# Easier for indexing
axes = axes.flatten()


for x, column in enumerate(column_plot):
    column_count = df[column].value_counts()

    sns.color_palette('pastel')
```

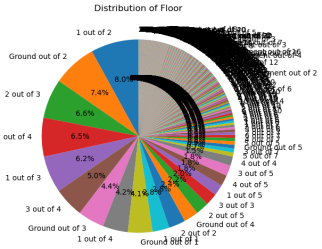
```
    axes[x].pie(column_count, labels = column_count.index, autopct = '%1.1f%%',␣
 ↪startangle=90)
    axes[x].set_title(f'Distribution of {column}')

plt.tight_layout()
plt.show()
```

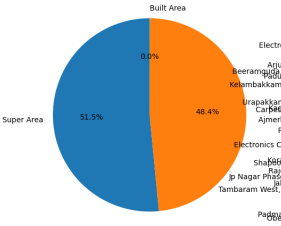
Pie Charts of column Distributions

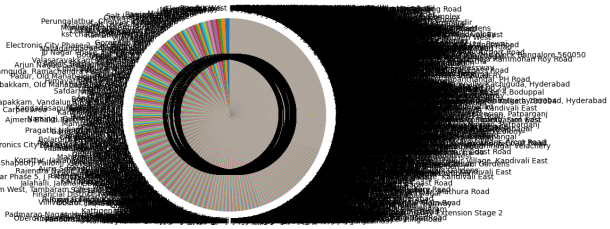


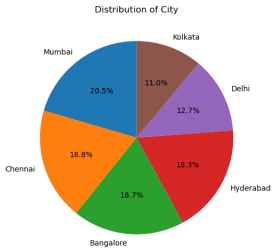


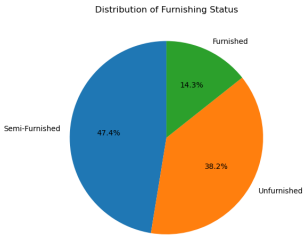


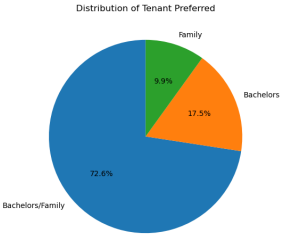


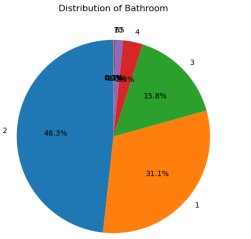


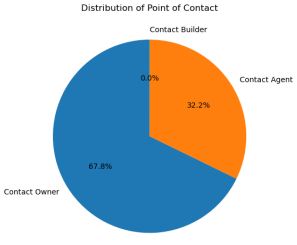


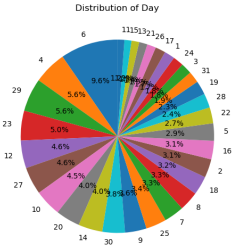


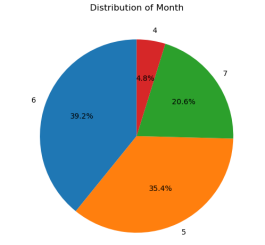


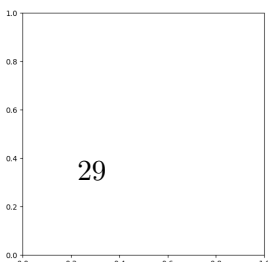




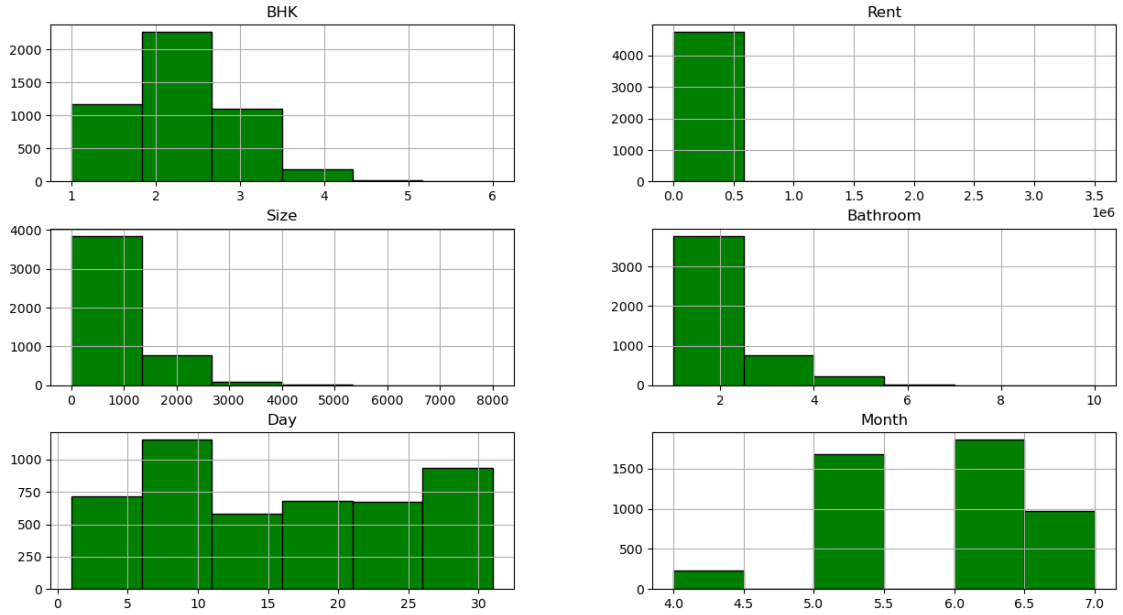








[79]:
```
df.hist(bins=6,figsize=(15,8), color='green', edgecolor='black')
plt.show()
```



[ ]: