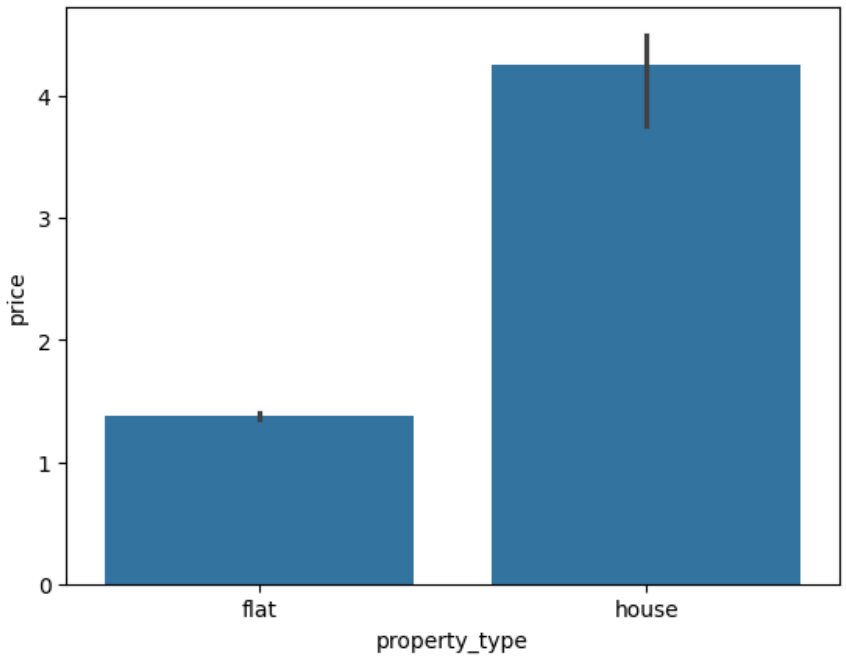
Capstone Project (EDA-Multivariate Analysis)

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

property_type vs price

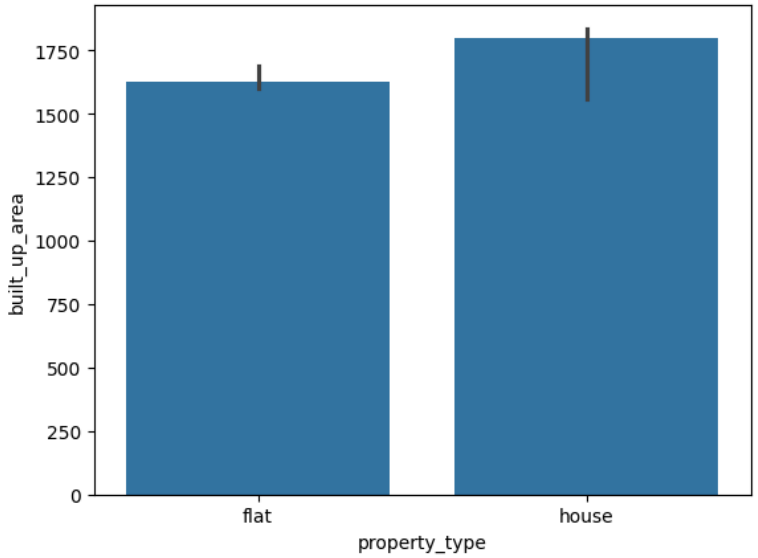
```
sns.barplot(x=df['property_type'], y=df['price'], estimator=np.median)
```



- Houses are more expensive

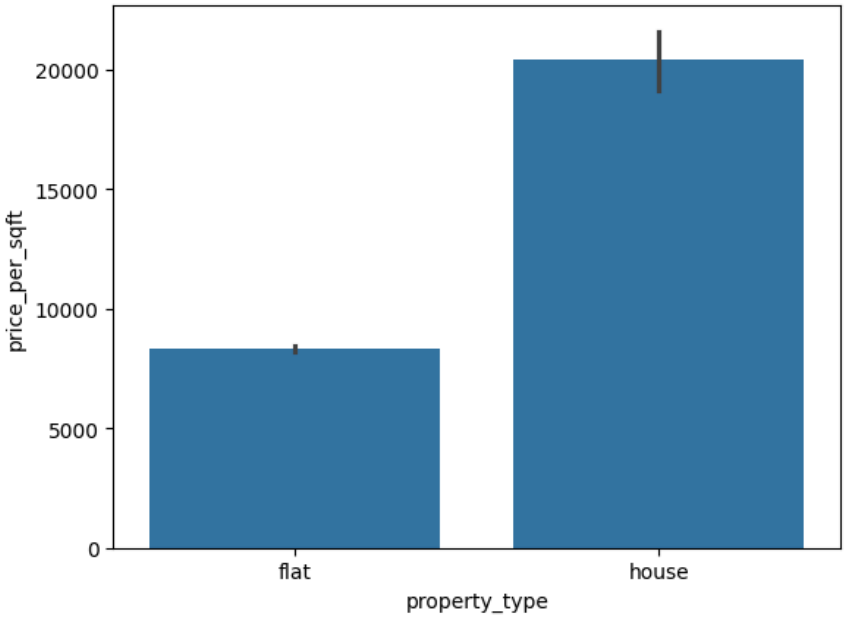
property_type vs area

```
sns.barplot(x=df['property_type'], y=df['built_up_area'], estimator=np.median)
```

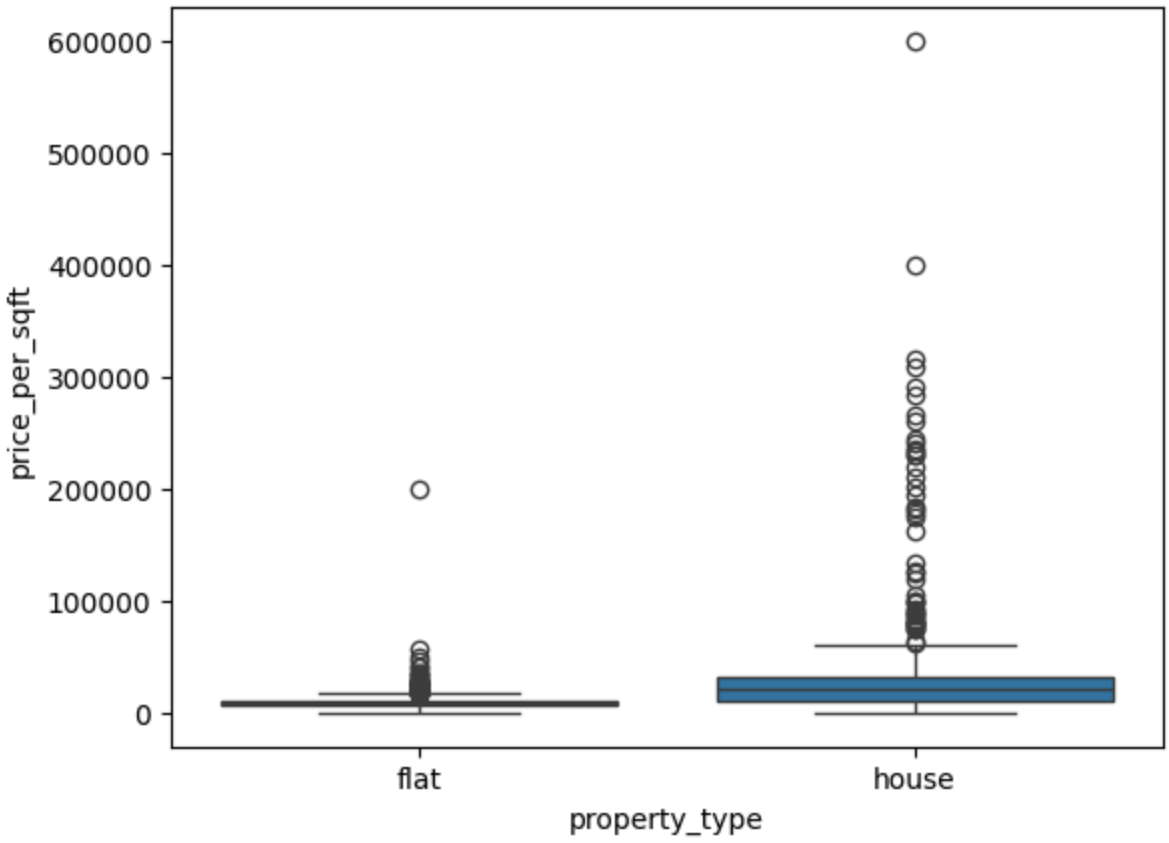


property_type vs price_per_sqft

```
sns.barplot(x=df['property_type'], y=df['price_per_sqft'], estimator=np.median)
```



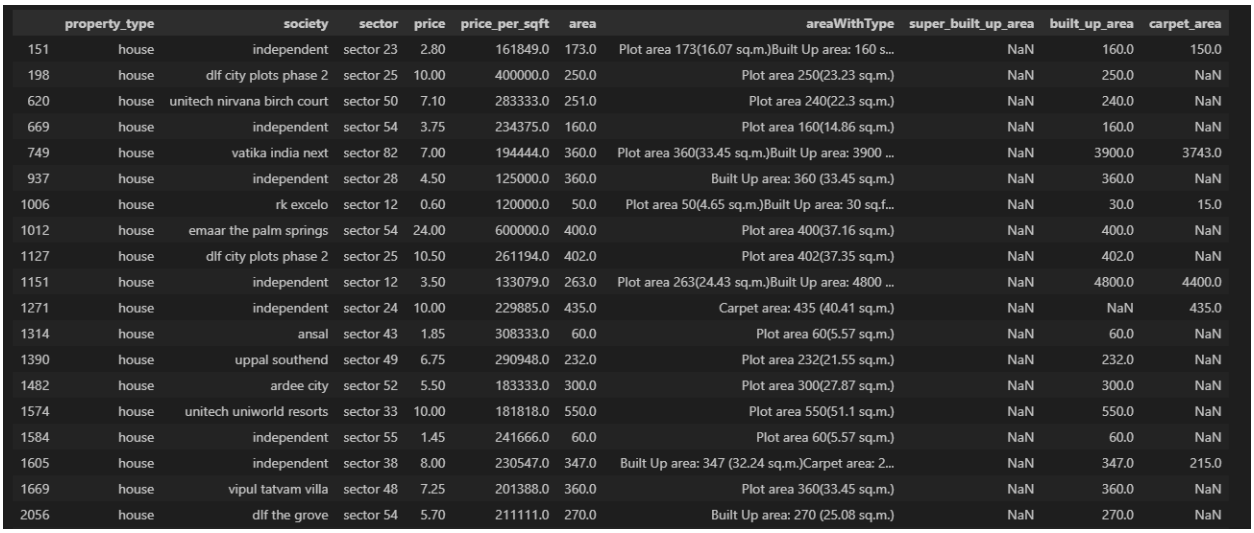

```
sns.boxplot(x=df['property_type'], y=df['price_per_sqft'])
```



- There are outliers

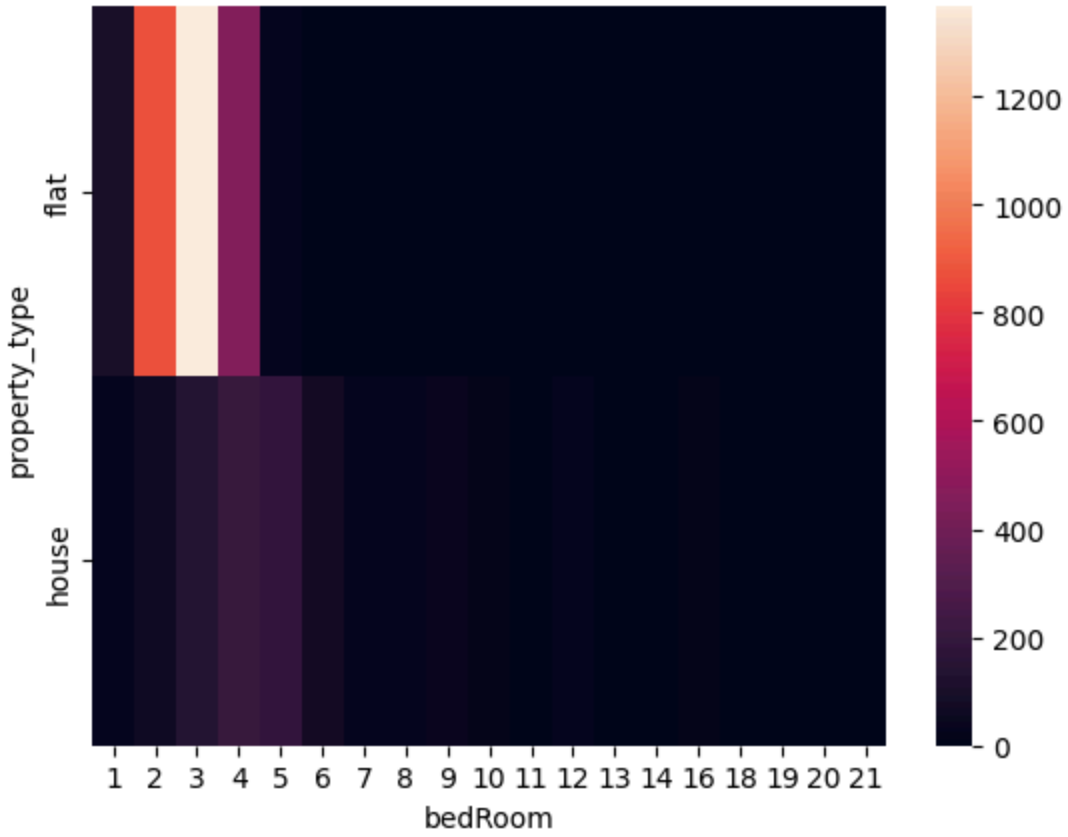
Check outliers:

```
# check outliers
df[df['price_per_sqft'] > 100000][['property_type','society','sector','price','price_per_sqft','area','areaWithType', 'super_built_up_area', 'built_up_area', 'carpet_area']]
```



	property_type	society	sector	price	price_per_sqft	area	areaWithType	super_built_up_area	built_up_area	carpet_area
151	house	independent	sector 23	2.80	161849.0	173.0	Plot area 173(16.07 sq.m.)Built Up area: 160 s...	NaN	160.0	150.0
198	house	dlf city plots phase 2	sector 25	10.00	400000.0	250.0	Plot area 250(23.23 sq.m.)	NaN	250.0	NaN
620	house	unitech nirvana birch court	sector 50	7.10	283333.0	251.0	Plot area 240(22.3 sq.m.)	NaN	240.0	NaN
669	house	independent	sector 54	3.75	234375.0	160.0	Plot area 160(14.86 sq.m.)	NaN	160.0	NaN
749	house	vatika india next	sector 82	7.00	194444.0	360.0	Plot area 360(33.45 sq.m.)Built Up area: 3900 ...	NaN	3900.0	3743.0
937	house	independent	sector 28	4.50	125000.0	360.0	Built Up area: 360 (33.45 sq.m.)	NaN	360.0	NaN
1006	house	rk excelo	sector 12	0.60	120000.0	50.0	Plot area 50(4.65 sq.m.)Built Up area: 30 sq.f...	NaN	30.0	15.0
1012	house	emaar the palm springs	sector 54	24.00	600000.0	400.0	Plot area 400(37.16 sq.m.)	NaN	400.0	NaN
1127	house	dlf city plots phase 2	sector 25	10.50	261194.0	402.0	Plot area 402(37.35 sq.m.)	NaN	402.0	NaN
1151	house	independent	sector 12	3.50	133079.0	263.0	Plot area 263(24.43 sq.m.)Built Up area: 4800 ...	NaN	4800.0	4400.0
1271	house	independent	sector 24	10.00	229885.0	435.0	Carpet area: 435 (40.41 sq.m.)	NaN	NaN	435.0
1314	house	ansal	sector 43	1.85	308333.0	60.0	Plot area 60(5.57 sq.m.)	NaN	60.0	NaN
1390	house	uppal southend	sector 49	6.75	290948.0	232.0	Plot area 232(21.55 sq.m.)	NaN	232.0	NaN
1482	house	ardee city	sector 52	5.50	183333.0	300.0	Plot area 300(27.87 sq.m.)	NaN	300.0	NaN
1574	house	unitech uniworld resorts	sector 33	10.00	181818.0	550.0	Plot area 550(51.1 sq.m.)	NaN	550.0	NaN
1584	house	independent	sector 55	1.45	241666.0	60.0	Plot area 60(5.57 sq.m.)	NaN	60.0	NaN
1605	house	independent	sector 38	8.00	230547.0	347.0	Built Up area: 347 (32.24 sq.m.)Carpet area: 2...	NaN	347.0	215.0
1669	house	vipul tatvam villa	sector 48	7.25	201388.0	360.0	Plot area 360(33.45 sq.m.)	NaN	360.0	NaN
2056	house	dlf the grove	sector 54	5.70	211111.0	270.0	Built Up area: 270 (25.08 sq.m.)	NaN	270.0	NaN

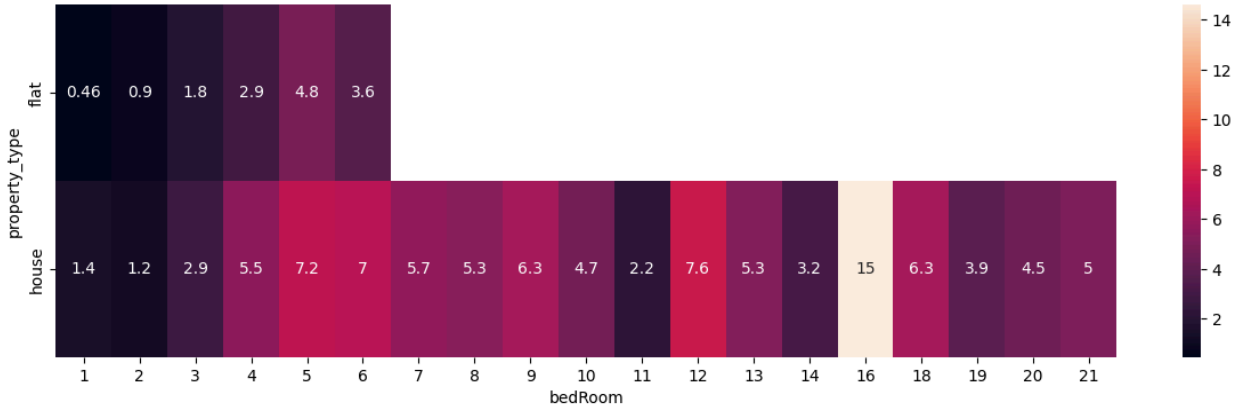
```
sns.heatmap(pd.crosstab(df['property_type'],df['bedRoom']))
```




```
pd.crosstab(df['property_type'],df['bedRoom']).T
```

property_type	flat	house
bedRoom		
1	94	30
2	873	68
3	1367	129
4	453	207
5	28	182
6	2	72
7	0	28
8	0	30
9	0	41
10	0	20
11	0	1
12	0	28
13	0	4
14	0	1
16	0	12
18	0	2
19	0	2
20	0	1
21	0	1

```
plt.figure(figsize=(15,4))
sns.heatmap(pd.pivot_table(df,index='property_type',columns='bedRoom',values='price',aggfunc='mean'),annot=True)
```



Sector Analysis

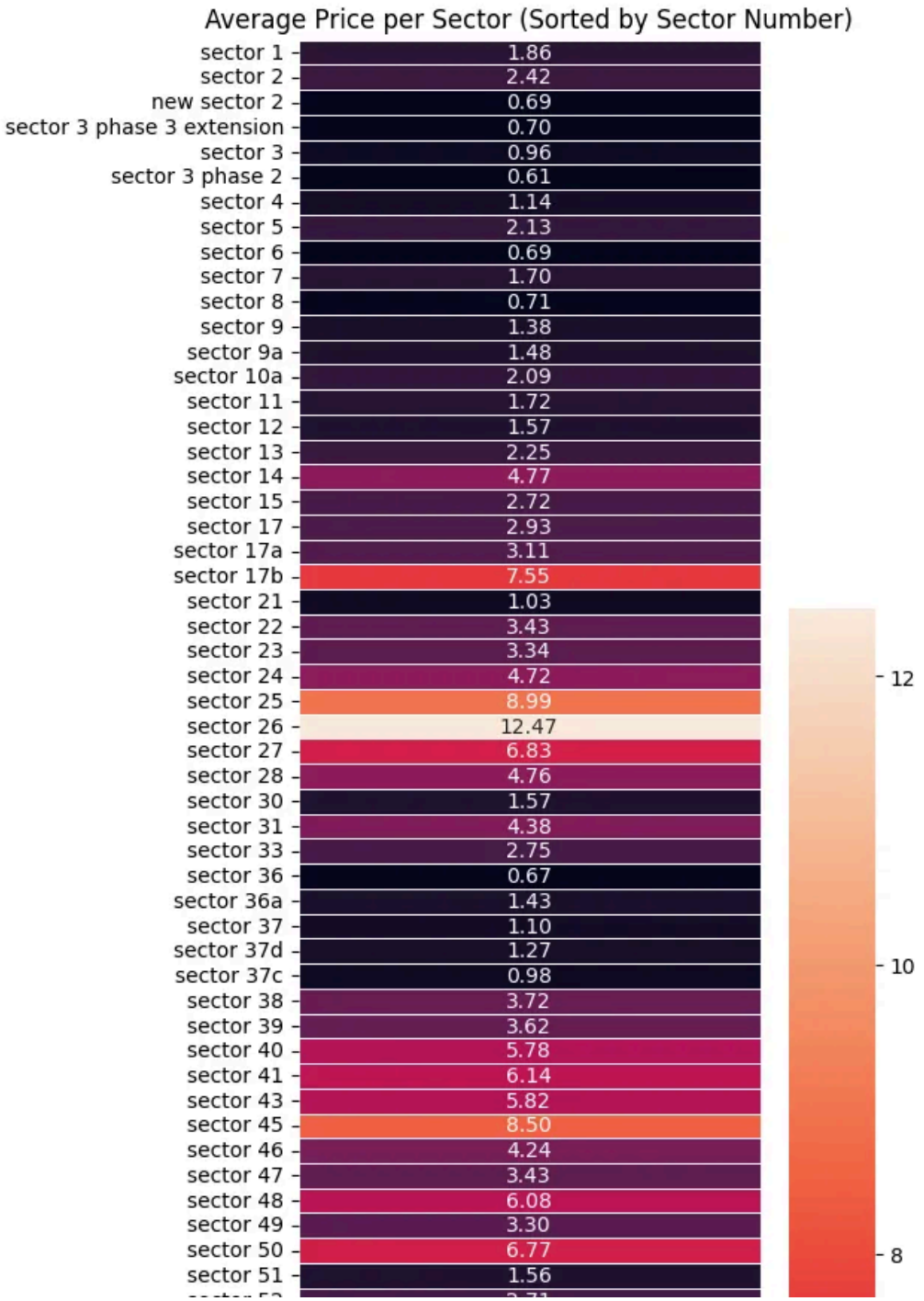
```
# sector analysis
import re
# Group by 'sector' and calculate the average price
avg_price_per_sector = df.groupby('sector')['price'].mean().reset_index()

# Function to extract sector numbers
def extract_sector_number(sector_name):
    match = re.search(r'\d+', sector_name)
    if match:
        return int(match.group())
    else:
        return float('inf')  # Return a large number for non-numbered sectors

avg_price_per_sector['sector_number'] = avg_price_per_sector['sector'].appl
y(extract_sector_number)

# Sort by sector number
avg_price_per_sector_sorted_by_sector = avg_price_per_sector.sort_values(by
='sector_number')
```

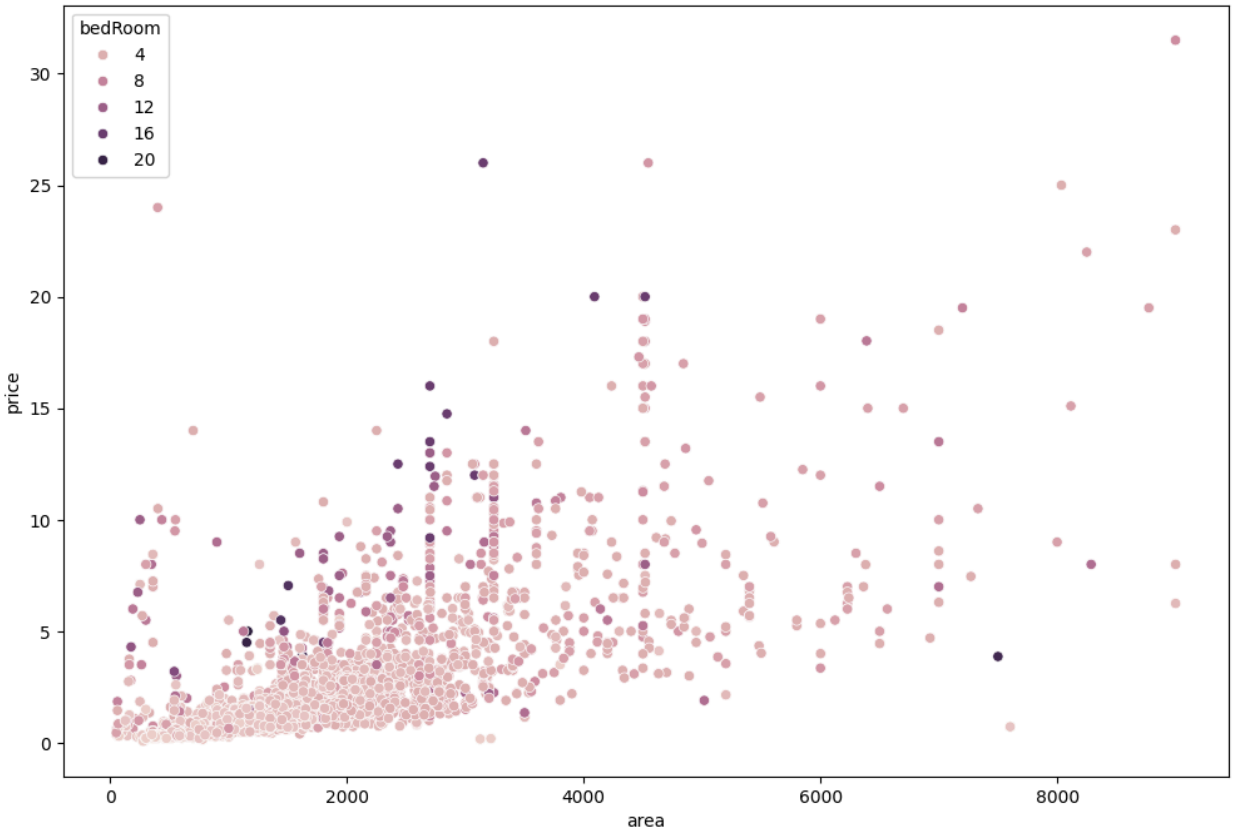
```
# Plot the heatmap
plt.figure(figsize=(5, 25))
sns.heatmap(avg_price_per_sector_sorted_by_sector.set_index('sector')[['price']], annot=True, fmt=".2f", linewidths=.5)
plt.title('Average Price per Sector (Sorted by Sector Number)')
plt.xlabel('Average Price')
plt.ylabel('Sector')
plt.show()
```



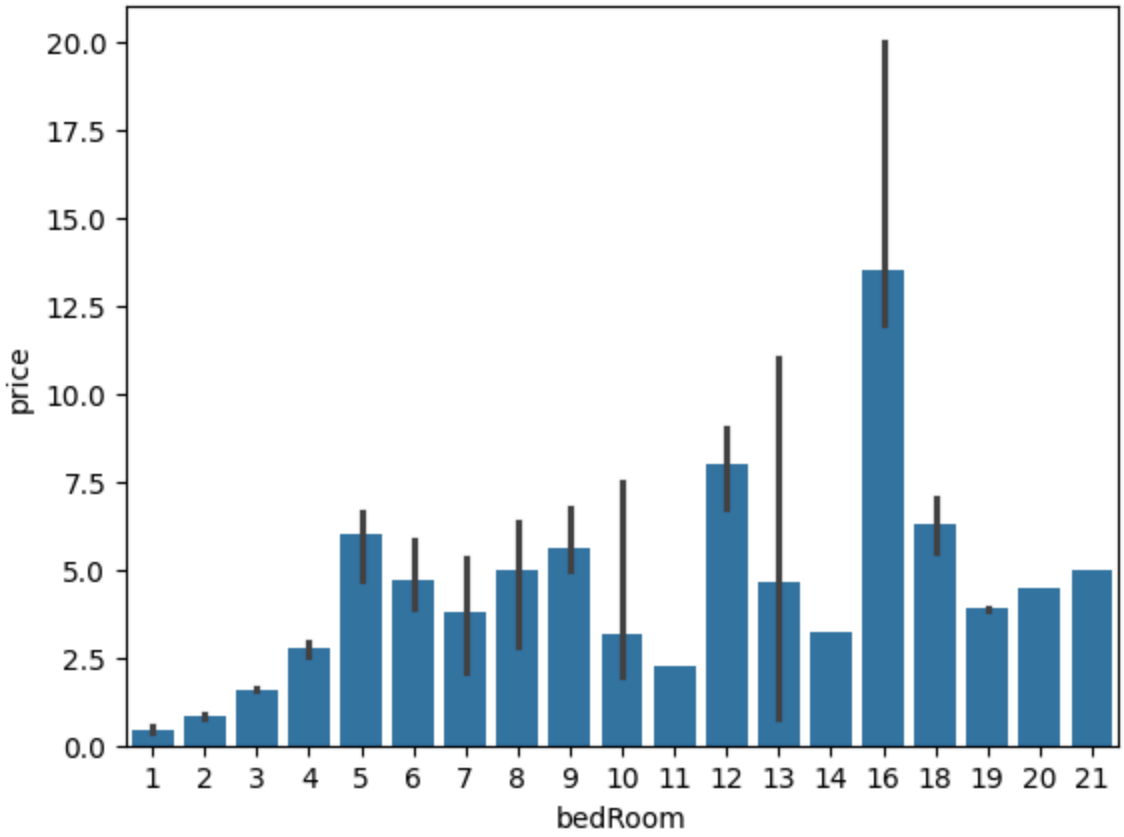
Price

```
plt.figure(figsize=(12,8))
sns.scatterplot(x=df[df['area']<10000]['area'],y=df['price'],hue=df['bedRoo
```

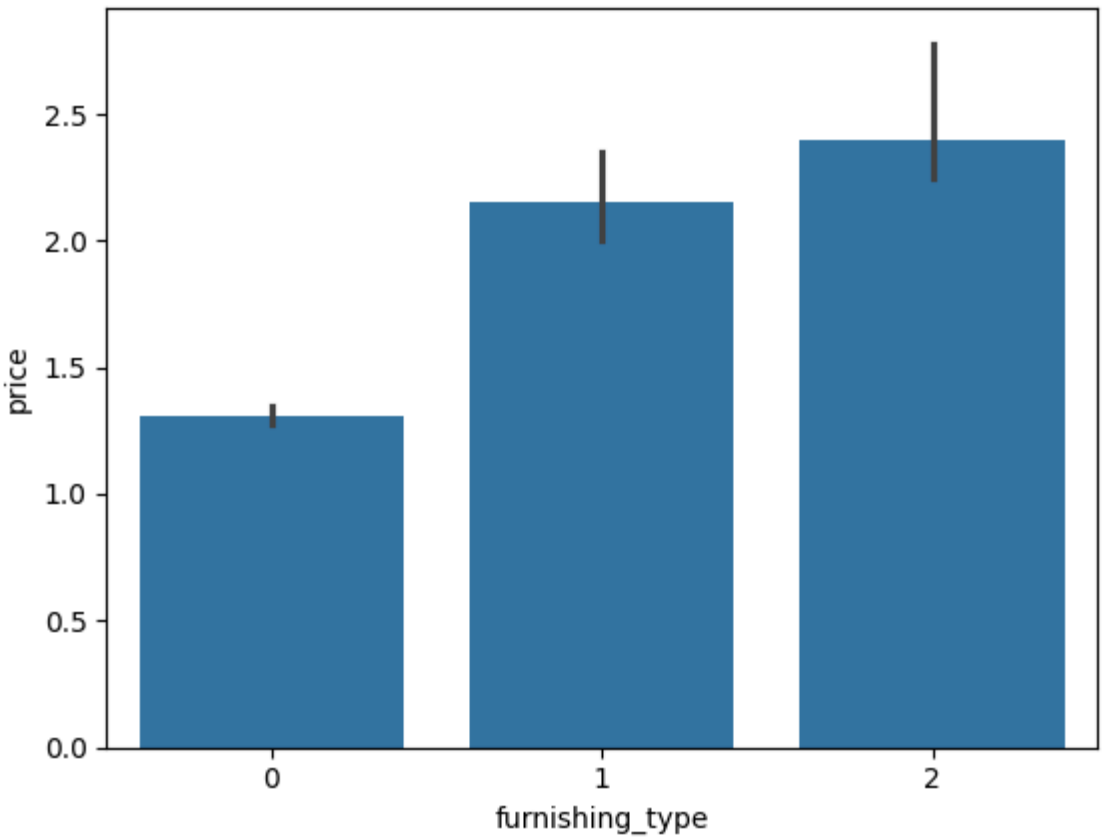
```
m'])
```




```
sns.barplot(x=df['bedRoom'],y=df['price'],estimator=np.median)
```



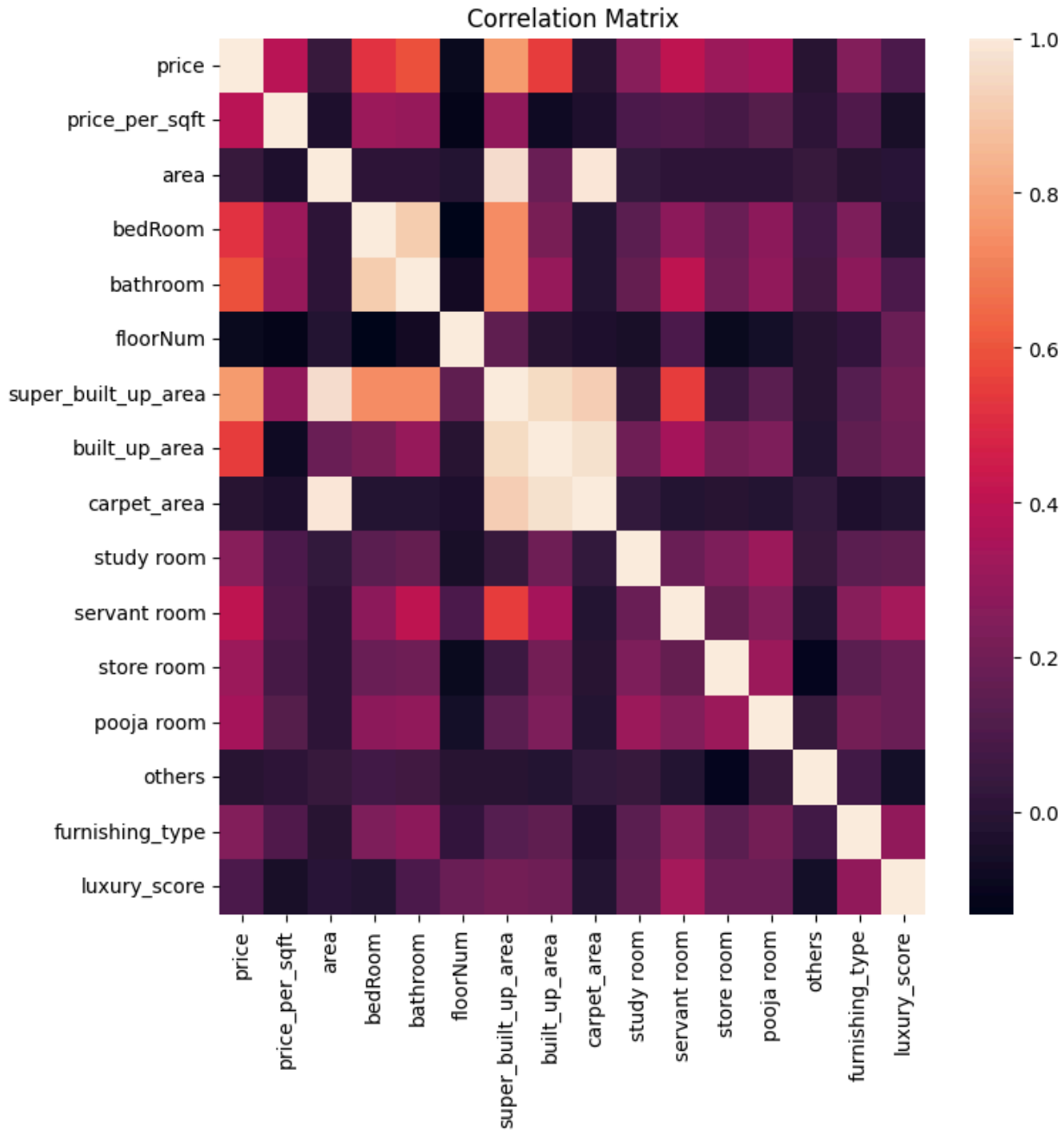

```
sns.barplot(x=df['furnishing_type'],y=df['price'],estimator=np.median)
```



Correlation

```
numeric_df = df.select_dtypes(include=['number'])

# Plot the heatmap
plt.figure(figsize=(8,8))
sns.heatmap(numeric_df.corr(), )
plt.title('Correlation Matrix')
plt.show()
```



`df.select_dtypes(include=['number'])` :

We selected only numeric columns.