dataset

November 5, 2023

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

[2]:
```
data=pd.read_csv(r"C:\Users\shara\Downloads\archive\House Price India.csv")
type(data)
data.dtypes
```

[2]:
```
id                                        int64
Date                                      int64
number of bedrooms                        int64
number of bathrooms                     float64
living area                               int64
lot area                                  int64
number of floors                        float64
waterfront present                        int64
number of views                           int64
condition of the house                    int64
grade of the house                        int64
Area of the house(excluding basement)     int64
Area of the basement                      int64
Built Year                                int64
Renovation Year                           int64
Postal Code                               int64
Lattitude                               float64
Longitude                               float64
living_area_renov                         int64
lot_area_renov                            int64
Number of schools nearby                  int64
Distance from the airport                 int64
Price                                     int64
dtype: object
```

[4]:
```
cat=[]
num=[]
for column in data.columns:
```

```
    if data[column].nunique() > 10:
        num.append(column)
    else:
        cat.append(column)
```

[5]:
```
cat
```

[5]:
```
['number of floors',
 'waterfront present',
 'number of views',
 'condition of the house',
 'grade of the house',
 'Number of schools nearby']
```

[6]:
```
num
```

[6]:
```
['id',
 'Date',
 'number of bedrooms',
 'number of bathrooms',
 'living area',
 'lot area',
 'Area of the house(excluding basement)',
 'Area of the basement',
 'Built Year',
 'Renovation Year',
 'Postal Code',
 'Lattitude',
 'Longitude',
 'living_area_renov',
 'lot_area_renov',
 'Distance from the airport',
 'Price']
```

[7]:
```
data.describe()#descriptive statistics
```

[7]:
```
                  id          Date  number of bedrooms  number of bathrooms  \
count   1.462000e+04  14620.000000        14620.000000         14620.000000
mean    6.762821e+09  42604.538646            3.379343             2.129583
std     6.237575e+03     67.347991            0.938719             0.769934
min     6.762810e+09  42491.000000            1.000000             0.500000
25%     6.762815e+09  42546.000000            3.000000             1.750000
50%     6.762821e+09  42600.000000            3.000000             2.250000
75%     6.762826e+09  42662.000000            4.000000             2.500000
max     6.762832e+09  42734.000000           33.000000             8.000000

         living area      lot area  number of floors  waterfront present  \
```

count	14620.000000	1.462000e+04	14620.000000	14620.000000
mean	2098.262996	1.509328e+04	1.502360	0.007661
std	928.275721	3.791962e+04	0.540239	0.087193
min	370.000000	5.200000e+02	1.000000	0.000000
25%	1440.000000	5.010750e+03	1.000000	0.000000
50%	1930.000000	7.620000e+03	1.500000	0.000000
75%	2570.000000	1.080000e+04	2.000000	0.000000
max	13540.000000	1.074218e+06	3.500000	1.000000

	number of views	condition of the house	…	Built Year	\
count	14620.000000	14620.000000	…	14620.000000	
mean	0.233105	3.430506	…	1970.926402	
std	0.766259	0.664151	…	29.493625	
min	0.000000	1.000000	…	1900.000000	
25%	0.000000	3.000000	…	1951.000000	
50%	0.000000	3.000000	…	1975.000000	
75%	0.000000	4.000000	…	1997.000000	
max	4.000000	5.000000	…	2015.000000	

	Renovation Year	Postal Code	Lattitude	Longitude	\
count	14620.000000	14620.000000	14620.000000	14620.000000	
mean	90.924008	122033.062244	52.792848	-114.404007	
std	416.216661	19.082418	0.137522	0.141326	
min	0.000000	122003.000000	52.385900	-114.709000	
25%	0.000000	122017.000000	52.707600	-114.519000	
50%	0.000000	122032.000000	52.806400	-114.421000	
75%	0.000000	122048.000000	52.908900	-114.315000	
max	2015.000000	122072.000000	53.007600	-113.505000	

	living_area_renov	lot_area_renov	Number of schools nearby	\
count	14620.000000	14620.000000	14620.000000	
mean	1996.702257	12753.500068	2.012244	
std	691.093366	26058.414467	0.817284	
min	460.000000	651.000000	1.000000	
25%	1490.000000	5097.750000	1.000000	
50%	1850.000000	7620.000000	2.000000	
75%	2380.000000	10125.000000	3.000000	
max	6110.000000	560617.000000	3.000000	

	Distance from the airport	Price
count	14620.000000	1.462000e+04
mean	64.950958	5.389322e+05
std	8.936008	3.675324e+05
min	50.000000	7.800000e+04
25%	57.000000	3.200000e+05
50%	65.000000	4.500000e+05
75%	73.000000	6.450000e+05

```
max                    80.000000  7.700000e+06

[8 rows x 23 columns]
```

[8]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14620 entries, 0 to 14619
Data columns (total 23 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   id                                     14620 non-null  int64
 1   Date                                   14620 non-null  int64
 2   number of bedrooms                     14620 non-null  int64
 3   number of bathrooms                    14620 non-null  float64
 4   living area                            14620 non-null  int64
 5   lot area                               14620 non-null  int64
 6   number of floors                       14620 non-null  float64
 7   waterfront present                     14620 non-null  int64
 8   number of views                        14620 non-null  int64
 9   condition of the house                 14620 non-null  int64
 10  grade of the house                     14620 non-null  int64
 11  Area of the house(excluding basement)  14620 non-null  int64
 12  Area of the basement                   14620 non-null  int64
 13  Built Year                             14620 non-null  int64
 14  Renovation Year                        14620 non-null  int64
 15  Postal Code                            14620 non-null  int64
 16  Lattitude                              14620 non-null  float64
 17  Longitude                              14620 non-null  float64
 18  living_area_renov                      14620 non-null  int64
 19  lot_area_renov                         14620 non-null  int64
 20  Number of schools nearby               14620 non-null  int64
 21  Distance from the airport              14620 non-null  int64
 22  Price                                  14620 non-null  int64
dtypes: float64(4), int64(19)
memory usage: 2.6 MB
```

[9]:
```
data.isnull()#handling missing values
```

[9]:

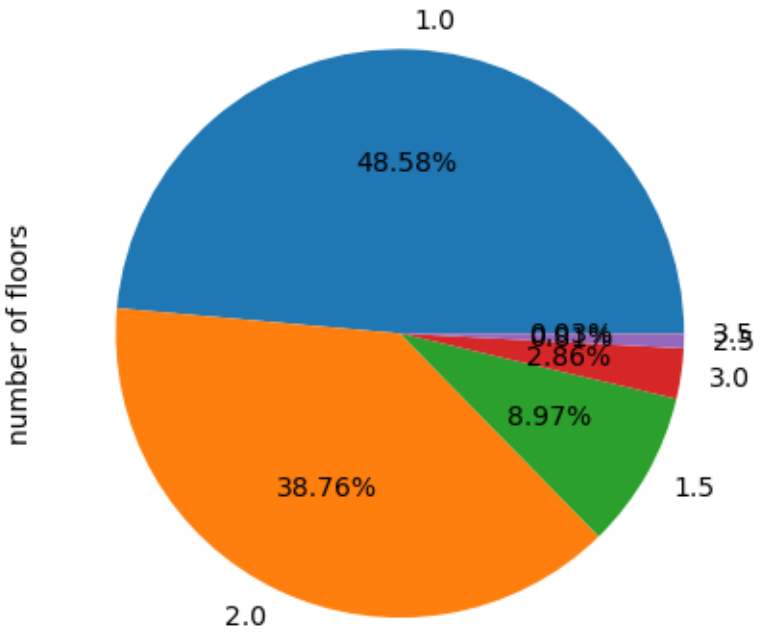
	id	Date	number of bedrooms	number of bathrooms	living area	\
0	False	False	False	False	False	
1	False	False	False	False	False	
2	False	False	False	False	False	
3	False	False	False	False	False	
4	False	False	False	False	False	
…	…	…	…	…	…	
14615	False	False	False	False	False	

```
14616  False  False               False                False        False
14617  False  False               False                False        False
14618  False  False               False                False        False
14619  False  False               False                False        False

       lot area  number of floors  waterfront present  number of views  \
0         False             False               False            False
1         False             False               False            False
2         False             False               False            False
3         False             False               False            False
4         False             False               False            False
…         …                 …                   …                …
14615     False             False               False            False
14616     False             False               False            False
14617     False             False               False            False
14618     False             False               False            False
14619     False             False               False            False

       condition of the house  …  Built Year  Renovation Year  Postal Code  \
0                       False  …       False            False        False
1                       False  …       False            False        False
2                       False  …       False            False        False
3                       False  …       False            False        False
4                       False  …       False            False        False
…                       …  …           …                …            …
14615                   False  …       False            False        False
14616                   False  …       False            False        False
14617                   False  …       False            False        False
14618                   False  …       False            False        False
14619                   False  …       False            False        False

       Lattitude  Longitude  living_area_renov  lot_area_renov  \
0          False      False              False           False
1          False      False              False           False
2          False      False              False           False
3          False      False              False           False
4          False      False              False           False
…          …          …                  …               …
14615      False      False              False           False
14616      False      False              False           False
14617      False      False              False           False
14618      False      False              False           False
14619      False      False              False           False

       Number of schools nearby  Distance from the airport  Price
0                         False                      False  False
1                         False                      False  False
```

```
2                          False                        False  False
3                          False                        False  False
4                          False                        False  False
…                          …                        …      …
14615                      False                        False  False
14616                      False                        False  False
14617                      False                        False  False
14618                      False                        False  False
14619                      False                        False  False

[14620 rows x 23 columns]
```

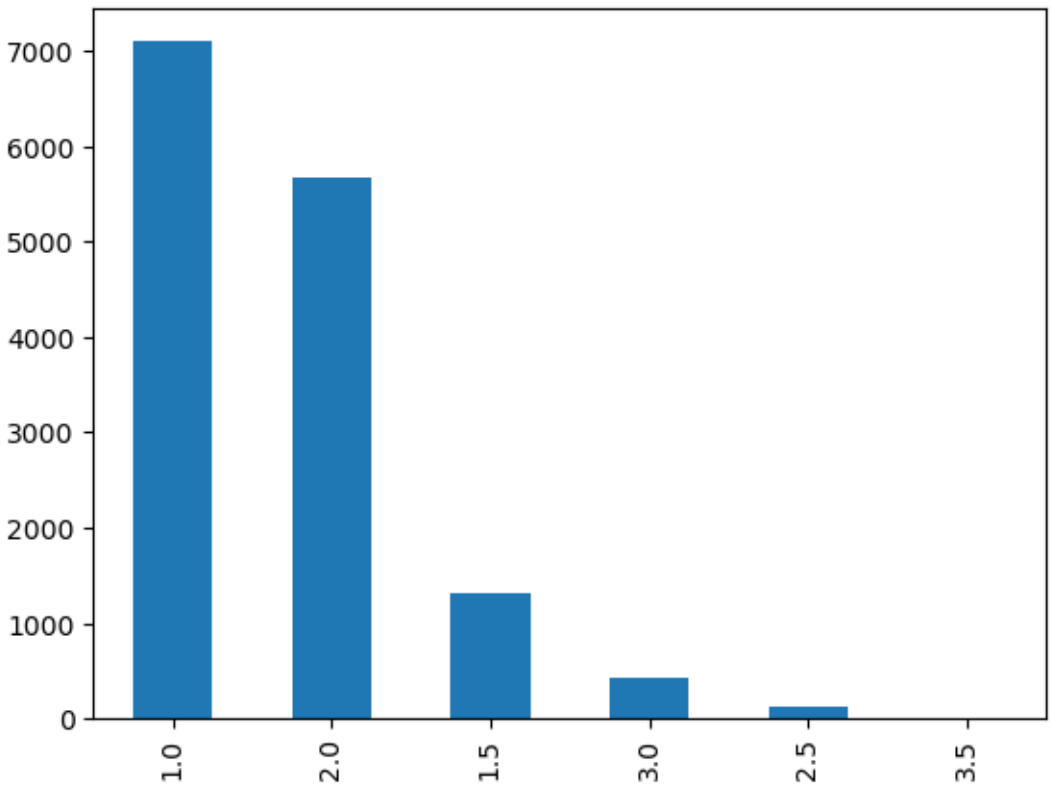
[10]:
```
data['number of floors'].value_counts().plot(kind="pie",autopct="%1.
↪2f%%")#univariate analysis
```

[10]: <Axes: ylabel='number of floors'>



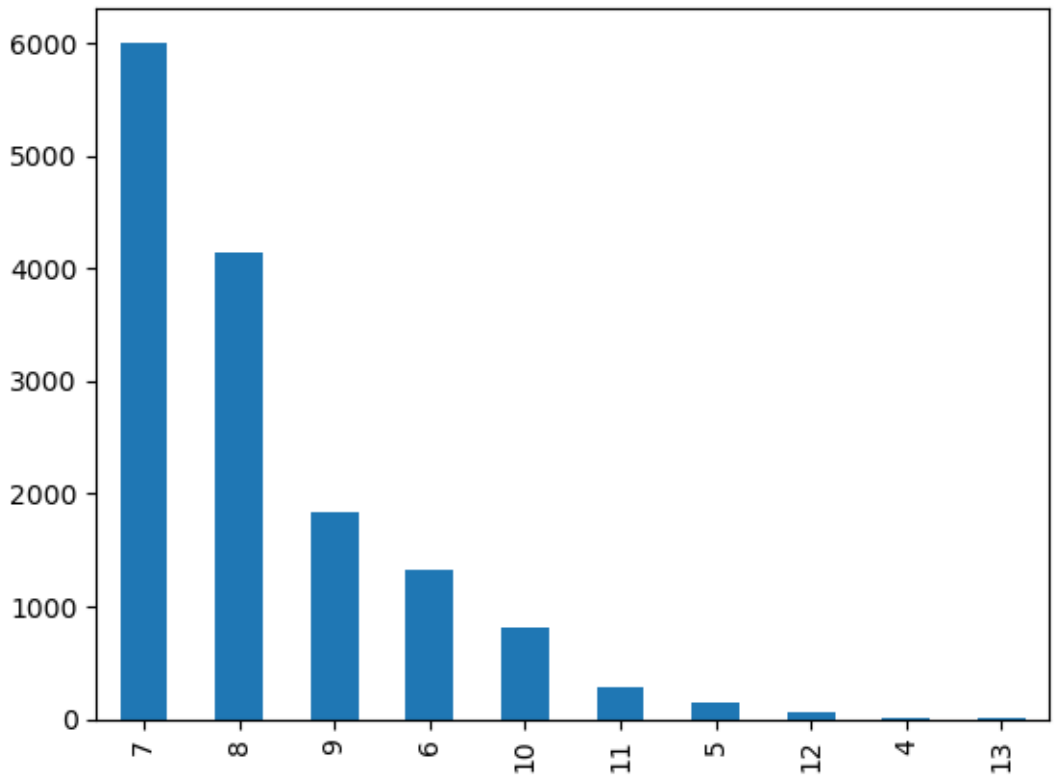
[11]:
```
data['number of floors'].value_counts().plot(kind="bar")
```

[11]: <Axes: >



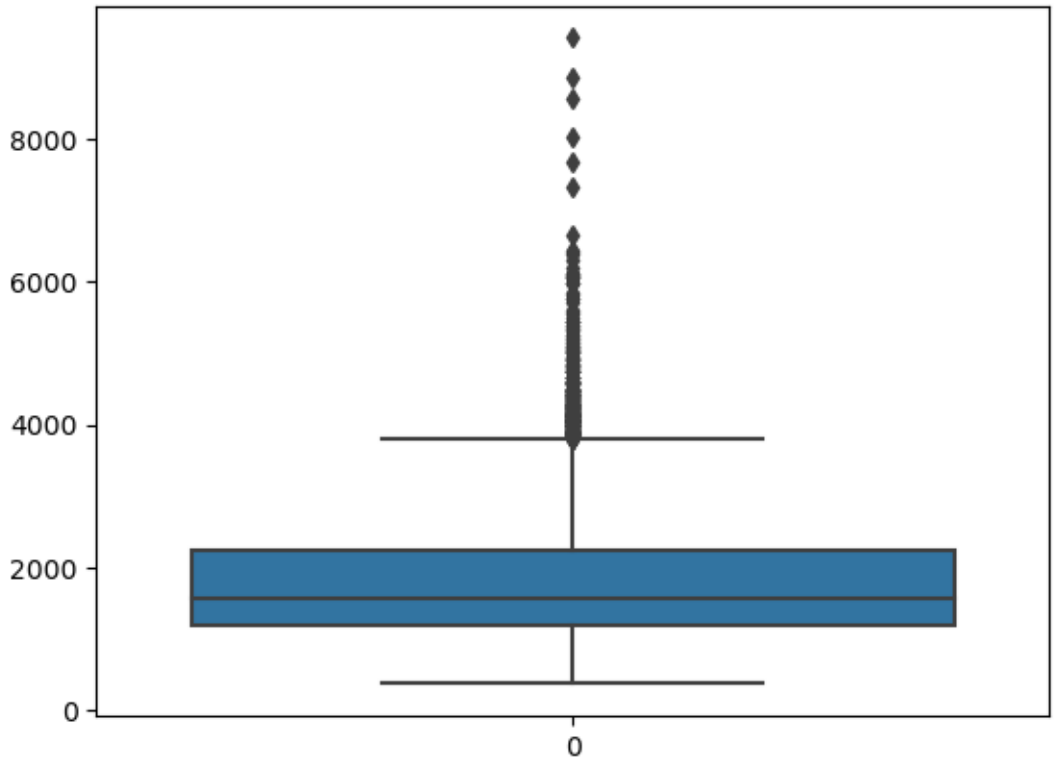
[12]:
```
data['grade of the house'].value_counts().plot(kind="bar")
```

[12]: <Axes: >



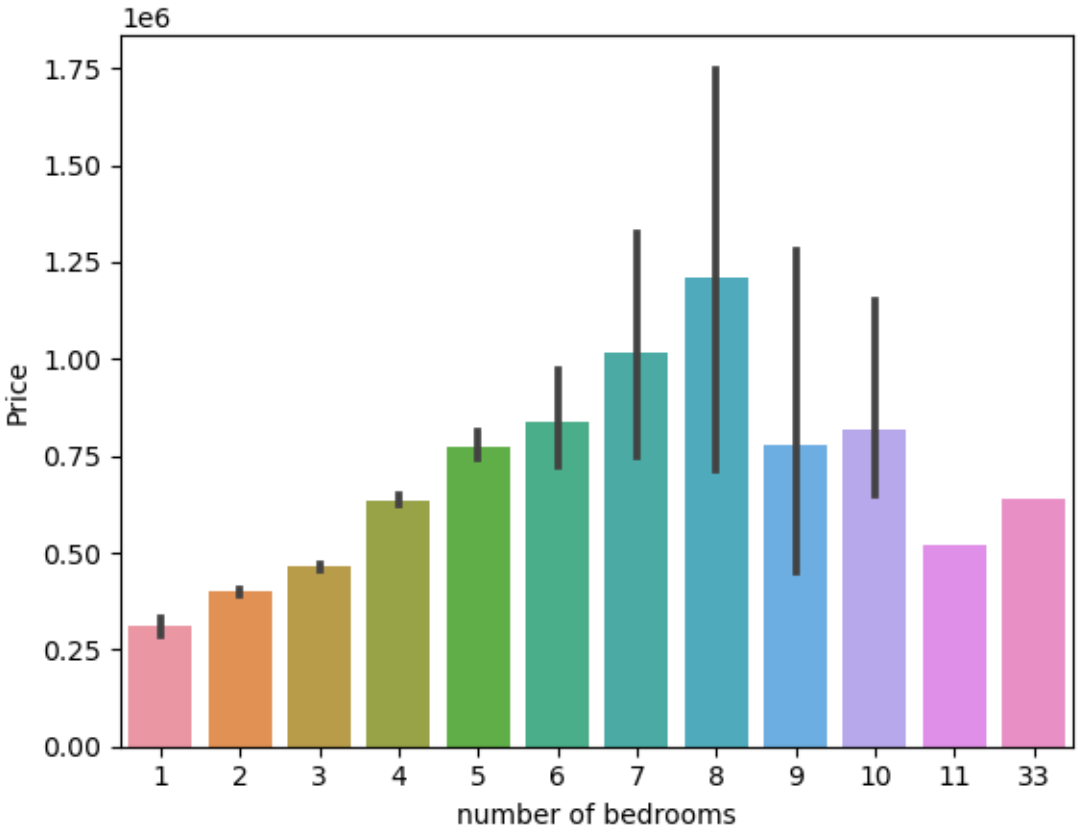
[13]:
```
sns.boxplot(data['Area of the house(excluding basement)'])
```

[13]: <Axes: >

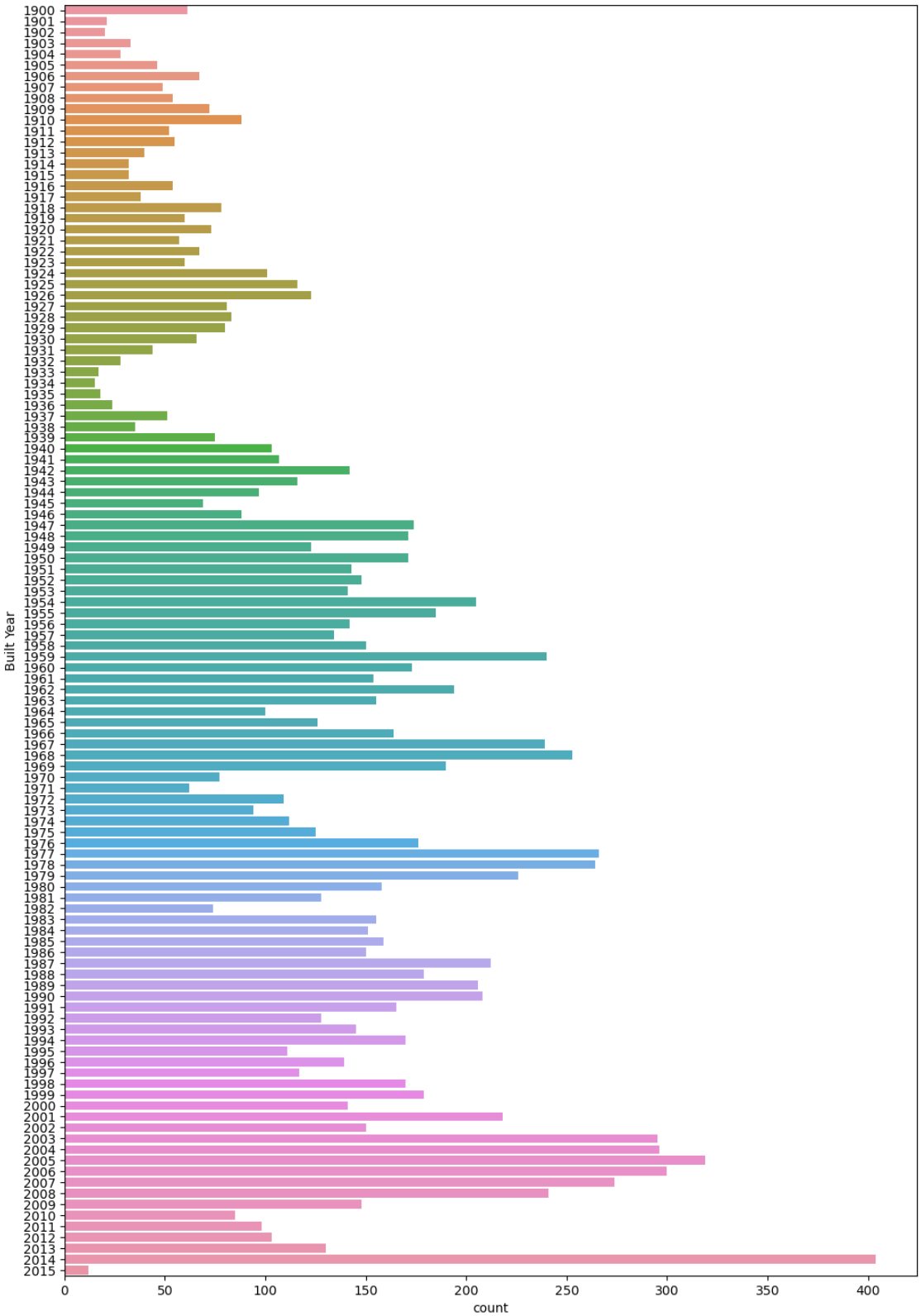


[14]:
```
sns.barplot(data=data,x='number of bedrooms',y='Price')#bivariate analysis
```

[14]:
```
<Axes: xlabel='number of bedrooms', ylabel='Price'>
```

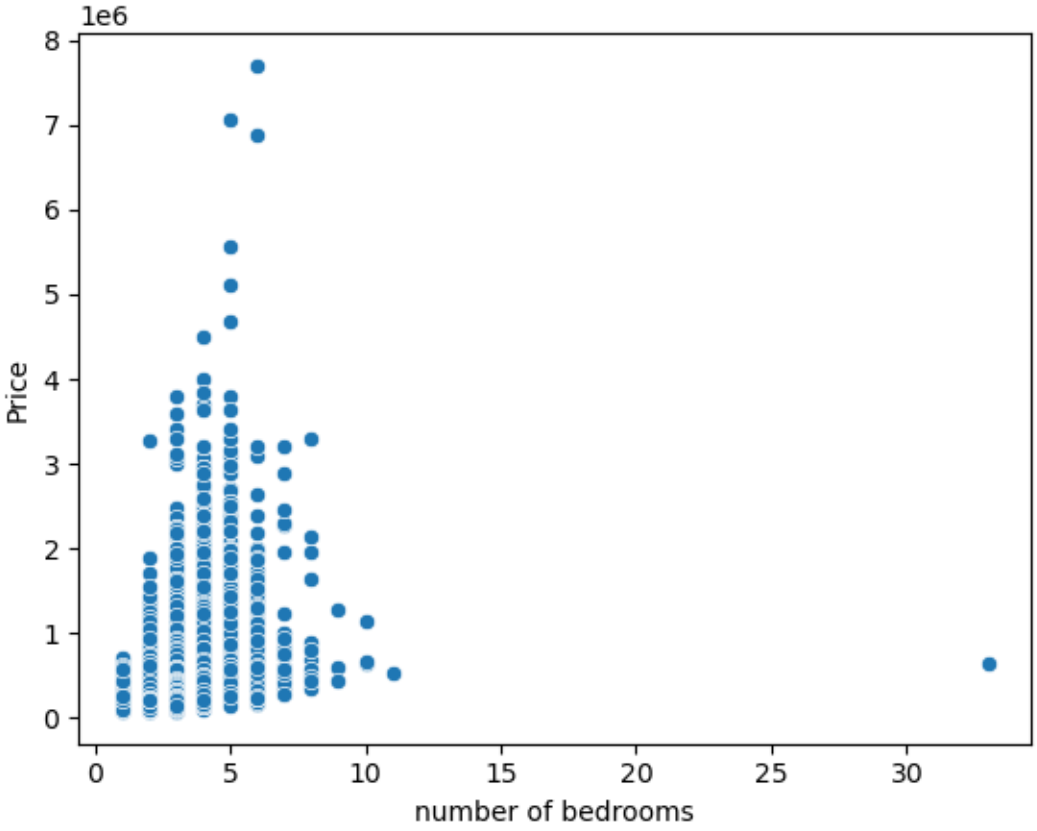


[15]:
```
plt.figure(figsize=(12,18))
sns.countplot(data=data,y='Built Year')
plt.show()
```

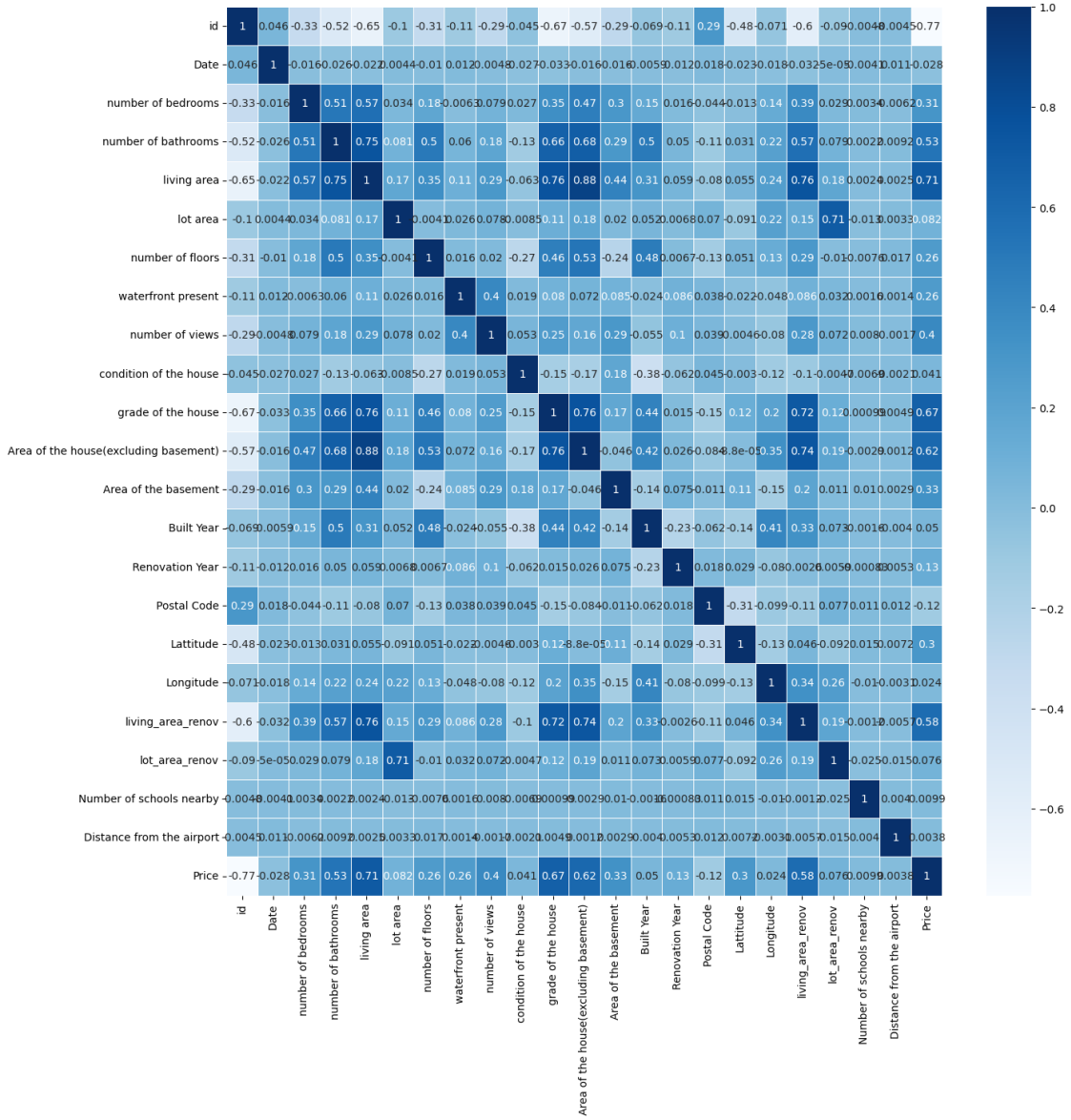

Built Year
count
0
50
100
150
200
250
300
350
400

[16]:
```
sns.scatterplot(data=data,x='number of bedrooms',y='Price')
```

[16]:
```
<Axes: xlabel='number of bedrooms', ylabel='Price'>
```

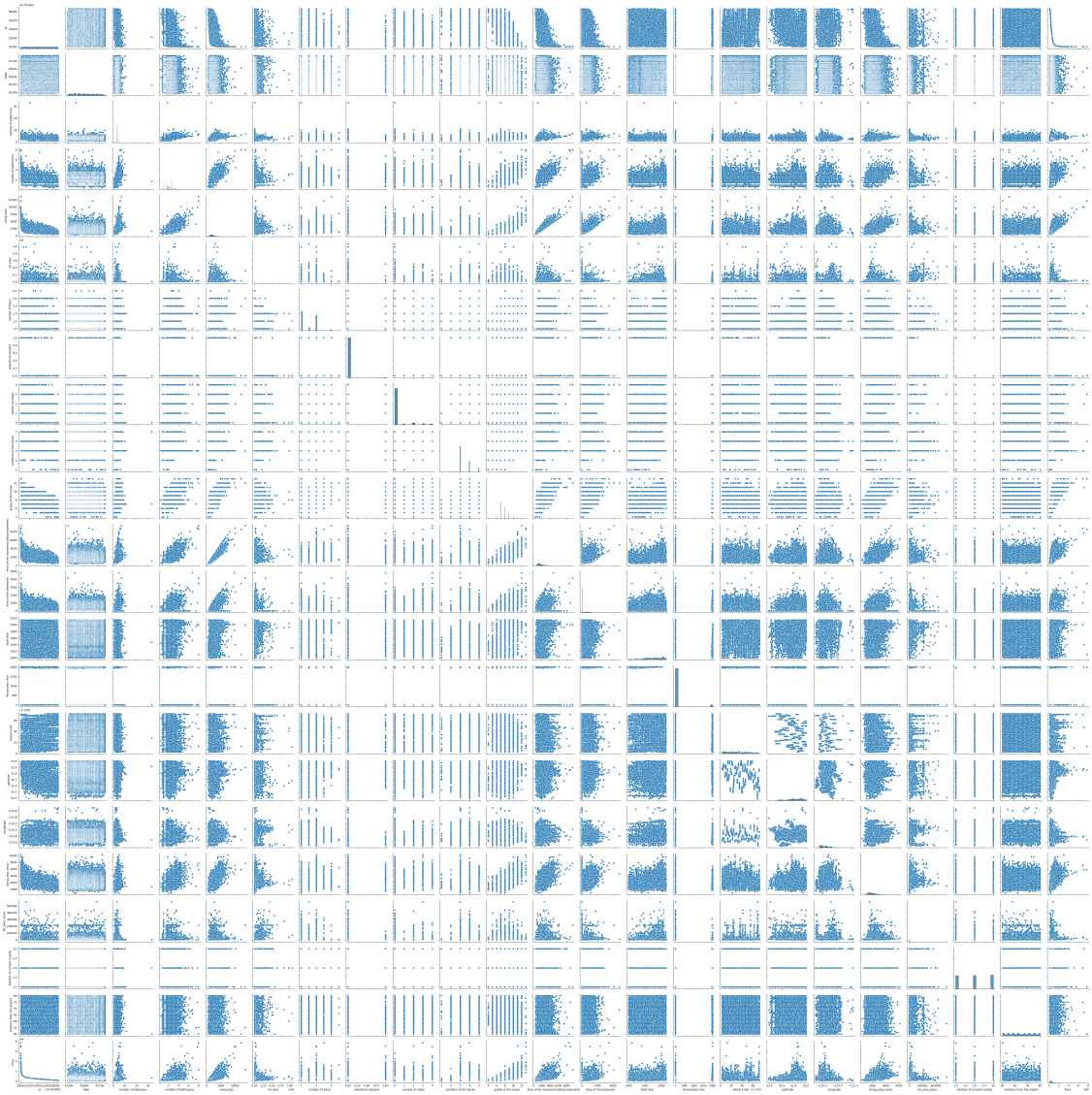


[17]:
```
plt.figure(figsize=(15,15))#multivariate analysis
sns.heatmap(data.corr(),linewidths=0.5,annot=True,cmap='Blues')
plt.show()
```



[ ]:

[ ]: <seaborn.axisgrid.PairGrid at 0x211dfb1d990>



[ ]: