Assignment-1 Data Preparation

```
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
```

1. Load dataset

```
dataset = pd.read_csv("House_Price_dataset.csv")
dataset
```

```
        price  area  bedrooms  bathrooms  stories mainroad guestroom
basement  \
0    13300000  7420       4.0        2.0      3.0      yes        no
no
1    12250000  8960       4.0        4.0      4.0      yes        no
no
2    12250000  9960       3.0        2.0      2.0      yes        no
yes
3    12215000  7500       4.0        2.0      2.0      yes        no
yes
4    11410000  7420       4.0        1.0      2.0      yes       yes
yes
..        ...   ...       ...        ...      ...      ...       ...
...
540   1820000  3000       2.0        1.0      1.0      yes        no
yes
541   1767150  2400       3.0        1.0      1.0       no        no
no
542   1750000  3620       2.0        1.0      1.0      yes        no
no
543   1750000  2910       3.0        1.0      1.0       no        no
no
544   1750000  3850       3.0        1.0      2.0      yes        no
no

    hotwaterheating airconditioning  parking prefarea furnishingstatus

0                no              yes      2.0      yes        furnished

1                no              yes      3.0       no        furnished

2                no               no      2.0      yes   semi-furnished
```

```
3                 no              yes      3.0      yes        furnished
4                 no              yes      2.0       no        furnished
..               ...              ...      ...      ...              ...
540               no               no      2.0       no      unfurnished
541               no               no      0.0       no   semi-furnished
542               no               no      0.0       no      unfurnished
543               no               no      0.0       no        furnished
544               no               no      0.0       no      unfurnished

[545 rows x 13 columns]
```

2. Find the shape of data

```
dataset.shape
```

```
(545, 13)
```

3. Find the summary of data

```
dataset.describe()
```

```
              price          area    bedrooms   bathrooms     stories
\
count  5.450000e+02    545.000000  540.000000  540.000000  543.000000
mean   4.766729e+06   5150.541284    2.961111    1.285185    1.804788
std    1.870440e+06   2170.141023    0.738779    0.502464    0.869011
min    1.750000e+06   1650.000000    1.000000    1.000000    1.000000
25%    3.430000e+06   3600.000000    2.000000    1.000000    1.000000
50%    4.340000e+06   4600.000000    3.000000    1.000000    2.000000
75%    5.740000e+06   6360.000000    3.000000    2.000000    2.000000
max    1.330000e+07  16200.000000    6.000000    4.000000    4.000000
```

```
          parking
count  541.000000
mean     0.691312
std      0.860070
min      0.000000
25%      0.000000
50%      0.000000
75%      1.000000
max      3.000000
```

4. Find the data type of each column

```
dataset.dtypes
```

```
price                int64
area                 int64
bedrooms           float64
bathrooms          float64
stories            float64
mainroad            object
guestroom           object
basement            object
hotwaterheating     object
airconditioning     object
parking            float64
prefarea            object
furnishingstatus    object
dtype: object
```

5. Find Missing Values

```
dataset
```

```
        price  area  bedrooms  bathrooms  stories mainroad guestroom
basement  \
0    13300000  7420       4.0        2.0      3.0      yes        no
no
1    12250000  8960       4.0        4.0      4.0      yes        no
no
2    12250000  9960       3.0        2.0      2.0      yes        no
yes
3    12215000  7500       4.0        2.0      2.0      yes        no
yes
4    11410000  7420       4.0        1.0      2.0      yes       yes
yes
..        ...   ...       ...        ...      ...      ...       ...
```

```
...
540   1820000  3000       2.0        1.0      1.0      yes        no
yes
541   1767150  2400       3.0        1.0      1.0       no        no
no
542   1750000  3620       2.0        1.0      1.0      yes        no
no
543   1750000  2910       3.0        1.0      1.0       no        no
no
544   1750000  3850       3.0        1.0      2.0      yes        no
no

    hotwaterheating airconditioning  parking prefarea furnishingstatus

0                no              yes      2.0      yes        furnished

1                no              yes      3.0       no        furnished

2                no               no      2.0      yes   semi-furnished

3                no              yes      3.0      yes        furnished

4                no              yes      2.0       no        furnished

..              ...              ...      ...      ...              ...

540              no               no      2.0       no      unfurnished

541              no               no      0.0       no   semi-furnished

542              no               no      0.0       no      unfurnished

543              no               no      0.0       no        furnished

544              no               no      0.0       no      unfurnished


[545 rows x 13 columns]

dataset.isna()

     price   area  bedrooms  bathrooms  stories  mainroad
guestroom  \
0    False  False     False      False    False     False      False

1    False  False     False      False    False     False      False

2    False  False     False      False    False     False      False

3    False  False     False      False    False     False      False
```

```
4    False  False     False      False    False     False      False

..     ...    ...       ...        ...      ...       ...        ...

540  False  False     False      False    False     False      False

541  False  False     False      False    False     False      False

542  False  False     False      False    False     False      False

543  False  False     False      False    False     False      False

544  False  False     False      False    False     False      False


     basement  hotwaterheating  airconditioning  parking  prefarea  \
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
..        ...             ...              ...      ...       ...
540     False           False            False    False     False
541     False           False            False    False     False
542     False           False            False    False     False
543     False           False            False    False     False
544     False           False            False    False     False

     furnishingstatus
0               False
1               False
2               False
3               False
4               False
..                ...
540             False
541             False
542             False
543             False
544             False

[545 rows x 13 columns]

dataset.isna().sum()

price               0
area                0
bedrooms            5
bathrooms           5
stories             2
mainroad            0
```

```
guestroom           0
basement            0
hotwaterheating     0
airconditioning     0
parking             4
prefarea            0
furnishingstatus    0
dtype: int64
```

6. Finding out Zero's

```
(dataset==0).sum()
```

```
price                 0
area                  0
bedrooms              0
bathrooms             0
stories               0
mainroad              0
guestroom             0
basement              0
hotwaterheating       0
airconditioning       0
parking             297
prefarea              0
furnishingstatus      0
dtype: int64
```

7. Find Mean

```
"""
numerical_dataset = ["Price", "Area", "Bedrooms", "Bathrooms",
"Stories", "Parking"]
for data in numerical_dataset:
    print(f"{data} = {np.mean(dataset[data.lower()])}")
"""

dataset.mean(numeric_only=True)
```

```
price        4.766729e+06
area         5.150541e+03
bedrooms     2.961111e+00
bathrooms    1.285185e+00
stories      1.804788e+00
parking      6.913124e-01
dtype: float64
```

8. Replace the missing values

```
dataset.isna().sum()
```

```
price               0
area                0
bedrooms            5
bathrooms           5
stories             2
mainroad            0
guestroom           0
basement            0
hotwaterheating     0
airconditioning     0
parking             4
prefarea            0
furnishingstatus    0
dtype: int64
```

```
for c_name, c_content in dataset.items():
    if pd.isnull(c_content).sum():
        c_content.fillna(c_content.mean(), inplace=True)
```

```
dataset.isna().sum()
```

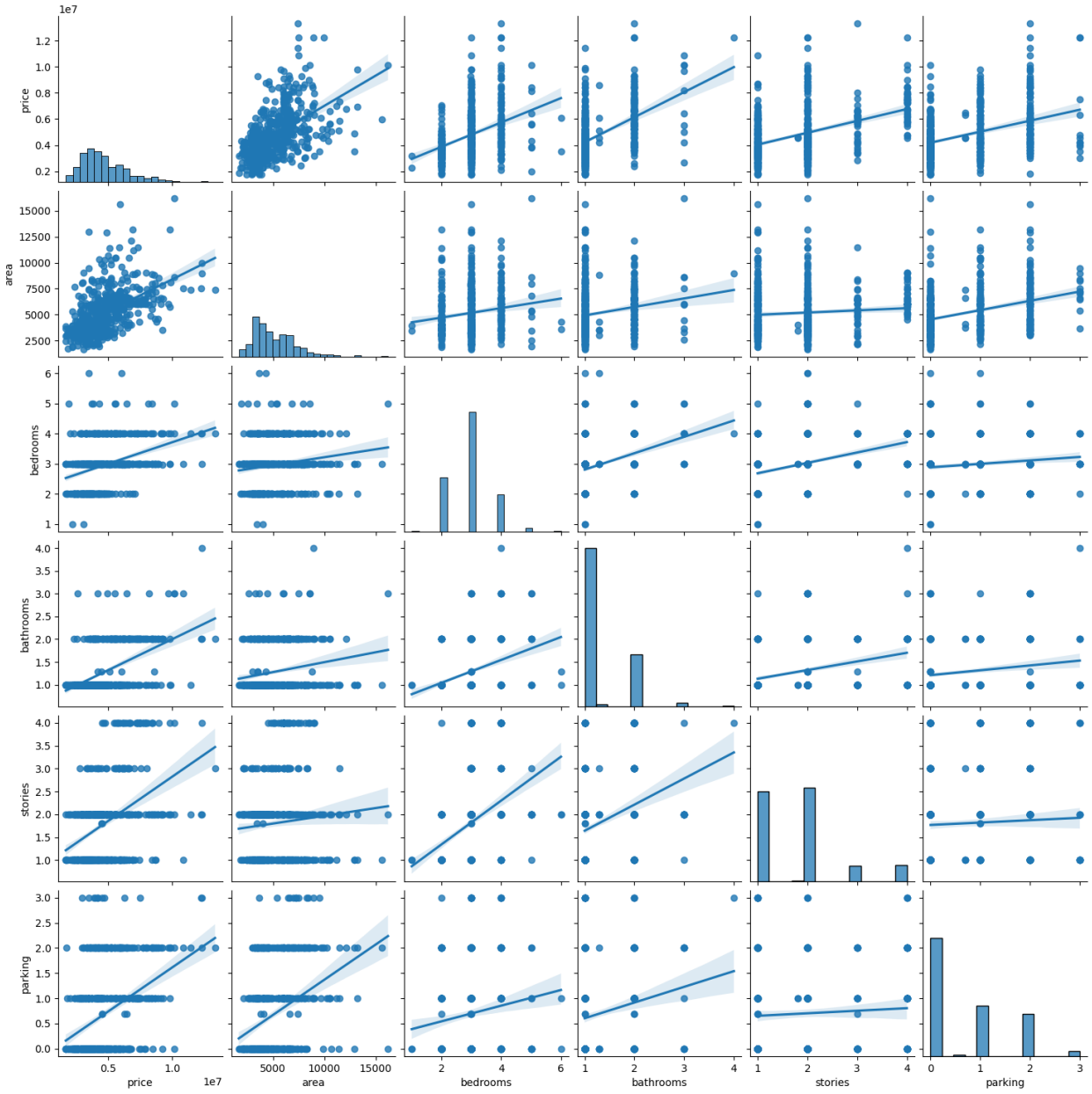
```
price               0
area                0
bedrooms            0
bathrooms           0
stories             0
mainroad            0
guestroom           0
basement            0
hotwaterheating     0
airconditioning     0
parking             0
prefarea            0
furnishingstatus    0
dtype: int64
```

9. Draw the pair plot

A pair plot, also known as a scatterplot matrix, is a matrix of graphs that enables the visualization of the relationship between each pair of variables in a dataset

```
sns.pairplot(dataset, kind="reg")
```

```
<seaborn.axisgrid.PairGrid at 0x7f21d93f9610>
```



10. Divide the dataset into training (75%) and testing (25%).

```
X = dataset.drop("price", axis=1)
y = dataset["price"]
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.25)

len(X_train), len(X_test)
```

```
(408, 137)
```

11. Create subsets as per the given instructions

1. Create the subset with all the columns and first 100 rows

```
first_100 = dataset[:100]
first_100
```

```
       price  area  bedrooms  bathrooms  stories mainroad guestroom
basement  \
0   13300000  7420  4.000000        2.0      3.0      yes        no
no
1   12250000  8960  4.000000        4.0      4.0      yes        no
no
2   12250000  9960  3.000000        2.0      2.0      yes        no
yes
3   12215000  7500  4.000000        2.0      2.0      yes        no
yes
4   11410000  7420  4.000000        1.0      2.0      yes       yes
yes
..       ...   ...       ...        ...      ...      ...       ...
...
95   6300000  4100  3.000000        2.0      3.0      yes        no
no
96   6300000  9000  3.000000        1.0      1.0      yes        no
yes
97   6300000  6400  3.000000        1.0      1.0      yes       yes
yes
98   6293000  6600  2.961111        2.0      3.0      yes        no
no
99   6265000  6000  4.000000        1.0      3.0      yes       yes
yes

   hotwaterheating airconditioning   parking prefarea furnishingstatus

0               no              yes  2.000000      yes        furnished

1               no              yes  3.000000       no        furnished

2               no               no  2.000000      yes   semi-furnished

3               no              yes  3.000000      yes        furnished

4               no              yes  2.000000       no        furnished

..             ...              ...       ...      ...              ...
```

```
95              no             yes  2.000000       no   semi-furnished

96              no              no  1.000000      yes        furnished

97              no             yes  1.000000      yes   semi-furnished

98              no             yes  0.691312      yes      unfurnished

99              no              no  0.000000      yes      unfurnished


[100 rows x 13 columns]
```

2. Create the subset with all the rows and columns where status is furnished

```
furnished = dataset[dataset["furnishingstatus"] == "furnished"]
furnished
```

```
        price  area  bedrooms  bathrooms  stories mainroad guestroom
basement  \
0    13300000  7420       4.0        2.0      3.0      yes        no
no
1    12250000  8960       4.0        4.0      4.0      yes        no
no
3    12215000  7500       4.0        2.0      2.0      yes        no
yes
4    11410000  7420       4.0        1.0      2.0      yes       yes
yes
8     9870000  8100       4.0        1.0      2.0      yes       yes
yes
..        ...   ...       ...        ...      ...      ...       ...
...
509   2590000  3600       2.0        2.0      2.0      yes        no
yes
512   2520000  3000       2.0        1.0      2.0      yes        no
no
522   2380000  2475       3.0        1.0      2.0      yes        no
no
523   2380000  2787       4.0        2.0      2.0      yes        no
no
543   1750000  2910       3.0        1.0      1.0       no        no
no

    hotwaterheating airconditioning  parking prefarea furnishingstatus

0                no             yes      2.0      yes        furnished

1                no             yes      3.0       no        furnished
```

```
3                 no              yes      3.0      yes         furnished
4                 no              yes      2.0       no         furnished
8                 no              yes      2.0      yes         furnished
..               ...              ...      ...      ...               ...
509               no               no      1.0       no         furnished
512               no               no      0.0       no         furnished
522               no               no      0.0       no         furnished
523               no               no      0.0       no         furnished
543               no               no      0.0       no         furnished

[140 rows x 13 columns]
```

3. Create the subset with only five important columns and all rows

```
imp_data = pd.DataFrame(dataset, columns=["area", "guestroom",
"parking", "furnishingstatus", "price"])
imp_data
```

```
     area guestroom  parking furnishingstatus     price
0    7420        no      2.0        furnished  13300000
1    8960        no      3.0        furnished  12250000
2    9960        no      2.0   semi-furnished  12250000
3    7500        no      3.0        furnished  12215000
4    7420       yes      2.0        furnished  11410000
..    ...       ...      ...              ...       ...
540  3000        no      2.0      unfurnished   1820000
541  2400        no      0.0   semi-furnished   1767150
542  3620        no      0.0      unfurnished   1750000
543  2910        no      0.0        furnished   1750000
544  3850        no      0.0      unfurnished   1750000

[545 rows x 5 columns]
```

4. Create the subset with all the samples where area > 1000

```
area_greater_1000 = dataset[dataset["area"] > 1000]
area_greater_1000
```

```
        price  area  bedrooms  bathrooms  stories mainroad guestroom
basement  \
0    13300000  7420       4.0        2.0      3.0      yes        no
no
1    12250000  8960       4.0        4.0      4.0      yes        no
no
2    12250000  9960       3.0        2.0      2.0      yes        no
yes
3    12215000  7500       4.0        2.0      2.0      yes        no
yes
4    11410000  7420       4.0        1.0      2.0      yes       yes
yes
..        ...   ...       ...        ...      ...      ...       ...
...
540   1820000  3000       2.0        1.0      1.0      yes        no
yes
541   1767150  2400       3.0        1.0      1.0       no        no
no
542   1750000  3620       2.0        1.0      1.0      yes        no
no
543   1750000  2910       3.0        1.0      1.0       no        no
no
544   1750000  3850       3.0        1.0      2.0      yes        no
no

    hotwaterheating airconditioning  parking prefarea furnishingstatus

0                no             yes      2.0      yes        furnished

1                no             yes      3.0       no        furnished

2                no              no      2.0      yes   semi-furnished

3                no             yes      3.0      yes        furnished

4                no             yes      2.0       no        furnished

..              ...             ...      ...      ...              ...

540              no              no      2.0       no      unfurnished

541              no              no      0.0       no   semi-furnished

542              no              no      0.0       no      unfurnished

543              no              no      0.0       no        furnished

544              no              no      0.0       no      unfurnished


[545 rows x 13 columns]
```

5. Subset with the rows 10 to 150 with no guestroom and no of bathrooms >=2

```
temp_dataset = dataset[10:151]
subset = temp_dataset[(temp_dataset["guestroom"] == "no") &
(temp_dataset["bathrooms"] >= 2)]
subset
```

```
       price   area  bedrooms  bathrooms  stories mainroad guestroom
basement  \
12   9310000   6550  4.000000        2.0      2.0      yes         no
no
13   9240000   3500  4.000000        2.0      2.0      yes         no
no
14   9240000   7800  2.961111        2.0      2.0      yes         no
no
17   8960000   8500  3.000000        2.0      4.0      yes         no
no
19   8855000   6420  3.000000        2.0      2.0      yes         no
no
28   8400000   7950  5.000000        2.0      2.0      yes         no
yes
29   8400000   5500  4.000000        2.0      2.0      yes         no
yes
30   8400000   7475  3.000000        2.0      4.0      yes         no
no
32   8295000   4880  4.000000        2.0      2.0      yes         no
no
35   8080940   7000  3.000000        2.0      4.0      yes         no
no
36   8043000   7482  3.000000        2.0      3.0      yes         no
no
37   7980000   9000  4.000000        2.0      4.0      yes         no
no
39   7910000   6000  4.000000        2.0      4.0      yes         no
no
41   7840000   6360  3.000000        2.0      4.0      yes         no
no
42   7700000   6480  3.000000        2.0      4.0      yes         no
no
43   7700000   6000  4.000000        2.0      4.0      yes         no
no
44   7560000   6000  4.000000        2.0      4.0      yes         no
no
45   7560000   6000  3.000000        2.0      3.0      yes         no
no
46   7525000   6000  3.000000        2.0      4.0      yes         no
no
48   7455000   4300  3.000000        2.0      2.0      yes         no
```

```
yes
50   7420000   7440  3.000000         2.0       4.0       yes          no
no
53   7350000   5150  3.000000         2.0       4.0       yes          no
no
62   7070000   6240  4.000000         2.0       2.0       yes          no
no
63   7035000   6360  4.000000         2.0       3.0       yes          no
no
65   6930000   8880  3.000000         2.0       2.0       yes          no
yes
67   6895000   7700  3.000000         2.0       1.0       yes          no
no
69   6790000  12090  4.000000         2.0       2.0       yes          no
no
70   6790000   4000  3.000000         2.0       2.0       yes          no
yes
71   6755000   6000  4.000000         2.0       4.0       yes          no
no
73   6685000   6600  2.000000         2.0       4.0       yes          no
yes
75   6650000   4260  4.000000         2.0       2.0       yes          no
no
76   6650000   6420  3.000000         2.0       3.0       yes          no
no
77   6650000   6500  3.000000         2.0       3.0       yes          no
no
81   6615000   4000  3.000000         2.0       2.0       yes          no
yes
82   6615000  10500  3.000000         2.0       1.0       yes          no
yes
83   6580000   6000  3.000000         2.0       4.0       yes          no
no
85   6510000   8250  3.000000         2.0       3.0       yes          no
no
89   6440000   8580  5.000000         3.0       2.0       yes          no
no
93   6300000   7200  3.000000         2.0       1.0       yes          no
yes
94   6300000   6000  4.000000         2.0       4.0       yes          no
no
95   6300000   4100  3.000000         2.0       3.0       yes          no
no
98   6293000   6600  2.961111         2.0       3.0       yes          no
no
100  6230000   6600  3.000000         2.0       1.0       yes          no
yes
106  6160000   5450  4.000000         2.0       1.0       yes          no
yes
```

```
122  5950000   6254  4.000000         2.0      1.0      yes        no
yes
123  5950000   7320  4.000000         2.0      2.0      yes        no
no
124  5950000   6525  3.000000         2.0      4.0      yes        no
no
127  5880000   6500  3.000000         2.0      3.0      yes        no
no
135  5775000   6000  3.000000         2.0      4.0      yes        no
no
136  5740000   5400  4.000000         2.0      2.0      yes        no
no
140  5740000   5800  3.000000         2.0      4.0      yes        no
no
142  5600000  10500  4.000000         2.0      2.0      yes        no
no
143  5600000   4800  5.000000         2.0      3.0       no        no
yes
147  5600000   5500  3.000000         2.0      2.0      yes        no
no
149  5600000   6600  4.000000         2.0      1.0      yes        no
yes

    hotwaterheating airconditioning   parking prefarea
furnishingstatus
12               no              yes  1.000000      yes   semi-
furnished
13              yes               no  2.000000       no
furnished
14               no               no  0.000000      yes   semi-
furnished
17               no              yes  2.000000       no
furnished
19               no              yes  1.000000      yes   semi-
furnished
28              yes               no  2.000000       no
unfurnished
29               no              yes  1.000000      yes   semi-
furnished
30               no              yes  2.000000       no
unfurnished
32               no              yes  1.000000      yes
furnished
35               no              yes  2.000000       no
furnished
36              yes               no  1.000000      yes
furnished
37               no              yes  2.000000       no
furnished
```

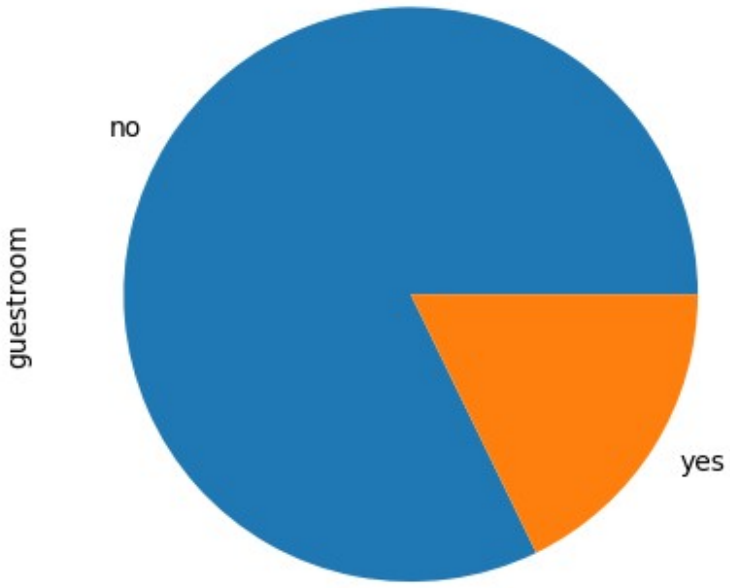
```
39                no              yes  1.000000        no   semi-
furnished
41                no              yes  0.000000       yes
furnished
42                no              yes  2.000000        no
unfurnished
43                no               no  2.000000        no   semi-
furnished
44                no              yes  1.000000        no
furnished
45                no              yes  0.000000        no   semi-
furnished
46                no              yes  1.000000        no
furnished
48                no               no  1.000000        no
unfurnished
50                no               no  1.000000       yes
unfurnished
53                no              yes  2.000000        no   semi-
furnished
62                no              yes  1.000000        no
furnished
63                no              yes  2.000000       yes
furnished
65                no              yes  1.000000        no
furnished
67                no               no  2.000000        no
unfurnished
69                no               no  2.000000       yes
furnished
70                no              yes  0.000000       yes   semi-
furnished
71                no              yes  0.000000        no
unfurnished
73                no               no  0.000000       yes
furnished
75               yes               no  0.000000        no   semi-
furnished
76                no              yes  0.000000       yes
furnished
77                no              yes  0.000000       yes
furnished
81                no              yes  1.000000        no   semi-
furnished
82                no              yes  1.000000       yes
furnished
83                no              yes  0.000000        no   semi-
furnished
85                no              yes  0.000000        no
```

```
furnished
89                no                no  2.000000        no
furnished
93                no               yes  3.000000        no    semi-
furnished
94                no                no  1.000000        no    semi-
furnished
95                no               yes  2.000000        no    semi-
furnished
98                no               yes  0.691312       yes
unfurnished
100               no               yes  0.000000       yes
unfurnished
106               no               yes  0.000000       yes    semi-
furnished
122               no                no  1.000000       yes    semi-
furnished
123               no                no  0.000000        no
furnished
124               no                no  1.000000        no
furnished
127               no               yes  0.000000        no
unfurnished
135               no               yes  0.000000        no
unfurnished
136               no               yes  2.000000        no
unfurnished
140               no               yes  0.000000        no
unfurnished
142               no                no  1.000000        no    semi-
furnished
143              yes                no  0.000000        no
unfurnished
147               no                no  1.000000        no    semi-
furnished
149               no                no  0.000000       yes    semi-
furnished
```

Draw pie chart showing no of guest-rooms

```
guestroom = dataset['guestroom'].value_counts()
guestroom.plot(kind="pie")
```

```
<AxesSubplot:ylabel='guestroom'>
```


guestroom
no
yes