In [2]:
```
import pandas as pd
import numpy as np
import seaborn as sns
```

In [ ]:
```
```

In [4]:
```
df = pd.read_csv("C:/Users/rutuj/Desktop/BE_Project/kc_house_data.csv")
df.head()
```

Out[4]:

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors
0	7129300520	20141013T000000	221900.0	3	1.00	1180	5650	1.0
1	6414100192	20141209T000000	538000.0	3	2.25	2570	7242	2.0
2	5631500400	20150225T000000	180000.0	2	1.00	770	10000	1.0
3	2487200875	20141209T000000	604000.0	4	3.00	1960	5000	1.0
4	1954400510	20150218T000000	510000.0	3	2.00	1680	8080	1.0

5 rows × 21 columns

In [5]:
```
df_original = df
```

In [7]:
```
df.columns
```

Out[7]:
```
Index(['id', 'date', 'price', 'bedrooms', 'bathrooms', 'sqft_living',
       'sqft_lot', 'floors', 'waterfront', 'view', 'condition', 'grade',
       'sqft_above', 'sqft_basement', 'yr_built', 'yr_renovated', 'zipcode',
       'lat', 'long', 'sqft_living15', 'sqft_lot15'],
      dtype='object')
```

In [8]:
```
type('id')
```

Out[8]:
```
str
```

In [9]:

```
df.describe()
```

Out[9]:

	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	
count	2.161300e+04	2.161300e+04	21613.000000	21613.000000	21613.000000	2.161300e+04	216
mean	4.580302e+09	5.400881e+05	3.370842	2.114757	2079.899736	1.510697e+04	
std	2.876566e+09	3.671272e+05	0.930062	0.770163	918.440897	4.142051e+04	
min	1.000102e+06	7.500000e+04	0.000000	0.000000	290.000000	5.200000e+02	
25%	2.123049e+09	3.219500e+05	3.000000	1.750000	1427.000000	5.040000e+03	
50%	3.904930e+09	4.500000e+05	3.000000	2.250000	1910.000000	7.618000e+03	
75%	7.308900e+09	6.450000e+05	4.000000	2.500000	2550.000000	1.068800e+04	
max	9.900000e+09	7.700000e+06	33.000000	8.000000	13540.000000	1.651359e+06	

In [10]:

```
print(df.dtypes)
```

```
id                 int64
date              object
price            float64
bedrooms           int64
bathrooms        float64
sqft_living        int64
sqft_lot           int64
floors           float64
waterfront         int64
view               int64
condition          int64
grade              int64
sqft_above         int64
sqft_basement      int64
yr_built           int64
yr_renovated       int64
zipcode            int64
lat              float64
long             float64
sqft_living15      int64
sqft_lot15         int64
dtype: object
```

In [11]:

```
df= df.drop(['id','sqft_living15','sqft_lot15'],axis=1)
```

In [12]:
```python
df
```

Out[12]:

	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfron
0	20141013T000000	221900.0	3	1.00	1180	5650	1.0	(
1	20141209T000000	538000.0	3	2.25	2570	7242	2.0	(
2	20150225T000000	180000.0	2	1.00	770	10000	1.0	(
3	20141209T000000	604000.0	4	3.00	1960	5000	1.0	(
4	20150218T000000	510000.0	3	2.00	1680	8080	1.0	(
...	...	...	...	...	...	...	...	..
21608	20140521T000000	360000.0	3	2.50	1530	1131	3.0	(
21609	20150223T000000	400000.0	4	2.50	2310	5813	2.0	(
21610	20140623T000000	402101.0	2	0.75	1020	1350	2.0	(
21611	20150116T000000	400000.0	3	2.50	1600	2388	2.0	(
21612	20141015T000000	325000.0	2	0.75	1020	1076	2.0	(

21613 rows × 18 columns

In [13]:
```python
df['date'].astype(str)
```

Out[13]:
```
0        20141013T000000
1        20141209T000000
2        20150225T000000
3        20141209T000000
4        20150218T000000
              ...
21608    20140521T000000
21609    20150223T000000
21610    20140623T000000
21611    20150116T000000
21612    20141015T000000
Name: date, Length: 21613, dtype: object
```

In [14]:

```
print(df.dtypes)
```

```
date              object
price            float64
bedrooms           int64
bathrooms        float64
sqft_living        int64
sqft_lot           int64
floors           float64
waterfront         int64
view               int64
condition          int64
grade              int64
sqft_above         int64
sqft_basement      int64
yr_built           int64
yr_renovated       int64
zipcode            int64
lat              float64
long             float64
dtype: object
```

In [15]:

```
df['date1'] = df['date'].str[0:8]
df['date2'] = df['date'].str[8:]
```

In [16]:

```
df.head()
```

Out[16]:

	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	vi
0	20141013T000000	221900.0	3	1.00	1180	5650	1.0	0	
1	20141209T000000	538000.0	3	2.25	2570	7242	2.0	0	
2	20150225T000000	180000.0	2	1.00	770	10000	1.0	0	
3	20141209T000000	604000.0	4	3.00	1960	5000	1.0	0	
4	20150218T000000	510000.0	3	2.00	1680	8080	1.0	0	

In [17]:

```
df['dy']=df['date1'].str[0:4]
df['dm']=df['date1'].str[4:6]
df['dd']=df['date1'].str[6:8]
```

In [18]:
```
df.head()
```

Out[18]:

	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	vi
0	20141013T000000	221900.0	3	1.00	1180	5650	1.0	0	
1	20141209T000000	538000.0	3	2.25	2570	7242	2.0	0	
2	20150225T000000	180000.0	2	1.00	770	10000	1.0	0	
3	20141209T000000	604000.0	4	3.00	1960	5000	1.0	0	
4	20150218T000000	510000.0	3	2.00	1680	8080	1.0	0	

5 rows × 23 columns

In [19]:
```
df=df.drop(['date','date1','date2','dd'],axis=1)
```

In [20]:
```
df.head()
```

Out[20]:

	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	gra
0	221900.0	3	1.00	1180	5650	1.0	0	0	3	
1	538000.0	3	2.25	2570	7242	2.0	0	0	3	
2	180000.0	2	1.00	770	10000	1.0	0	0	3	
3	604000.0	4	3.00	1960	5000	1.0	0	0	5	
4	510000.0	3	2.00	1680	8080	1.0	0	0	3	

In [21]:
```
df['price']=df['price'].astype(int)
```

In [22]:

```
print(df.dtypes)
```

```
price              int32
bedrooms           int64
bathrooms        float64
sqft_living        int64
sqft_lot           int64
floors           float64
waterfront         int64
view               int64
condition          int64
grade              int64
sqft_above         int64
sqft_basement      int64
yr_built           int64
yr_renovated       int64
zipcode            int64
lat              float64
long             float64
dy                object
dm                object
dtype: object
```

In [23]:

```
df['dy']=df['dy'].astype(int)
df['dm']=df['dm'].astype(int)
```

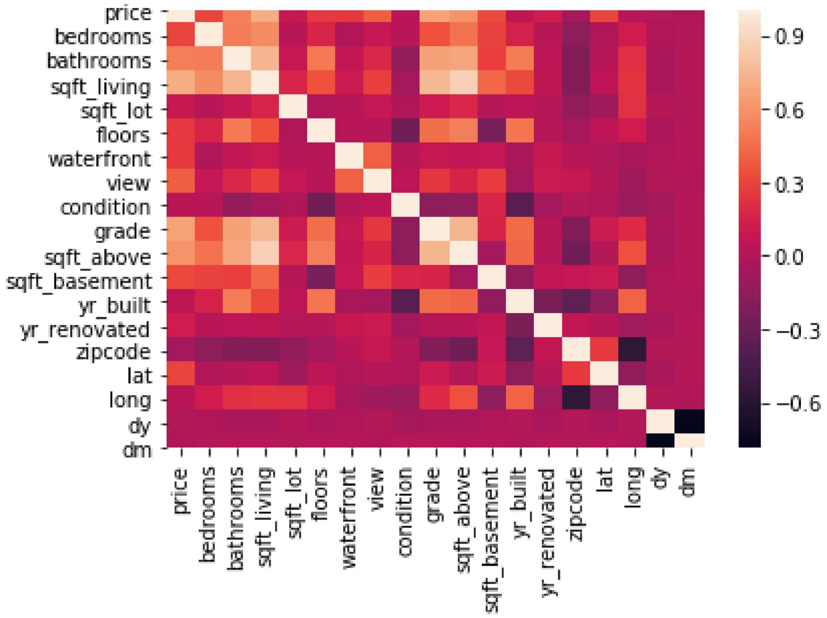
In [24]:

```
df3=df.corr()
```

In [25]:
```
import seaborn as sns

Var_Corr = df.corr()
# plot the heatmap and annotation on it
sns.heatmap(Var_Corr, xticklabels=Var_Corr.columns, yticklabels=Var_Corr.colum
ns, )
```

Out[25]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1c21a930d48>
```



In [26]:
```
df1=df.corr(method="pearson")
```

In [27]:
```
#df.corr(method="kendall")
```

In [28]:
```
#df1=df.corr(method="spearman")
```

```
In [29]: print(df3['price'])
```

```
price            1.000000
bedrooms         0.308350
bathrooms        0.525138
sqft_living      0.702035
sqft_lot         0.089661
floors           0.256794
waterfront       0.266369
view             0.397293
condition        0.036362
grade            0.667434
sqft_above       0.605567
sqft_basement    0.323816
yr_built         0.054012
yr_renovated     0.126434
zipcode         -0.053203
lat              0.307003
long             0.021626
dy               0.003576
dm              -0.010081
Name: price, dtype: float64
```

```
In [30]: df2= df3['price']
```

```
In [31]: df2
```

```
Out[31]: price            1.000000
         bedrooms         0.308350
         bathrooms        0.525138
         sqft_living      0.702035
         sqft_lot         0.089661
         floors           0.256794
         waterfront       0.266369
         view             0.397293
         condition        0.036362
         grade            0.667434
         sqft_above       0.605567
         sqft_basement    0.323816
         yr_built         0.054012
         yr_renovated     0.126434
         zipcode         -0.053203
         lat              0.307003
         long             0.021626
         dy               0.003576
         dm              -0.010081
         Name: price, dtype: float64
```

```
In [32]: df2=df2.sort_values(ascending= False)
In [33]: df2
```

```
Out[33]: price            1.000000
         sqft_living      0.702035
         grade            0.667434
         sqft_above       0.605567
         bathrooms        0.525138
         view             0.397293
         sqft_basement    0.323816
         bedrooms         0.308350
         lat              0.307003
         waterfront       0.266369
         floors           0.256794
         yr_renovated     0.126434
         sqft_lot         0.089661
         yr_built         0.054012
         condition        0.036362
         long             0.021626
         dy               0.003576
         dm              -0.010081
         zipcode         -0.053203
         Name: price, dtype: float64
```

```
In [ ]:
```

```
In [34]: import os
         print("current",os.getcwd())
```

```
         current C:\Users\rutuj\BE Project G26
```

```
In [35]: os.chdir("C:/Users/rutuj/Desktop/BE_Project")
```

```
In [36]: print("current",os.getcwd())
```

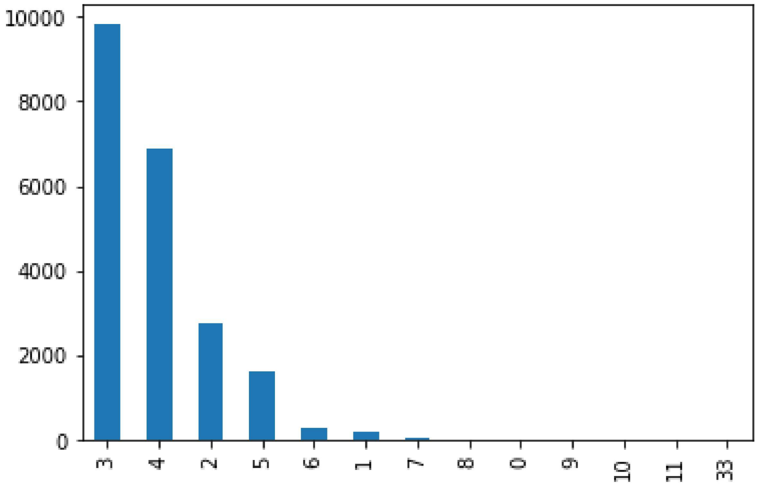
```
         current C:\Users\rutuj\Desktop\BE_Project
```

```
In [37]: df2.to_excel("corelation.xlsx")
```

```
In [38]: #ax = df2.plot.bar(x=)
```

```
In [39]: %matplotlib inline
In [40]: df['bedrooms'].value_counts().plot(kind='bar')

Out[40]: <matplotlib.axes._subplots.AxesSubplot at 0x1c21ae5a448>
```



```
In [41]: df['bedrooms'].value_counts(ascending=False)

Out[41]: 3     9824
         4     6882
         2     2760
         5     1601
         6      272
         1      199
         7       38
         8       13
         0       13
         9        6
         10       3
         11       1
         33       1
         Name: bedrooms, dtype: int64
```

```
In [42]: df.shape

Out[42]: (21613, 19)
```

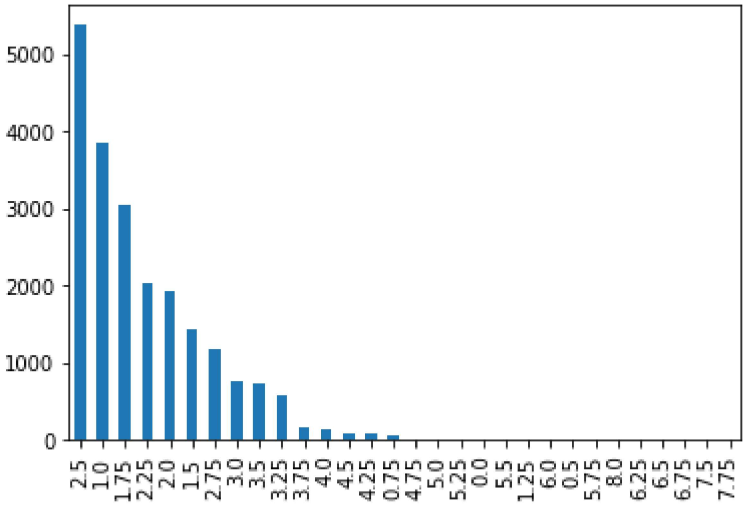
```
In [43]: df_bedroom =df[(df['bedrooms'] != 0) & (df['bedrooms'] != 11) & (df['bedrooms'
         ] != 33)]
```

```
In [44]: df_bedroom.shape

Out[44]: (21598, 19)
```

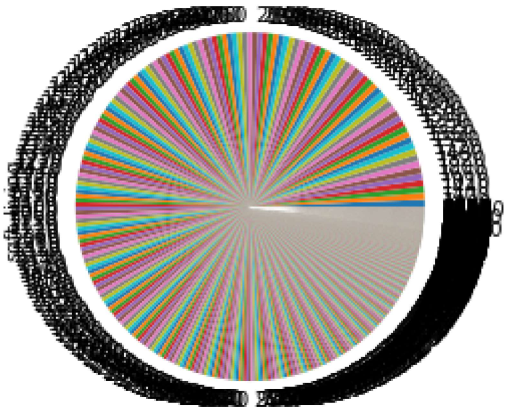
In [49]:

```
#bathrooms
df_bathroom = df['bathrooms'].value_counts().plot(kind='bar')
```



In [58]:

```
#sqft_living
df_bathroom = df['sqft_living'].value_counts().plot(kind='pie')
```



In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: