In [2]:

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

In [3]:

```
data= pd.read_csv("housing2.csv - housing2.csv.csv",sep=',',encoding="utf-8")
```

In [4]:

```
data
```

Out[4]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_
0	-122.23	37.88	41.0	880	129.0	322.0	126	8.3252	452600	N
1	-122.22	37.86	21.0	7099	1106.0	2401.0	1138	8.3014	358500	N
2	-122.24	37.85	52.0	1467	190.0	496.0	177	7.2574	352100	N
3	-122.25	37.85	52.0	1274	235.0	558.0	219	5.6431	341300	N
4	-122.25	37.85	NaN	1627	280.0	NaN	259	3.8462	342200	N
...	...	...	...	...	...	...	...	...	...	
20635	-121.09	39.48	25.0	1665	374.0	845.0	330	1.5603	78100	
20636	-121.21	39.49	18.0	697	150.0	356.0	114	2.5568	77100	
20637	-121.22	39.43	17.0	2254	485.0	1007.0	433	1.7000	92300	
20638	-121.32	39.43	18.0	1860	409.0	741.0	349	1.8672	84700	
20639	-121.24	39.37	16.0	2785	616.0	1387.0	530	2.3886	89400	

20640 rows × 11 columns

In [5]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20382 non-null  float64
 3   total_rooms         20640 non-null  int64  
 4   total_bedrooms      15758 non-null  float64
 5   population          20596 non-null  float64
 6   households          19335 non-null  object 
 7   median_income       17873 non-null  float64
 8   median_house_value  20640 non-null  int64  
 9   ocean_proximity     20640 non-null  object 
 10  gender              16620 non-null  object 
dtypes: float64(6), int64(2), object(3)
memory usage: 1.7+ MB
```

In [6]:

```
data.describe()
```

Out[6]:

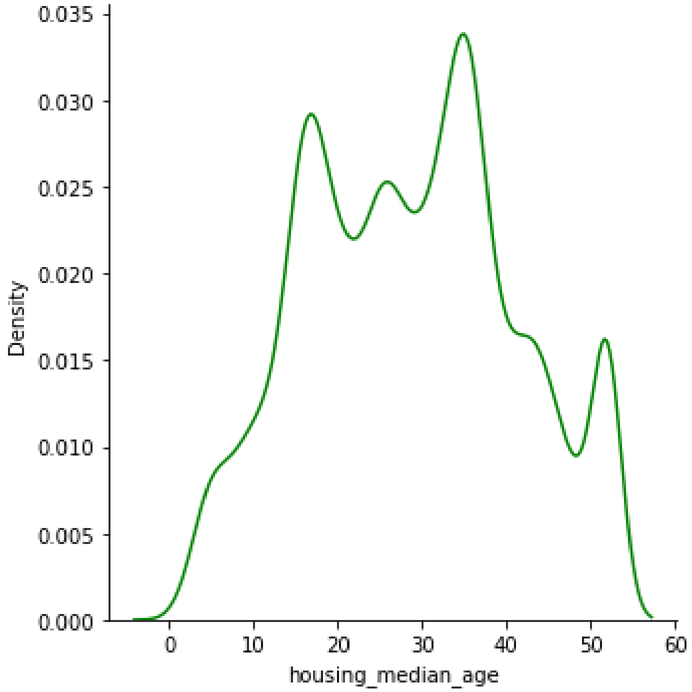
	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	median_income	median_house_value
count	20640.000000	20640.000000	20382.000000	20640.000000	15758.000000	20596.000000	17873.000000	20640.000000
mean	-119.569704	35.631861	28.676283	2635.763081	539.920104	1424.928724	3.939403	206855.816909
std	2.003532	2.135952	12.589284	2181.615252	419.834171	1132.237768	1.943517	115395.615874
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	0.499900	14999.000000
25%	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.000000	2.598600	119600.000000
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.000000	3.587100	179700.000000
75%	-118.010000	37.710000	37.000000	3148.000000	652.000000	1725.000000	4.830400	264725.000000
max	-114.310000	41.950000	52.000000	39320.000000	6210.000000	35682.000000	15.000100	500001.000000

In [7]:

```
sns.displot(data["housing_median_age"], kind="kde",color="green")
```

Out[7]:

```
<seaborn.axisgrid.FacetGrid at 0x3c6c9b4d00>
```



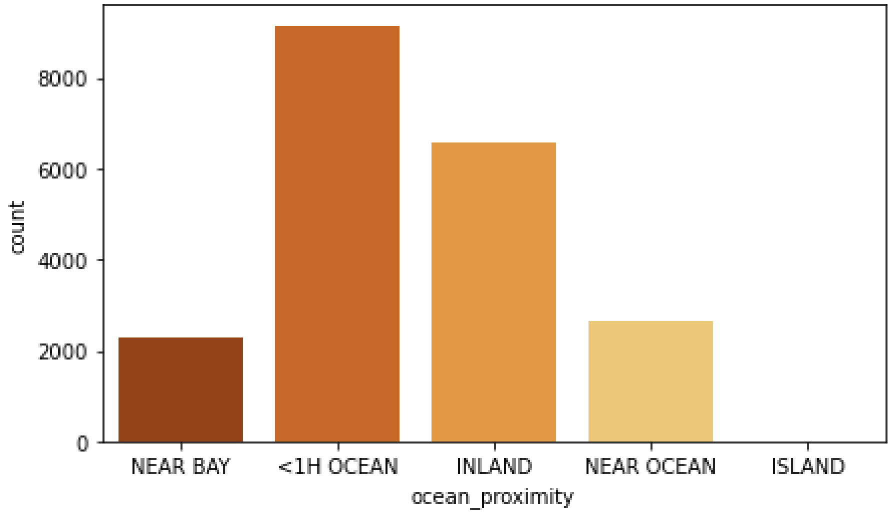
In [8]:
```
data.ocean_proximity.unique()
```

Out[8]:
```
array(['NEAR BAY', '<1H OCEAN', 'INLAND', 'NEAR OCEAN', 'ISLAND'],
      dtype=object)
```

In [9]:
```
plt.figure(figsize=(7,4))
sns.countplot(data['ocean_proximity'], palette = "YlOrBr_r")
```

```
C:\Users\hsd\anaconda3\anaconda3.64\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following va
riable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other a
rguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[9]:
```
<AxesSubplot:xlabel='ocean_proximity', ylabel='count'>
```



In [10]:

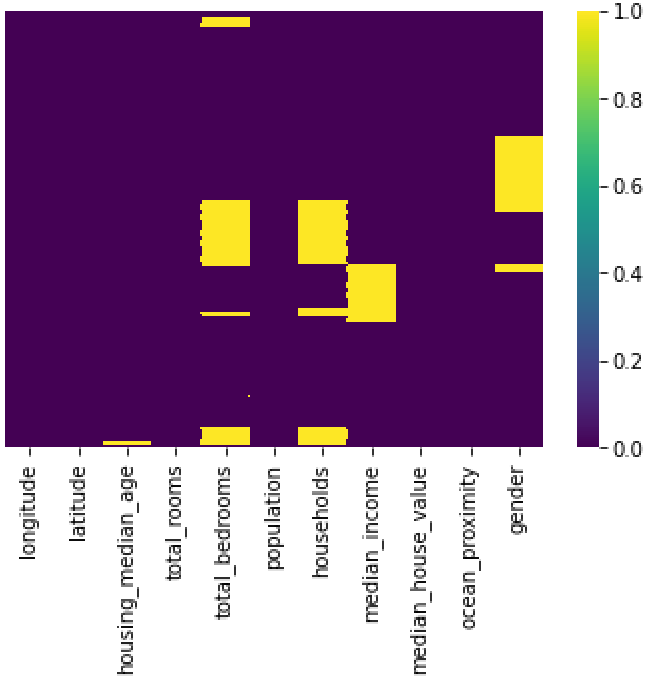
```
data.isnull().sum()
```

Out[10]:

```
longitude                0
latitude                 0
housing_median_age     258
total_rooms              0
total_bedrooms        4882
population              44
households            1305
median_income         2767
median_house_value       0
ocean_proximity          0
gender                4020
dtype: int64
```

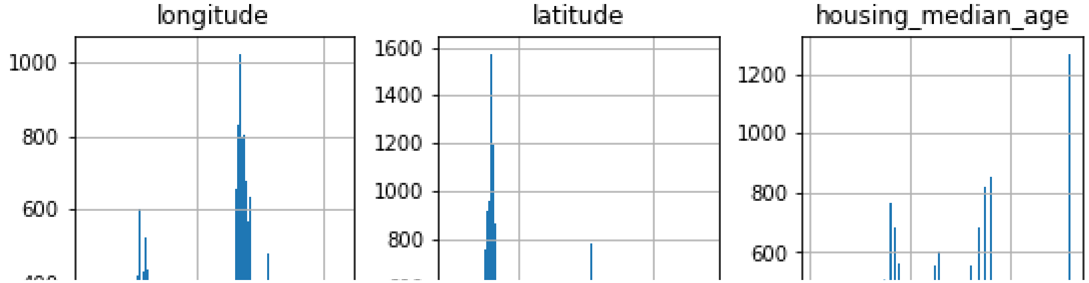
In [40]:

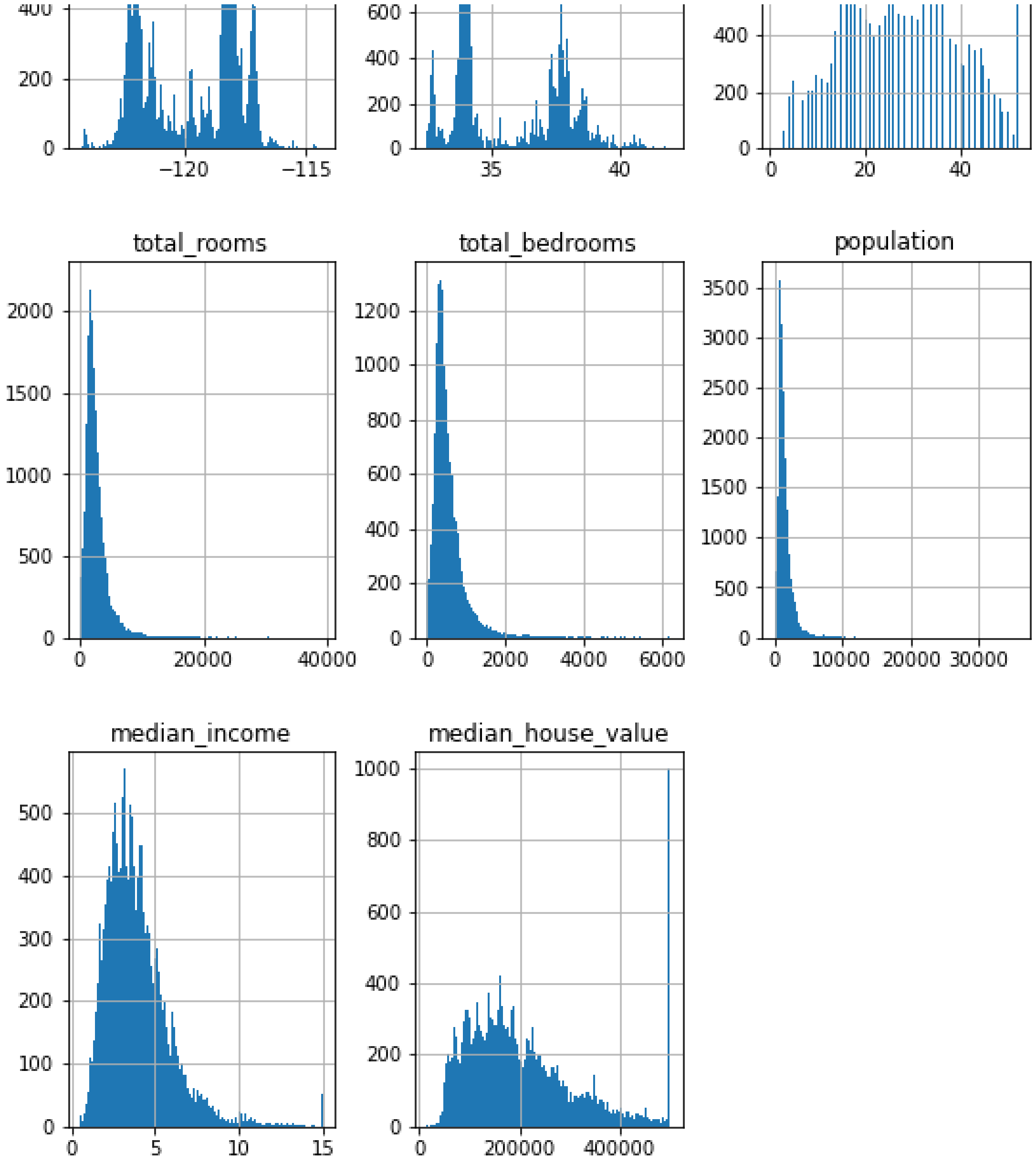
```
sns.heatmap(data.isnull(),cmap='viridis',yticklabels=False)

plt.show('missing data')
plt.show()
```



```
In [12]: data.hist(bins=140,figsize=(9,13))
```

```
Out[12]: array([[<AxesSubplot:title={'center':'longitude'}>,
                 <AxesSubplot:title={'center':'latitude'}>,
                 <AxesSubplot:title={'center':'housing_median_age'}>],
                [<AxesSubplot:title={'center':'total_rooms'}>,
                 <AxesSubplot:title={'center':'total_bedrooms'}>,
                 <AxesSubplot:title={'center':'population'}>],
                [<AxesSubplot:title={'center':'median_income'}>,
                 <AxesSubplot:title={'center':'median_house_value'}>,
                 <AxesSubplot:>]], dtype=object)
```



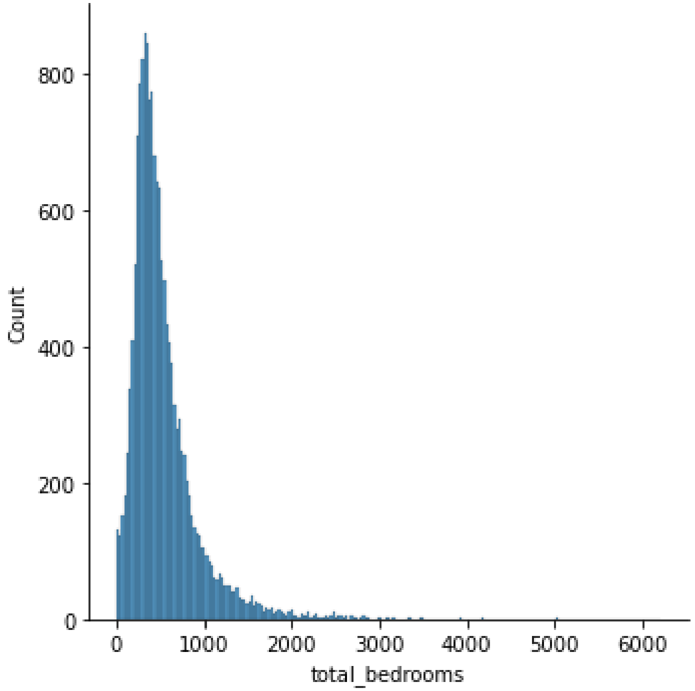


In [13]:

```
sns.displot(x='total_bedrooms',data=data)
```

```
<seaborn.axisgrid.FacetGrid at 0x3c6e2eb640>
```

Out[13]:



In [14]:

```
data.dropna()
```

Out[14]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_
0	-122.23	37.88	41.0	880	129.0	322.0	126	8.3252	452600	N
1	-122.22	37.86	21.0	7099	1106.0	2401.0	1138	8.3014	358500	N
2	-122.24	37.85	52.0	1467	190.0	496.0	177	7.2574	352100	N
3	-122.25	37.85	52.0	1274	235.0	558.0	219	5.6431	341300	N
17	-122.27	37.85	52.0	1228	293.0	648.0	303	2.1202	155500	N
...	...	...	...	...	...	...	...	...	...	
20635	-121.09	39.48	25.0	1665	374.0	845.0	330	1.5603	78100	
20636	-121.21	39.49	18.0	697	150.0	356.0	114	2.5568	77100	

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_
20637	-121.22	39.43	17.0	2254	485.0	1007.0	433	1.7000	92300	
20638	-121.32	39.43	18.0	1860	409.0	741.0	349	1.8672	84700	
20639	-121.24	39.37	16.0	2785	616.0	1387.0	530	2.3886	89400	

10177 rows × 11 columns

In [22]:

```
data['households']=data['households'].replace('no',np.nan)
```

In [23]:

```
data['households'].value_counts()
```

Out[23]:

```
282     47
375     46
306     45
380     45
335     42
        ..
1399     1
3073     1
985      1
2289     1
1843     1
Name: households, Length: 1702, dtype: int64
```

In [24]:

```
data.dtypes
```

Out[24]:

```
longitude             float64
latitude              float64
housing_median_age    float64
total_rooms             int64
total_bedrooms        float64
population            float64
households             object
median_income         float64
median_house_value      int64
ocean_proximity        object
gender                 object
dtype: object
```

In [25]:

```
data['households']=pd.to_numeric(data['households'])
```

In [26]:

```
data[pd.isnull(data['households'])]
```

Out[26]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_
7	-122.25	37.84	NaN	3104	NaN	NaN	NaN	3.1200	241400	N
8	-122.26	37.84	42.0	2555	NaN	NaN	NaN	2.0804	226700	N
9	-122.25	37.84	52.0	3549	NaN	NaN	NaN	3.6912	261100	N
10	-122.26	37.85	52.0	2202	NaN	NaN	NaN	3.2031	281500	N
11	-122.26	37.85	52.0	3503	NaN	NaN	NaN	3.2705	241800	N
...	...	...	...	...	...	...	...	...	...	
20627	-121.32	39.13	NaN	358	NaN	169.0	NaN	3.0000	162500	
20628	-121.48	39.10	NaN	2043	NaN	1018.0	NaN	2.5952	92400	
20629	-121.39	39.12	NaN	10035	NaN	6912.0	NaN	2.0943	108300	
20630	-121.32	39.29	NaN	2640	NaN	1257.0	NaN	3.5673	112000	
20631	-121.40	39.33	15.0	2655	493.0	1200.0	NaN	3.5179	107200	

4385 rows × 11 columns

In [42]:

```
data['housing_median_age'].replace(np.nan,data['housing_median_age'].mean(),inplace=True)
```

In [43]:

```
data.isnull().sum()
```

Out[43]:

```
longitude                0
latitude                 0
housing_median_age       0
total_rooms              0
total_bedrooms        4882
population              44
households            4385
median_income         2767
median_house_value       0
```

```
ocean_proximity          0
gender                4020
dtype: int64
```

In [44]:

```
data['total_bedrooms'].replace(np.nan,data['total_bedrooms'].mean(),inplace=True)
```

In [45]:

```
data.isnull().sum()
```

Out[45]:

```
longitude                0
latitude                 0
housing_median_age       0
total_rooms              0
total_bedrooms           0
population              44
households            4385
median_income         2767
median_house_value       0
ocean_proximity          0
gender                4020
dtype: int64
```

In [46]:

```
data['population'].replace(np.nan,data['population'].mean(),inplace=True)
```

In [47]:

```
data.isnull().sum()
```

Out[47]:

```
longitude                0
latitude                 0
housing_median_age       0
total_rooms              0
total_bedrooms           0
population               0
households            4385
median_income         2767
median_house_value       0
ocean_proximity          0
gender                4020
dtype: int64
```

In [48]:

```
data['households'].replace(np.nan,data['households'].mean(),inplace=True)
```

In [49]:

```
data.isnull().sum()
```

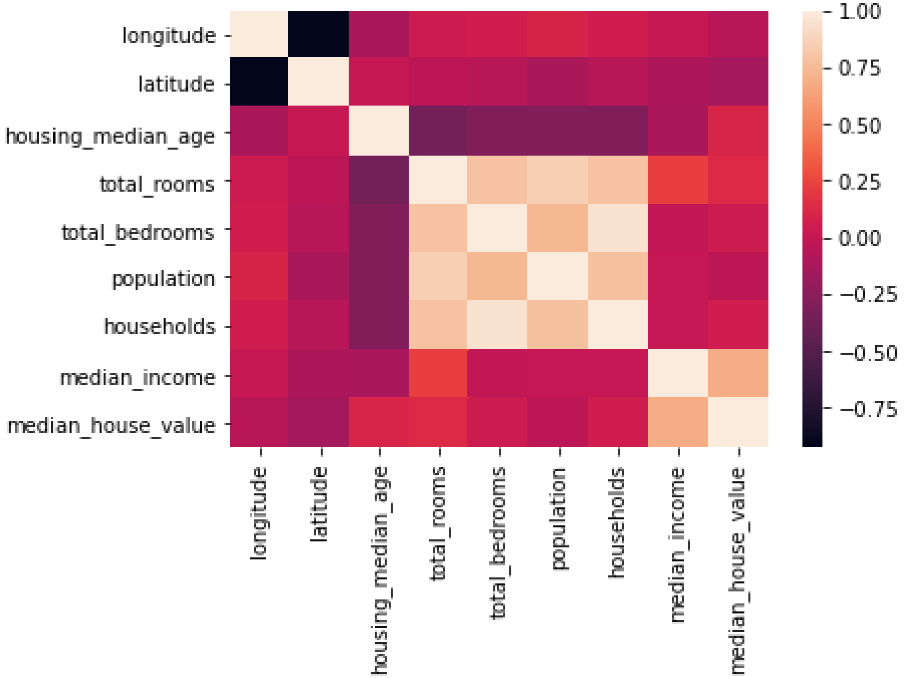
```
Out[49]: longitude                0
         latitude                 0
         housing_median_age       0
         total_rooms              0
         total_bedrooms           0
         population               0
         households               0
         median_income         2767
         median_house_value       0
         ocean_proximity          0
         gender                4020
         dtype: int64
```

```
In [50]: data.corr()
```

Out[50]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_hou
longitude	1.000000	-0.924664	-0.106884	0.044568	0.063468	0.100253	0.053426	0.011478	-
latitude	-0.924664	1.000000	0.009689	-0.036100	-0.054250	-0.109120	-0.057212	-0.103528	-
housing_median_age	-0.106884	0.009689	1.000000	-0.356480	-0.296786	-0.291137	-0.293285	-0.117026	
total_rooms	0.044568	-0.036100	-0.356480	1.000000	0.793059	0.856124	0.794263	0.220357	
total_bedrooms	0.063468	-0.054250	-0.296786	0.793059	1.000000	0.743033	0.947697	-0.009918	
population	0.100253	-0.109120	-0.291137	0.856124	0.743033	1.000000	0.782637	0.001809	-
households	0.053426	-0.057212	-0.293285	0.794263	0.947697	0.782637	1.000000	0.005795	
median_income	0.011478	-0.103528	-0.117026	0.220357	-0.009918	0.001809	0.005795	1.000000	
median_house_value	-0.045967	-0.144160	0.106648	0.134153	0.044949	-0.024351	0.058656	0.688625	

```
In [51]: sns.heatmap(data.corr())
         plt.show('Heat map correlation')
         plt.show()
```


longitude
latitude
housing_median_age
total_rooms
total_bedrooms
population
households
median_income
median_house_value
longitude
latitude
housing_median_age
total_rooms
total_bedrooms
population
households
median_income
median_house_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

In [ ]: