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

In [16]:
```
data = pd.read_csv("housing2.csv", sep=',',encoding="utf-8")
```

In [18]:
```
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 [3]:
```
data.describe()
```

Out[3]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	populati
count	20640.000000	20640.000000	20382.000000	20640.000000	15758.000000	20596.0000
mean	-119.569704	35.631861	28.676283	2635.763081	539.920104	1424.9287
std	2.003532	2.135952	12.589284	2181.615252	419.834171	1132.2377
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.0000
25%	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.0000
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.0000
75%	-118.010000	37.710000	37.000000	3148.000000	652.000000	1725.0000
max	-114.310000	41.950000	52.000000	39320.000000	6210.000000	35682.0000

In [19]:
```
data.columns
```

Out[19]:
```
Index(['longitude', 'latitude', 'housing_median_age', 'total_rooms',
       'total_bedrooms', 'population', 'households', 'median_income',
       'median_house_value', 'ocean_proximity', 'gender'],
      dtype='object')
```

In [20]:
```
1 data['ocean_proximity'].value_counts()
```

Out[20]:
```
<1H OCEAN     9136
INLAND        6551
NEAR OCEAN    2658
NEAR BAY      2290
ISLAND           5
Name: ocean_proximity, dtype: int64
```

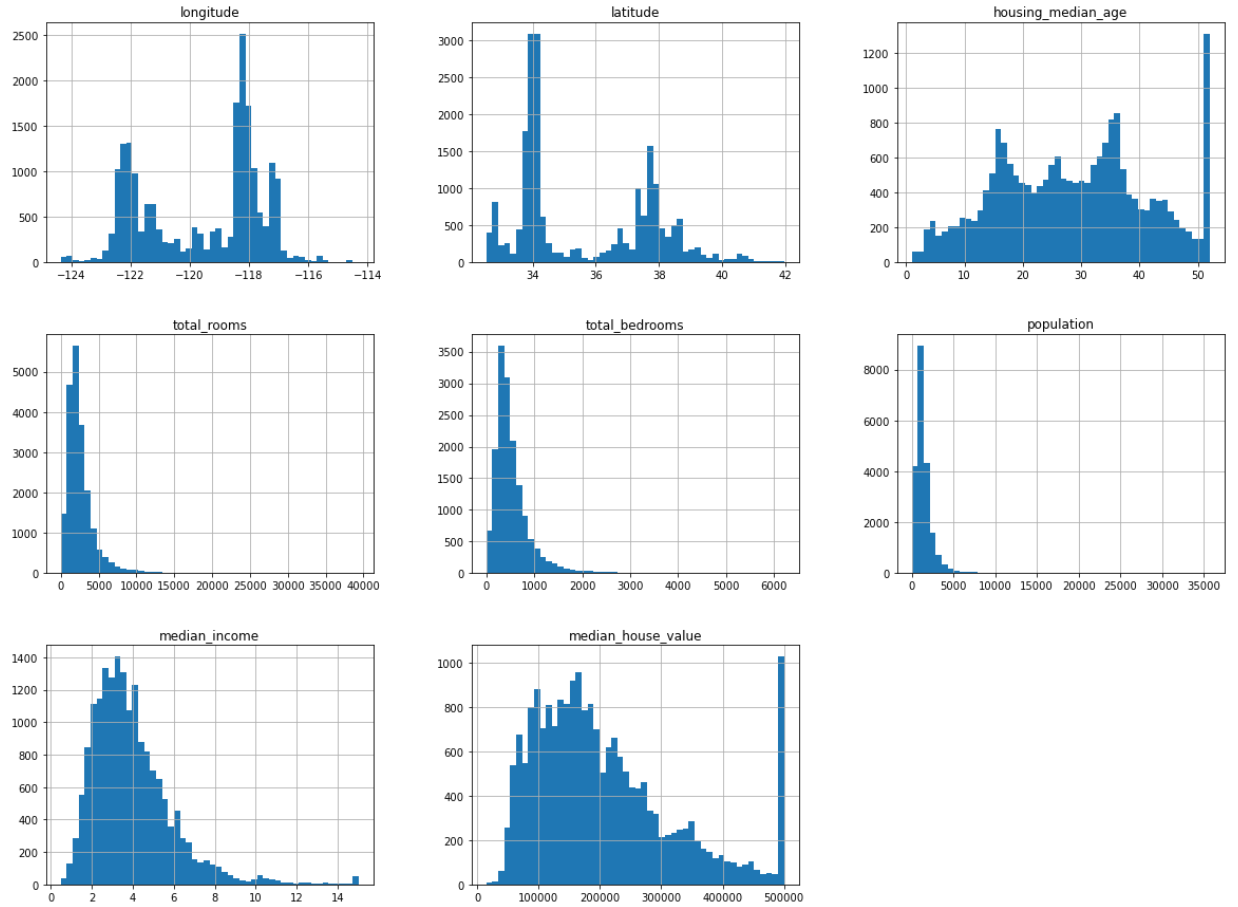
In [47]:
```
1 data['households'].value_counts()
```

Out[47]:
```
no      3080
282       47
375       46
306       45
380       45
         ...
2392       1
1577       1
2838       1
1747       1
2159       1
Name: households, Length: 1703, dtype: int64
```

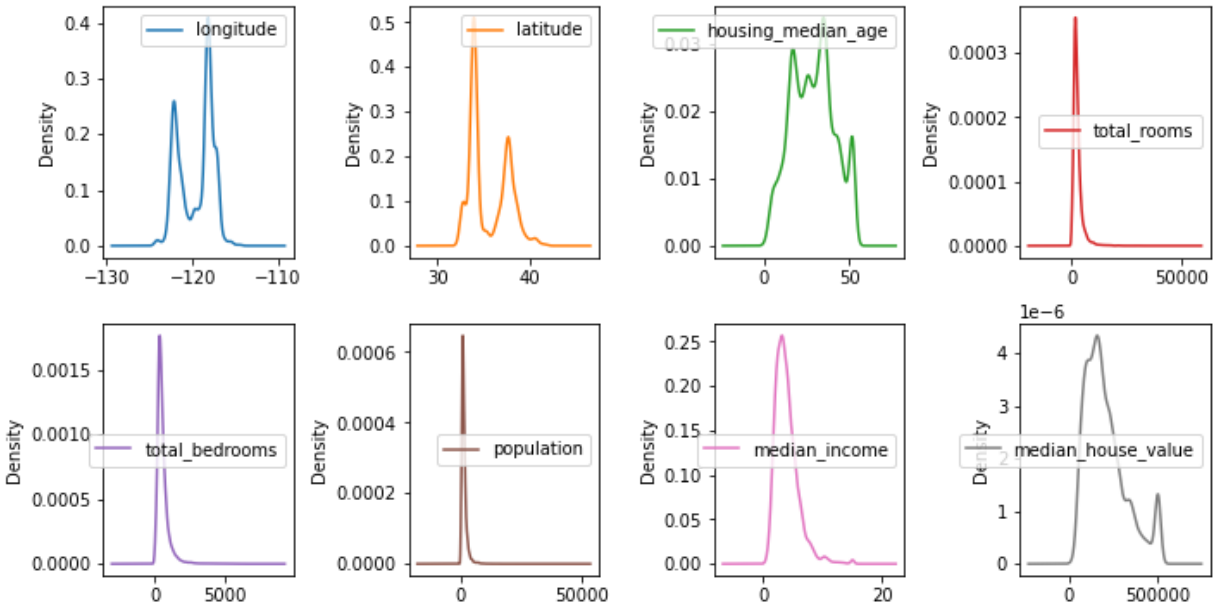
In [21]:
```
1 data['housing_median_age'].value_counts()
```

Out[21]:
```
52.0    1261
36.0     856
35.0     820
16.0     764
17.0     686
34.0     683
33.0     609
26.0     604
18.0     562
25.0     558
32.0     556
37.0     532
15.0     508
19.0     497
27.0     476
24.0     471
30.0     468
28.0     465
20.0     455
29.0     455
31.0     453
21.0     441
23.0     437
14.0     412
22.0     391
38.0     389
42.0     366
39.0     365
44.0     356
43.0     349
40.0     301
13.0     297
41.0     295
45.0     291
10.0     257
11.0     246
46.0     245
5.0      239
12.0     235
9.0      204
8.0      204
47.0     195
4.0      188
48.0     176
7.0      173
6.0      153
50.0     134
49.0     133
3.0       61
2.0       58
51.0      48
1.0        4
Name: housing_median_age, dtype: int64
```

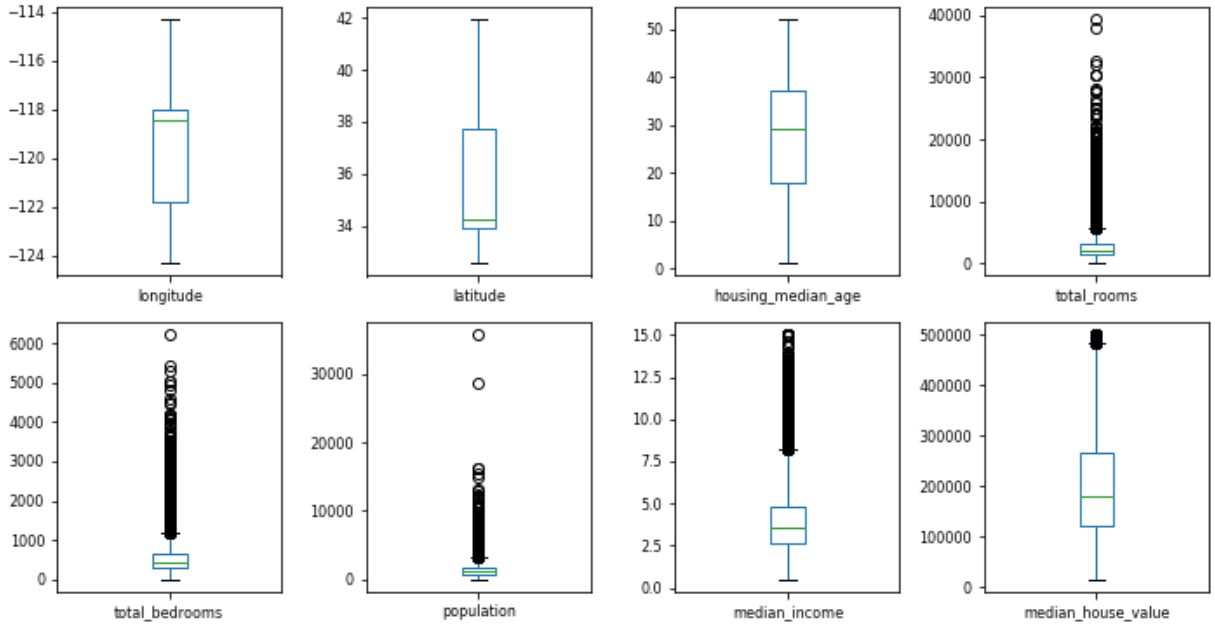
In [22]:
```
data.hist(bins=50, figsize=(20,15))
plt.show()
```



In [23]:
```
data.plot(kind='density',subplots=True,layout=(4,4),sharex=False,figsize=(10
plt.tight_layout()
```

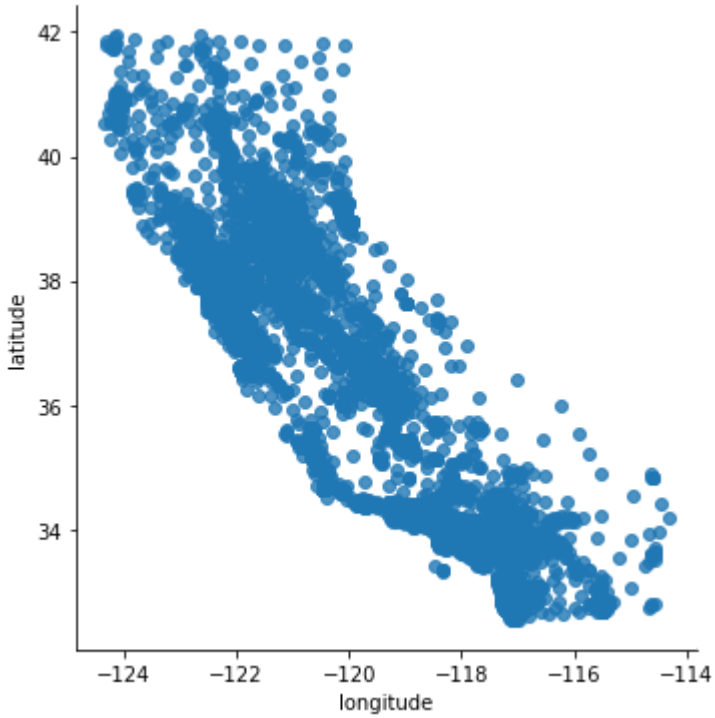


In [24]:
```
data.plot(kind='box', subplots=True, layout=(4,4), sharex=False,
fontsize=8,figsize=(10,10))
plt.tight_layout()
```



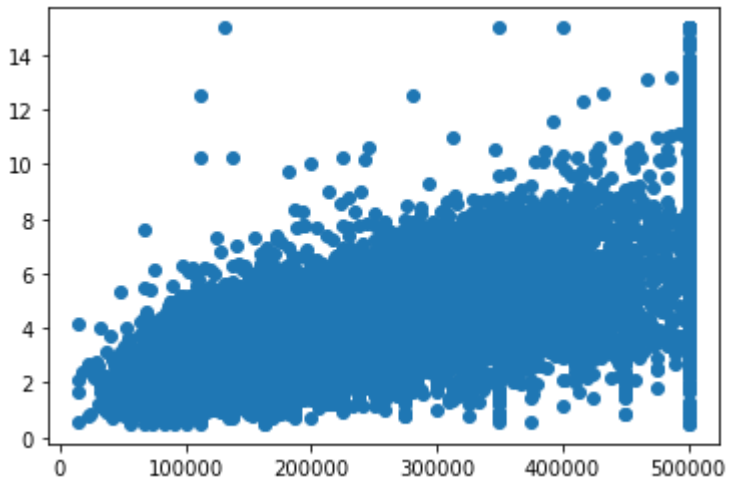
In [29]:
```
sns.lmplot('longitude', 'latitude', data=data, fit_reg=False, height=5)
plt.show()
```

```
C:\Users\Qebaa\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



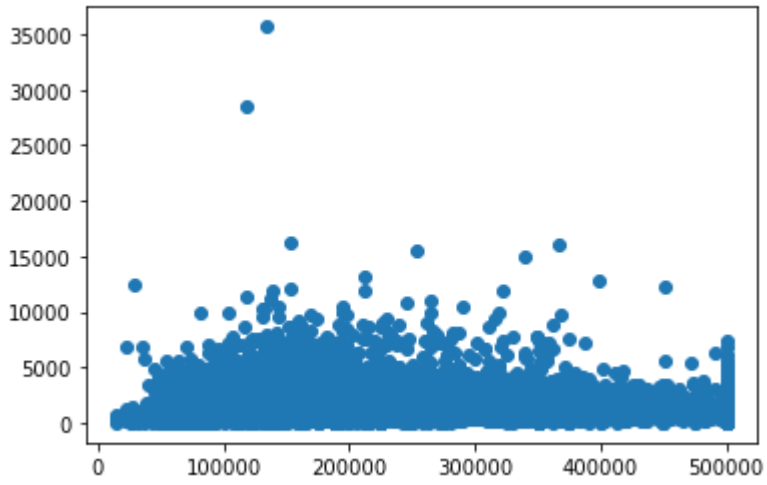
In [38]:
```
plt.scatter(data['median_house_value'], data['median_income'])
```

Out[38]:
```
<matplotlib.collections.PathCollection at 0x19e397462e0>
```



In [41]:
```
1 plt.scatter(data['median_house_value'], data['population'])
```

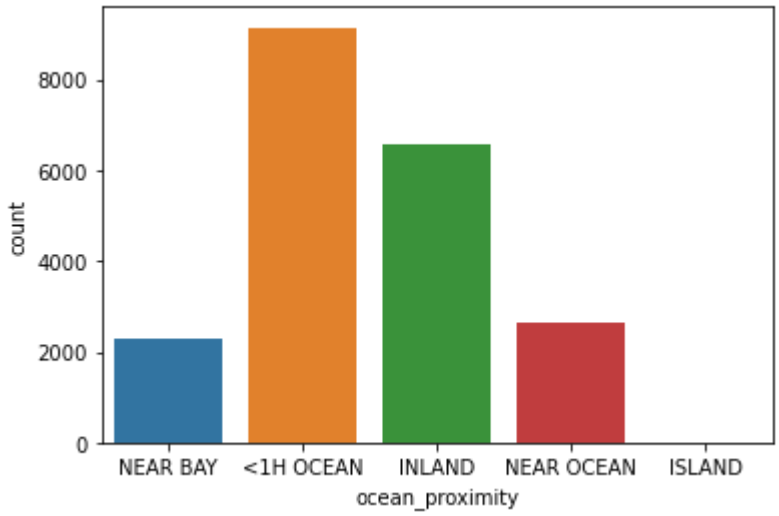
Out[41]:
```
<matplotlib.collections.PathCollection at 0x19e354a03d0>
```



In [25]:
```
1 sns.countplot(data['ocean_proximity'])
```

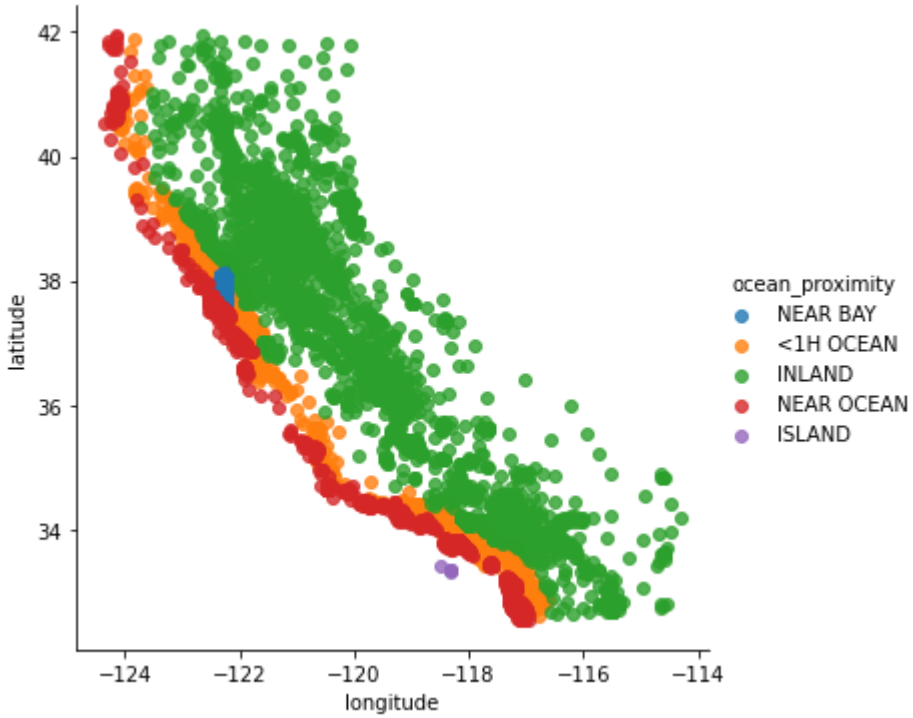
```
C:\Users\Qebaa\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWar
ning: Pass the following variable as a keyword arg: x. From version 0.12, the o
nly valid positional argument will be `data`, and passing other arguments witho
ut an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[25]:
```
<AxesSubplot:xlabel='ocean_proximity', ylabel='count'>
```

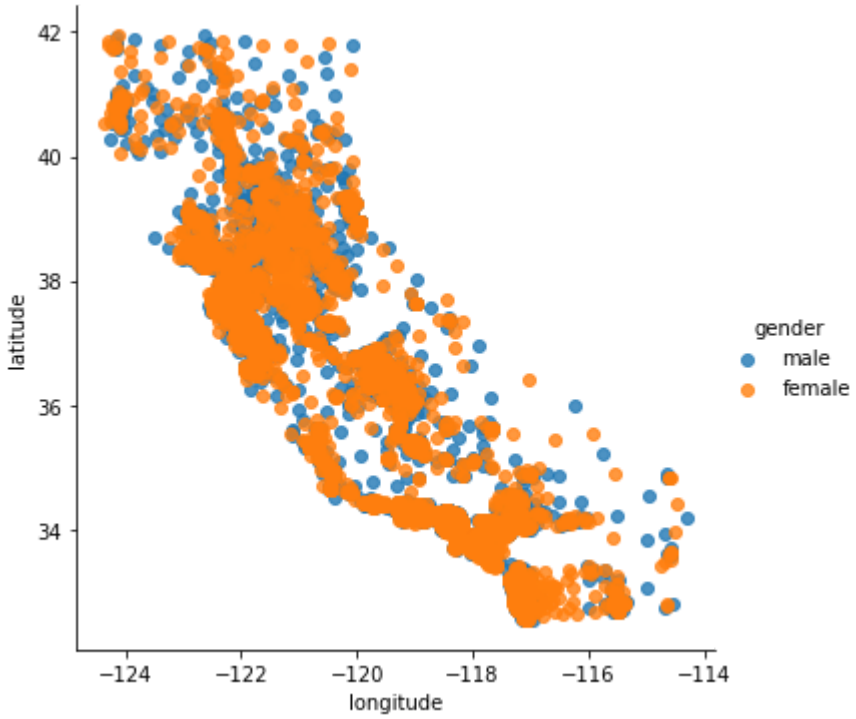


In [4]:
```
sns.lmplot('longitude', 'latitude', data=data, hue='ocean_proximity', fit_re
plt.show()
```

```
C:\Users\Qebaa\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWar
ning: Pass the following variables as keyword args: x, y. From version 0.12, th
e only valid positional argument will be `data`, and passing other arguments wi
thout an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

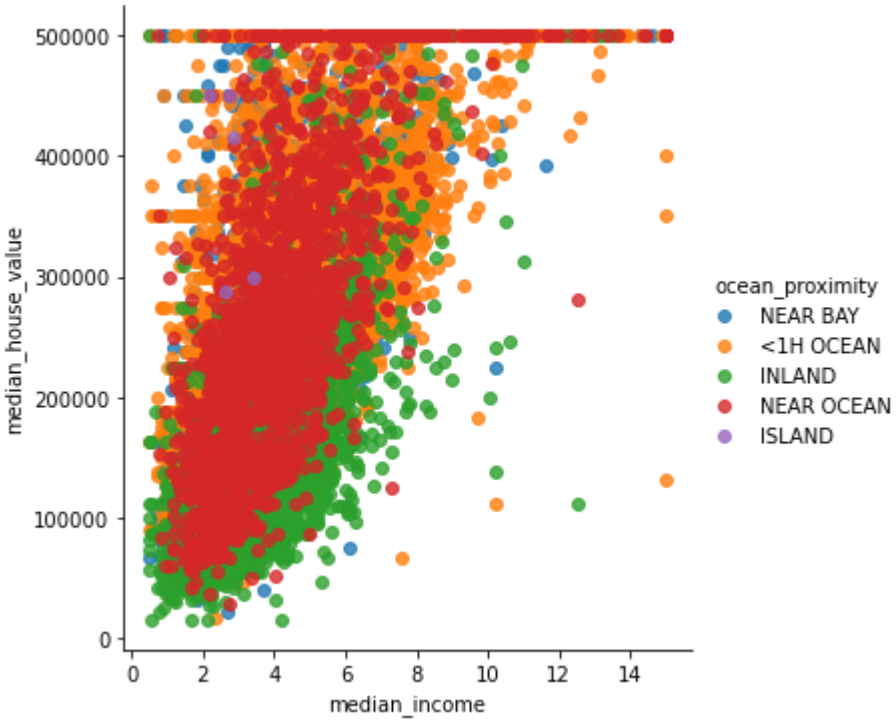


In [9]:
```
sns.lmplot('longitude', 'latitude', data=data, hue='gender', fit_reg=False,
plt.show()
```



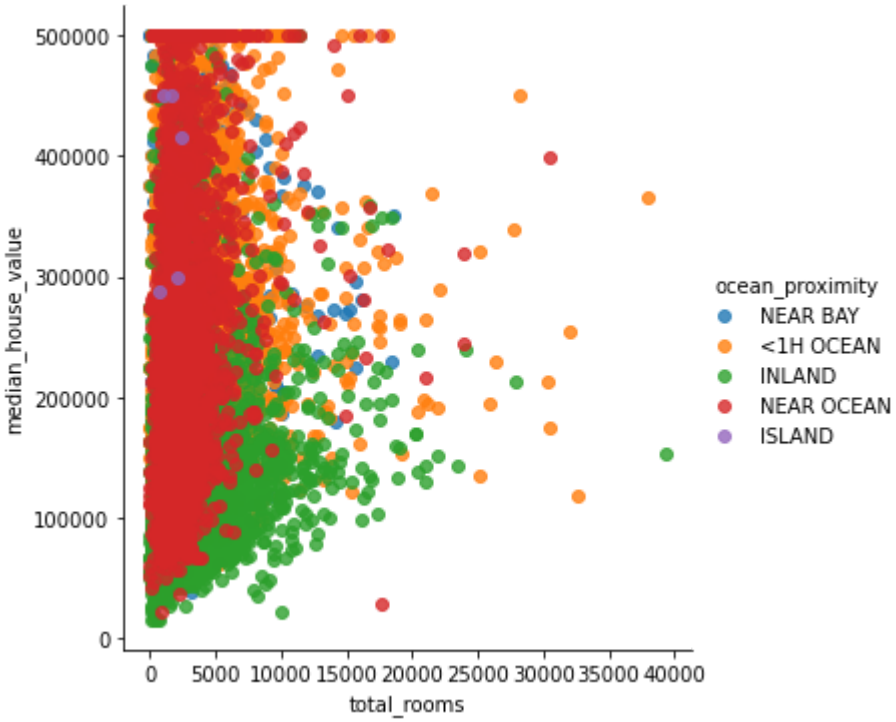
In [27]:
```
sns.lmplot('median_income','median_house_value' , data=data, hue='ocean_prox
plt.show()
```

```
C:\Users\Qebaa\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWar
ning: Pass the following variables as keyword args: x, y. From version 0.12, th
e only valid positional argument will be `data`, and passing other arguments wi
thout an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



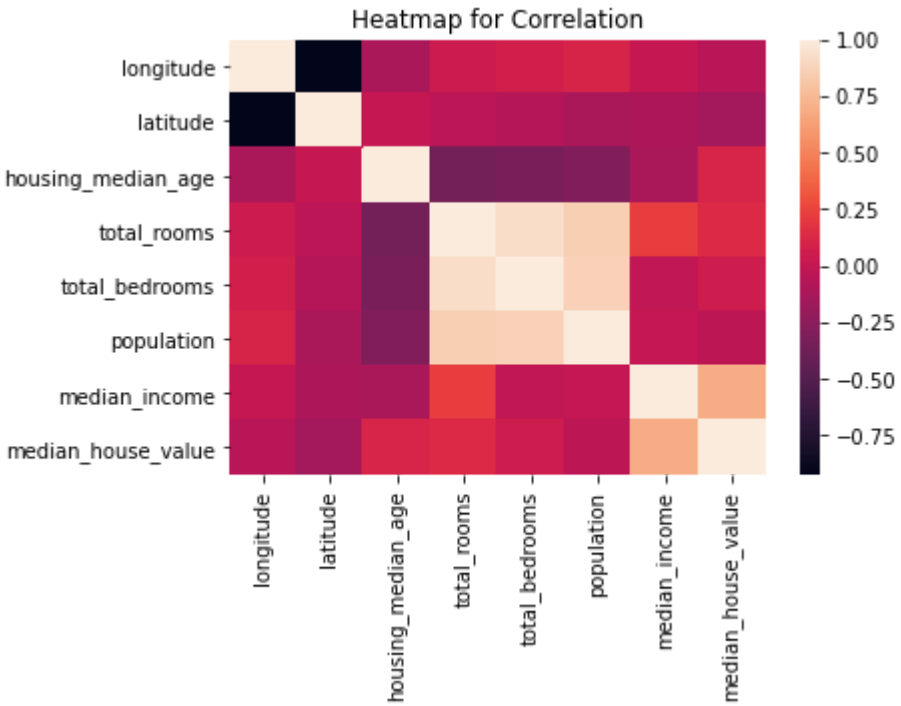
In [28]:
```
sns.lmplot( 'total_rooms','median_house_value', data=data, hue='ocean_proxim
plt.show()
```

```
C:\Users\Qebaa\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWar
ning: Pass the following variables as keyword args: x, y. From version 0.12, th
e only valid positional argument will be `data`, and passing other arguments wi
thout an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



In [30]:
```
1 sns.heatmap(data.corr()).set_title('Heatmap for Correlation')
```

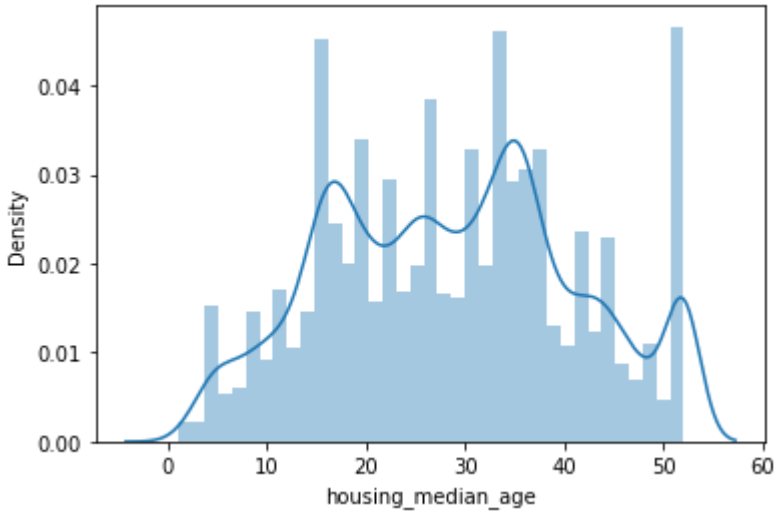
Out[30]:
```
Text(0.5, 1.0, 'Heatmap for Correlation')
```



In [31]:
```
1 sns.distplot(data['housing_median_age'])
```

```
C:\Users\Qebaa\anaconda3\lib\site-packages\seaborn\distributions.py:2551: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

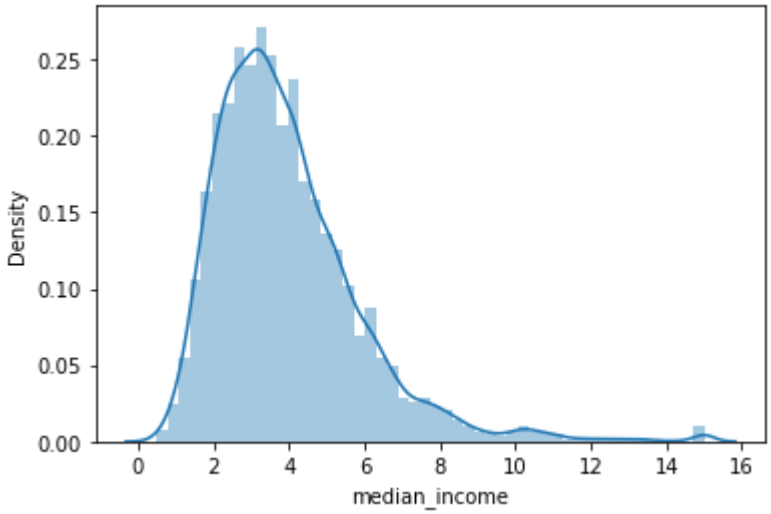
Out[31]: <AxesSubplot:xlabel='housing_median_age', ylabel='Density'>



In [33]:
```
1 sns.distplot(data['median_income'])
```

```
C:\Users\Qebaa\anaconda3\lib\site-packages\seaborn\distributions.py:2551: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

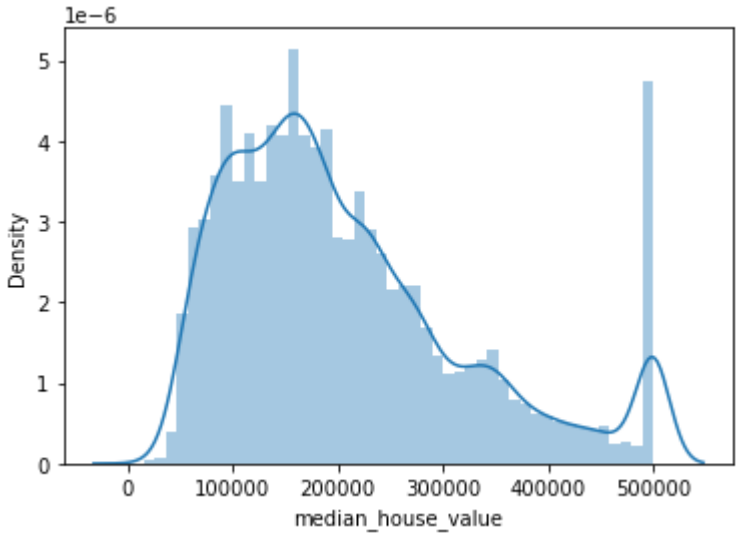
Out[33]:
```
<AxesSubplot:xlabel='median_income', ylabel='Density'>
```



In [34]:
```
1 sns.distplot(data['median_house_value'])
```

```
C:\Users\Qebaa\anaconda3\lib\site-packages\seaborn\distributions.py:2551: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[34]: <AxesSubplot:xlabel='median_house_value', ylabel='Density'>



In [43]:
```
1 data.sort_values(['median_income','median_house_value'],ascending=True)
```

Out[43]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	household
19800	-123.32	40.43	15.0	661	NaN	131.0	N
73	-122.29	37.81	46.0	12	4.0	18.0	
3258	-122.89	39.42	16.0	411	114.0	26.0	
19523	-121.01	37.65	52.0	178	53.0	152.0	
5213	-118.28	33.93	52.0	117	33.0	74.0	
...	...	...	...	...	...	...	
14545	-117.26	32.95	15.0	1036	149.0	395.0	1
14546	-117.26	32.95	15.0	1882	233.0	704.0	2
14806	-117.18	32.68	29.0	1539	344.0	556.0	2
14807	-117.18	32.69	52.0	1837	313.0	668.0	3
14810	-117.17	32.69	40.0	2236	331.0	767.0	3

20640 rows × 11 columns

In [44]:
```
1 data.sort_values(['housing_median_age','median_house_value'],ascending=True)
```

Out[44]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	household
12286	-116.95	33.86	1.0	6	2.0	8.0	
3130	-117.95	35.08	1.0	83	15.0	32.0	
19536	-120.93	37.65	1.0	2254	328.0	402.0	1
18972	-122.00	38.23	1.0	2062	343.0	872.0	2
2774	-115.80	33.26	2.0	96	18.0	30.0	
...	...	...	...	...	...	...	
20422	-118.90	34.14	NaN	1503	NaN	576.0	N
20426	-118.69	34.18	NaN	1177	NaN	415.0	N
20427	-118.80	34.19	NaN	15572	NaN	5495.0	N
20436	-118.69	34.21	NaN	3663	NaN	1179.0	N
20443	-118.85	34.27	NaN	187	NaN	130.0	N

20640 rows × 11 columns

In [45]:
```
1 data.sort_values(['housing_median_age','population'],ascending=True)
```

Out[45]:

longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	me
-116.95	33.86	1.0	6	2.0	8.0	2	
-117.95	35.08	1.0	83	15.0	32.0	15	
-120.93	37.65	1.0	2254	328.0	402.0	112	
-122.00	38.23	1.0	2062	343.0	872.0	268	
-121.96	37.74	2.0	200	20.0	25.0	9	
...	...	...	...	...	...	...	
-118.90	34.26	NaN	25187	NaN	11956.0	NaN	
-122.25	37.85	NaN	1627	280.0	NaN	259	
-122.25	37.85	NaN	919	213.0	NaN	193	
-122.25	37.84	NaN	2535	NaN	NaN	514	
-122.25	37.84	NaN	3104	NaN	NaN	NaN	

rows × 11 columns

In [46]:
```
1 data.sort_values(['median_house_value','ocean_proximity'],ascending=True)
```

Out[46]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	househol
2521	-122.74	39.71	16.0	255	73.0	85.0	
2799	-117.02	36.40	19.0	619	239.0	490.0	1
9188	-117.86	34.24	52.0	803	NaN	628.0	
19802	-123.17	40.31	36.0	98	NaN	18.0	Na
5887	-118.33	34.15	39.0	493	168.0	259.0	1
...	...	...	...	...	...	...	
20233	-119.29	34.24	27.0	4742	NaN	1682.0	Na
20272	-119.23	34.19	16.0	5297	NaN	1489.0	Na
20273	-119.23	34.17	18.0	6171	NaN	2164.0	Na
20322	-119.14	34.23	8.0	243	NaN	102.0	Na
20380	-118.83	34.14	16.0	1316	NaN	450.0	Na

20640 rows × 11 columns

In [49]:
```
1 new_val = pd.get_dummies(data.ocean_proximity)
```

In [50]:
```
1 new_val.head(5)
```

Out[50]:

	<1H OCEAN	INLAND	ISLAND	NEAR BAY	NEAR OCEAN
0	0	0	0	1	0
1	0	0	0	1	0
2	0	0	0	1	0
3	0	0	0	1	0
4	0	0	0	1	0

In [51]:
```
1 data[new_val.columns] = new_val
```

In [52]:
```
1 data.describe()
2
```

Out[52]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	populati
count	20640.000000	20640.000000	20382.000000	20640.000000	15758.000000	20596.0000
mean	-119.569704	35.631861	28.676283	2635.763081	539.920104	1424.9287
std	2.003532	2.135952	12.589284	2181.615252	419.834171	1132.2377
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.0000
25%	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.0000
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.0000
75%	-118.010000	37.710000	37.000000	3148.000000	652.000000	1725.0000
max	-114.310000	41.950000	52.000000	39320.000000	6210.000000	35682.0000

In [53]:
```
1 data.columns
```

Out[53]:
```
Index(['longitude', 'latitude', 'housing_median_age', 'total_rooms',
       'total_bedrooms', 'population', 'households', 'median_income',
       'median_house_value', 'ocean_proximity', 'gender', '<1H OCEAN',
       'INLAND', 'ISLAND', 'NEAR BAY', 'NEAR OCEAN'],
      dtype='object')
```

In [57]:
```
1 data=data[['longitude', 'latitude', 'housing_median_age', 'total_rooms',
2        'total_bedrooms', 'population', 'households', 'median_income',
3         'gender', '<1H OCEAN',
4        'INLAND', 'ISLAND', 'NEAR BAY', 'NEAR OCEAN','median_house_value']]
```

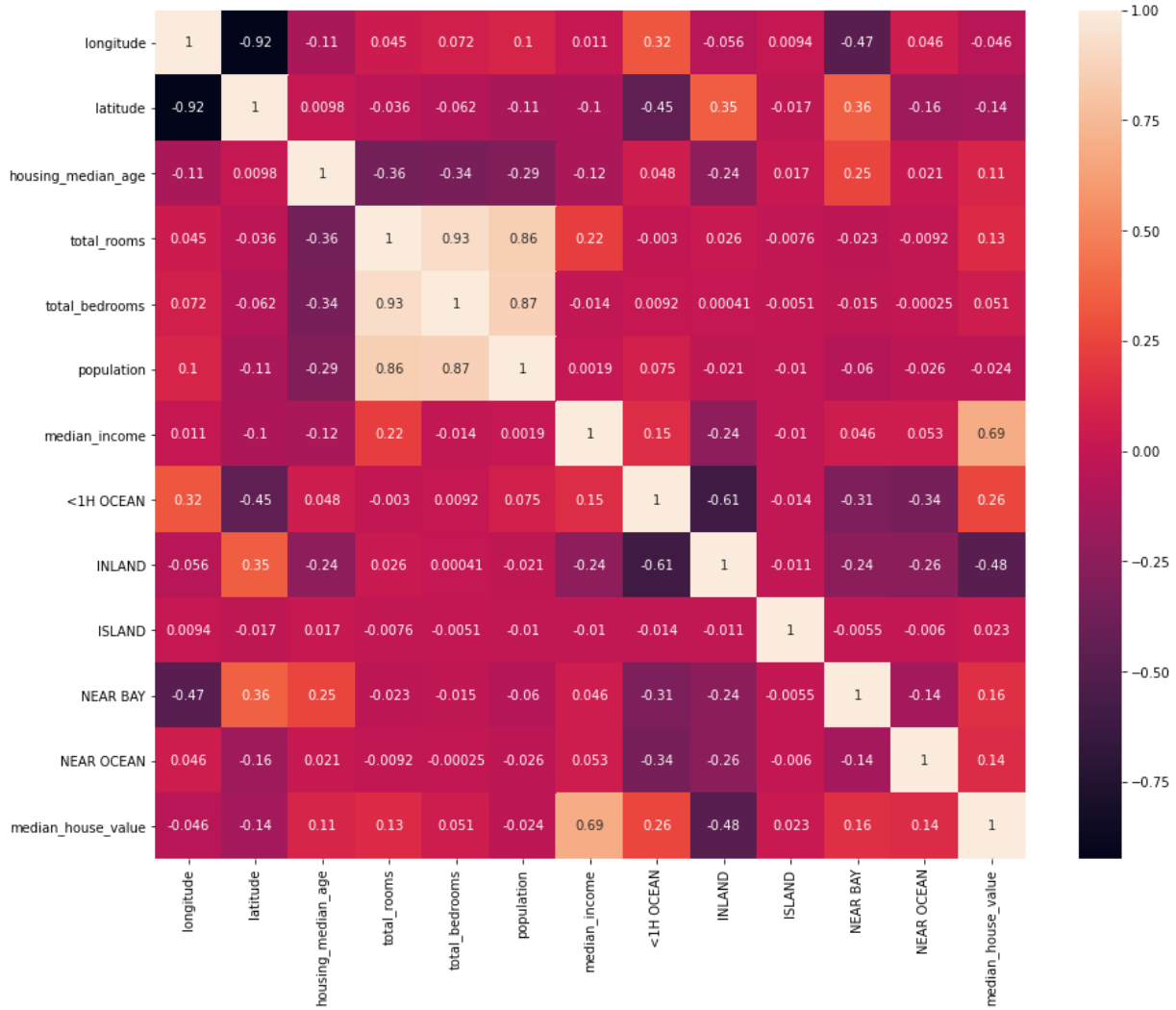
In [58]:
```
1 data.corr()
```

Out[58]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	popu
longitude	1.000000	-0.924664	-0.107447	0.044568	0.071610	0.1
latitude	-0.924664	1.000000	0.009773	-0.036100	-0.062301	-0.1
housing_median_age	-0.107447	0.009773	1.000000	-0.360441	-0.336824	-0.2
total_rooms	0.044568	-0.036100	-0.360441	1.000000	0.926709	0.8
total_bedrooms	0.071610	-0.062301	-0.336824	0.926709	1.000000	0.8
population	0.100390	-0.109222	-0.294245	0.856914	0.871130	1.0
median_income	0.011440	-0.103496	-0.117771	0.220391	-0.013803	0.0
<1H OCEAN	0.321121	-0.446969	0.048219	-0.003031	0.009239	0.0
INLAND	-0.055575	0.351166	-0.238743	0.025624	0.000414	-0.0
ISLAND	0.009446	-0.016572	0.017076	-0.007572	-0.005072	-0.0
NEAR BAY	-0.474489	0.358771	0.252832	-0.023022	-0.015300	-0.0
NEAR OCEAN	0.045509	-0.160818	0.021194	-0.009175	-0.000252	-0.0
median_house_value	-0.045967	-0.144160	0.107378	0.134153	0.050963	-0.0

In [61]:
```
plt.figure(figsize=(15,12))
sns.heatmap(data.corr(), annot=True)
```

Out[61]: <AxesSubplot:>



In [62]:
```
data['gender'].replace("male",0,inplace=True)
data['gender'].replace("female",1,inplace=True)
data.head()

```

```
C:\Users\Qebaa\anaconda3\lib\site-packages\pandas\core\series.py:4563: SettingW
ithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/sta
ble/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pyd
ata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-c
opy)
  return super().replace(
```

Out[62]:

ng_median_age	total_rooms	total_bedrooms	population	households	median_income	gender	< OCE
41.0	880	129.0	322.0	126	8.3252	0.0	
21.0	7099	1106.0	2401.0	1138	8.3014	1.0	
52.0	1467	190.0	496.0	177	7.2574	0.0	
52.0	1274	235.0	558.0	219	5.6431	1.0	
NaN	1627	280.0	NaN	259	3.8462	0.0	

In [63]:
```
data.isna().sum()
```

Out[63]:
```
longitude               0
latitude                0
housing_median_age    258
total_rooms             0
total_bedrooms       4882
population             44
households           1305
median_income        2767
gender               4020
<1H OCEAN               0
INLAND                  0
ISLAND                  0
NEAR BAY                0
NEAR OCEAN              0
median_house_value      0
dtype: int64
```

In [70]:
```
data = data.fillna(data.mean())
```

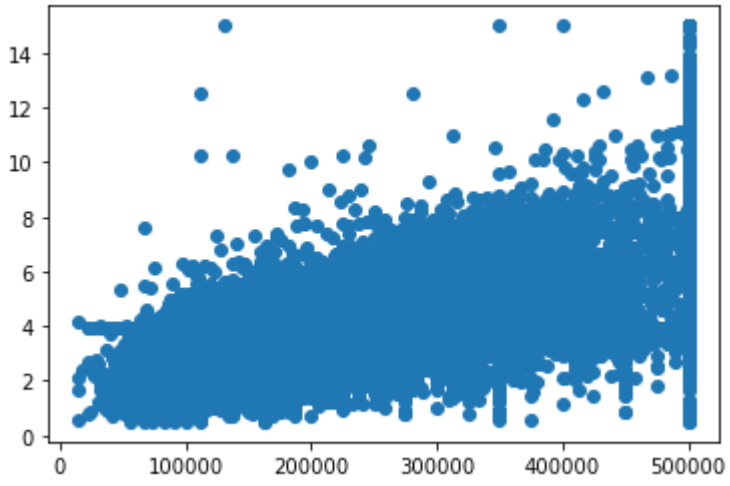
In [73]:
```
1 data.head(20)
```

Out[73]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	household
0	-122.23	37.88	41.000000	880	129.000000	322.000000	
1	-122.22	37.86	21.000000	7099	1106.000000	2401.000000	1
2	-122.24	37.85	52.000000	1467	190.000000	496.000000	
3	-122.25	37.85	52.000000	1274	235.000000	558.000000	
4	-122.25	37.85	28.676283	1627	280.000000	1424.928724	
5	-122.25	37.85	28.676283	919	213.000000	1424.928724	
6	-122.25	37.84	28.676283	2535	539.920104	1424.928724	
7	-122.25	37.84	28.676283	3104	539.920104	1424.928724	N
8	-122.26	37.84	42.000000	2555	539.920104	1424.928724	N
9	-122.25	37.84	52.000000	3549	539.920104	1424.928724	N
10	-122.26	37.85	52.000000	2202	539.920104	1424.928724	N
11	-122.26	37.85	52.000000	3503	539.920104	1424.928724	N
12	-122.26	37.85	52.000000	2491	539.920104	1424.928724	N
13	-122.26	37.84	52.000000	696	191.000000	345.000000	N
14	-122.26	37.85	52.000000	2643	626.000000	1212.000000	N
15	-122.26	37.85	50.000000	1120	283.000000	697.000000	N
16	-122.27	37.85	52.000000	1966	347.000000	793.000000	N
17	-122.27	37.85	52.000000	1228	293.000000	648.000000	
18	-122.26	37.84	50.000000	2239	455.000000	990.000000	
19	-122.27	37.84	52.000000	1503	298.000000	690.000000	

In [72]:
```
1 plt.scatter(data['median_house_value'], data['median_income'])
```

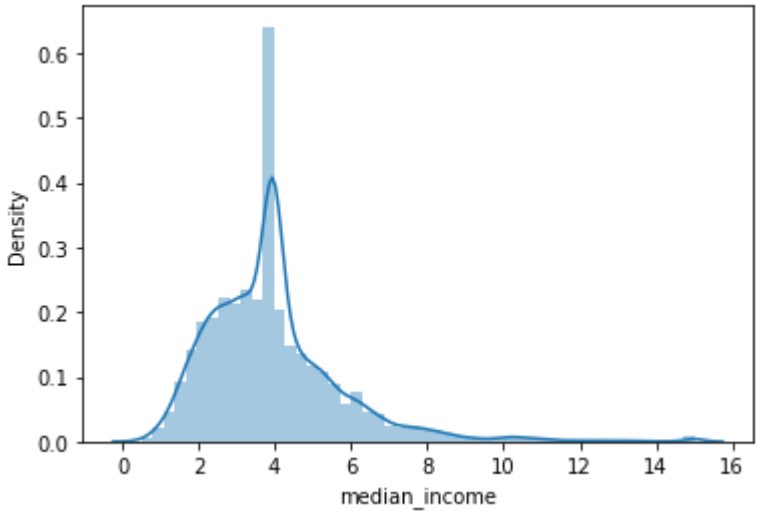
Out[72]:
```
<matplotlib.collections.PathCollection at 0x19e3577f8e0>
```



In [74]:
```
1 sns.distplot(data['median_income'])
```

```
C:\Users\Qebaa\anaconda3\lib\site-packages\seaborn\distributions.py:2551: Futur
eWarning: `distplot` is a deprecated function and will be removed in a future v
ersion. Please adapt your code to use either `displot` (a figure-level function
with similar flexibility) or `histplot` (an axes-level function for histogram
s).
  warnings.warn(msg, FutureWarning)
```

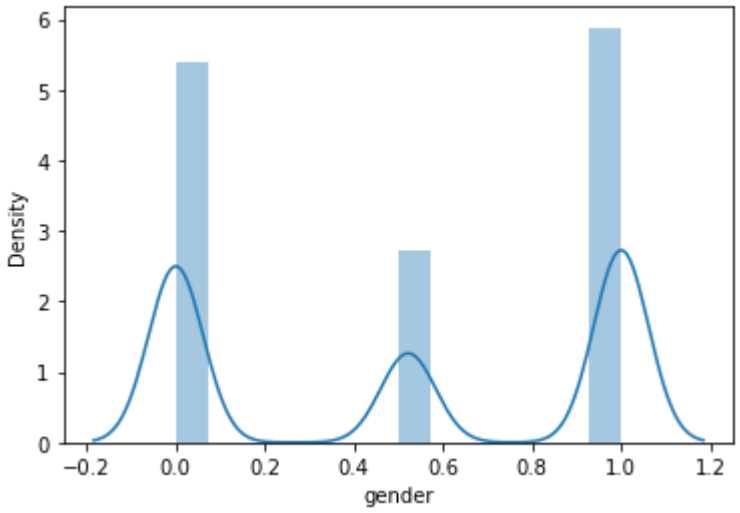
Out[74]: <AxesSubplot:xlabel='median_income', ylabel='Density'>



In [76]:
```
1 sns.distplot(data['gender'])
```

```
C:\Users\Qebaa\anaconda3\lib\site-packages\seaborn\distributions.py:2551: Futur
eWarning: `distplot` is a deprecated function and will be removed in a future v
ersion. Please adapt your code to use either `displot` (a figure-level function
with similar flexibility) or `histplot` (an axes-level function for histogram
s).
  warnings.warn(msg, FutureWarning)
```

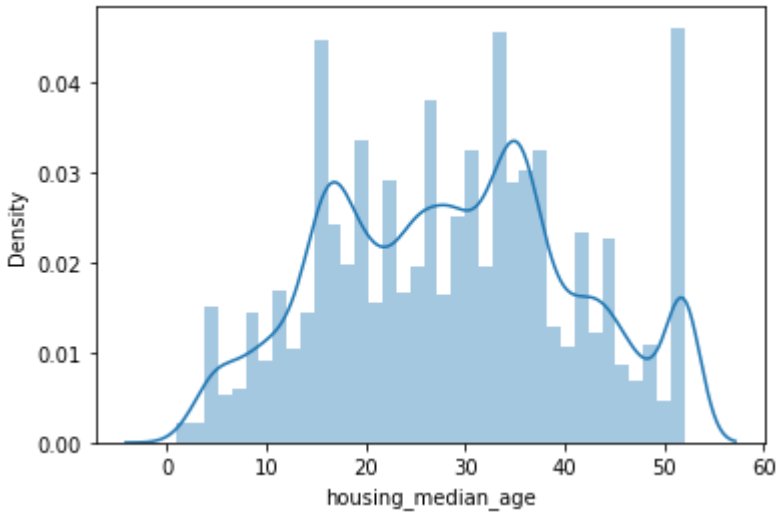
Out[76]: <AxesSubplot:xlabel='gender', ylabel='Density'>



In [77]:
```
1 sns.distplot(data['housing_median_age'])
```

```
C:\Users\Qebaa\anaconda3\lib\site-packages\seaborn\distributions.py:2551: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[77]: <AxesSubplot:xlabel='housing_median_age', ylabel='Density'>



In [ ]:
```
1
```