In [4]:
```
import pandas as pd
import numpy as np
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
#np.random.seed(101)
```

In [5]:
```
##reading the data
data= pd.read_csv("F:\machinfy\mohamed\working\working machinfy\housing10.cs
```

In [323]:
```
data.head()
```

Out[323]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households
0	-122.23	37.88	41.0	880	129.0	322.0	126
1	-122.22	37.86	21.0	7099	1106.0	2401.0	1138
2	-122.24	37.85	52.0	1467	190.0	496.0	177
3	-122.25	37.85	52.0	1274	235.0	558.0	219
4	-122.25	37.85	NaN	1627	280.0	NaN	259

In [324]:
```
data.tail()
```

Out[324]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	househol
20635	-121.09	39.48	25.0	1665	374.0	845.0	3
20636	-121.21	39.49	18.0	697	150.0	356.0	1
20637	-121.22	39.43	17.0	2254	485.0	1007.0	4
20638	-121.32	39.43	18.0	1860	409.0	741.0	3
20639	-121.24	39.37	16.0	2785	616.0	1387.0	5

In [325]:
```
1 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 [326]:
```
1 data.drop_duplicates()
```

Out[326]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	househol
0	-122.23	37.88	41.0	880	129.0	322.0	1
1	-122.22	37.86	21.0	7099	1106.0	2401.0	11
2	-122.24	37.85	52.0	1467	190.0	496.0	1
3	-122.25	37.85	52.0	1274	235.0	558.0	2
4	-122.25	37.85	NaN	1627	280.0	NaN	2
...	...	...	...	...	...	...	
20635	-121.09	39.48	25.0	1665	374.0	845.0	3
20636	-121.21	39.49	18.0	697	150.0	356.0	1
20637	-121.22	39.43	17.0	2254	485.0	1007.0	4
20638	-121.32	39.43	18.0	1860	409.0	741.0	3
20639	-121.24	39.37	16.0	2785	616.0	1387.0	5

20640 rows × 11 columns

In [327]:
```
1 data.describe(include='all')
```

Out[327]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	populat
count	20640.000000	20640.000000	20382.000000	20640.000000	15758.000000	20596.000
unique	NaN	NaN	NaN	NaN	NaN	N
top	NaN	NaN	NaN	NaN	NaN	N
freq	NaN	NaN	NaN	NaN	NaN	N
mean	-119.569704	35.631861	28.676283	2635.763081	539.920104	1424.928
std	2.003532	2.135952	12.589284	2181.615252	419.834171	1132.237
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000
25%	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.000
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.000
75%	-118.010000	37.710000	37.000000	3148.000000	652.000000	1725.000
max	-114.310000	41.950000	52.000000	39320.000000	6210.000000	35682.000

In [328]:
```
1 data.describe(include='all').T
```

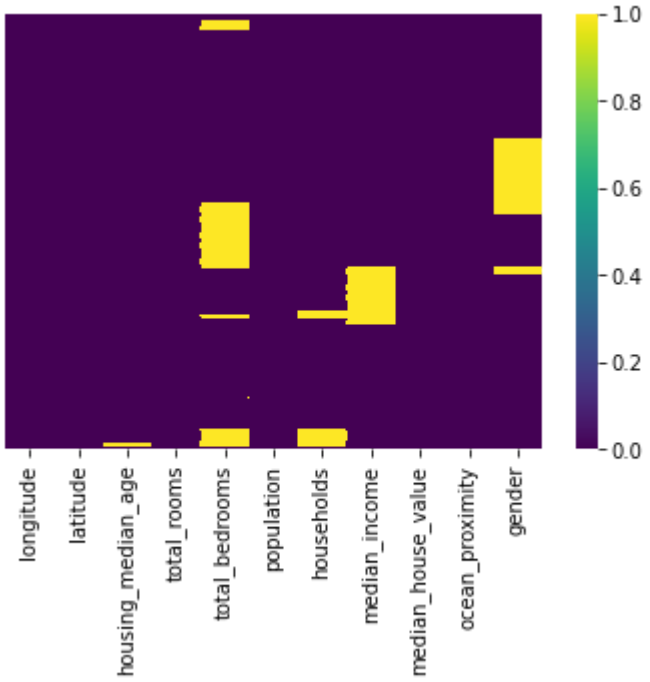
Out[328]:

	count	unique	top	freq	mean	std	min	25%	50%	
longitude	20640	NaN	NaN	NaN	-119.57	2.00353	-124.35	-121.8	-118.49	-1
latitude	20640	NaN	NaN	NaN	35.6319	2.13595	32.54	33.93	34.26	
housing_median_age	20382	NaN	NaN	NaN	28.6763	12.5893	1	18	29	
total_rooms	20640	NaN	NaN	NaN	2635.76	2181.62	2	1447.75	2127	
total_bedrooms	15758	NaN	NaN	NaN	539.92	419.834	1	296	435	
population	20596	NaN	NaN	NaN	1424.93	1132.24	3	787	1166	
households	19335	1703	no	3080	NaN	NaN	NaN	NaN	NaN	
median_income	17873	NaN	NaN	NaN	3.94068	1.94374	0.4999	2.5986	3.5871	4.
median_house_value	20640	NaN	NaN	NaN	206856	115396	14999	119600	179700	26
ocean_proximity	20640	5	<1H OCEAN	9136	NaN	NaN	NaN	NaN	NaN	
gender	16620	2	female	8673	NaN	NaN	NaN	NaN	NaN	

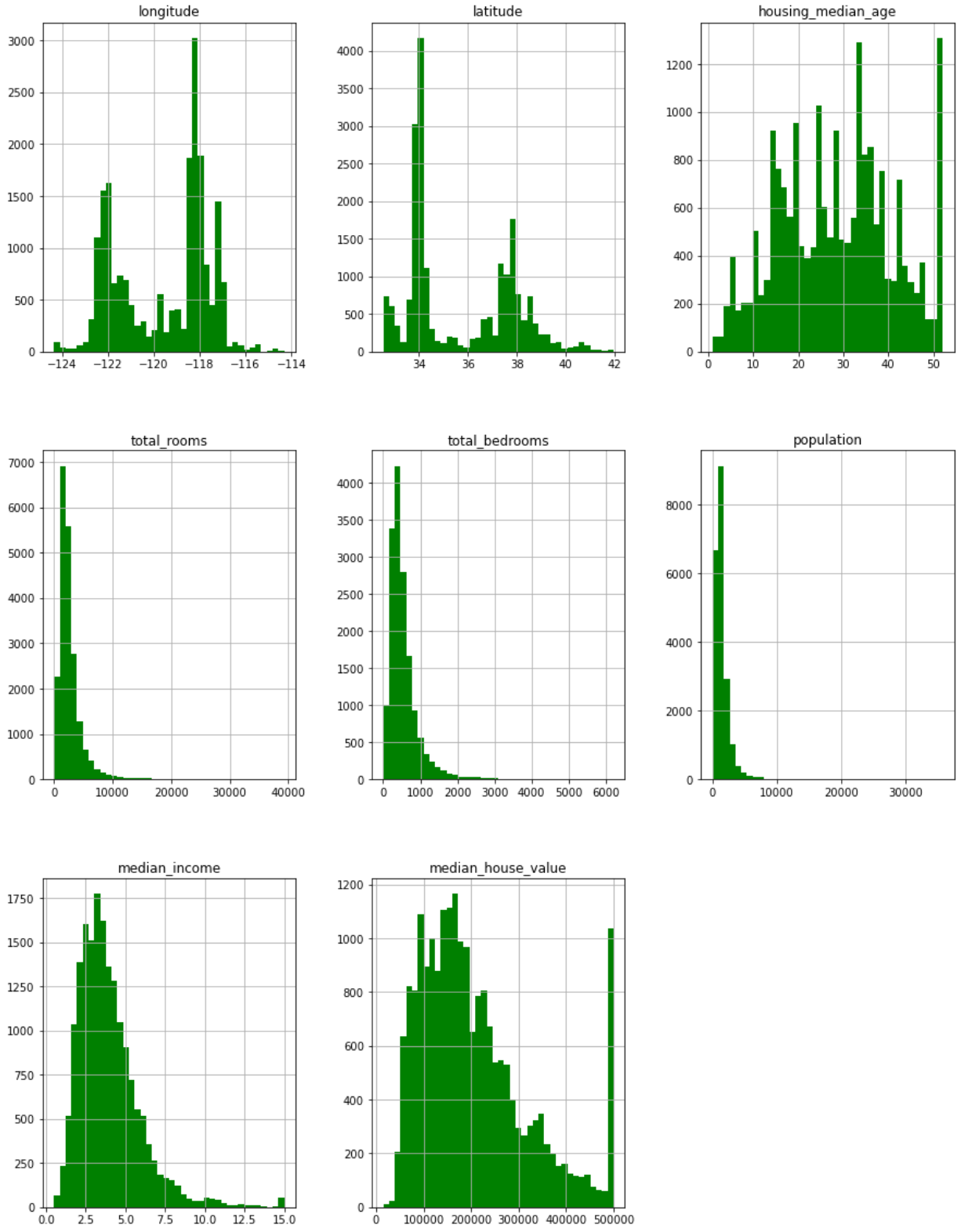
In [329]:
```
1 data.isnull().sum()
```

Out[329]:
```
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
```

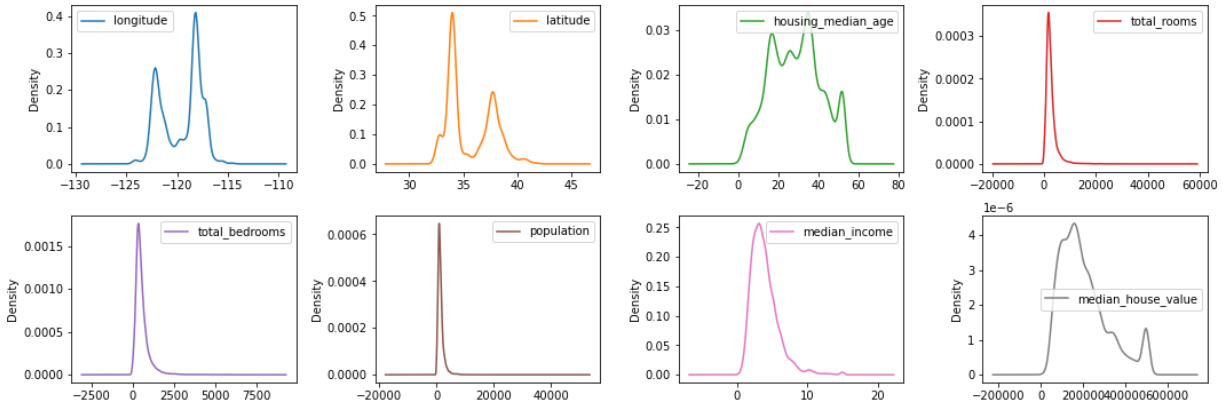
In [330]:
```
1 sns.heatmap(data.isnull(), cmap='viridis', cbar=True ,yticklabels =False)
2 plt.show()
```



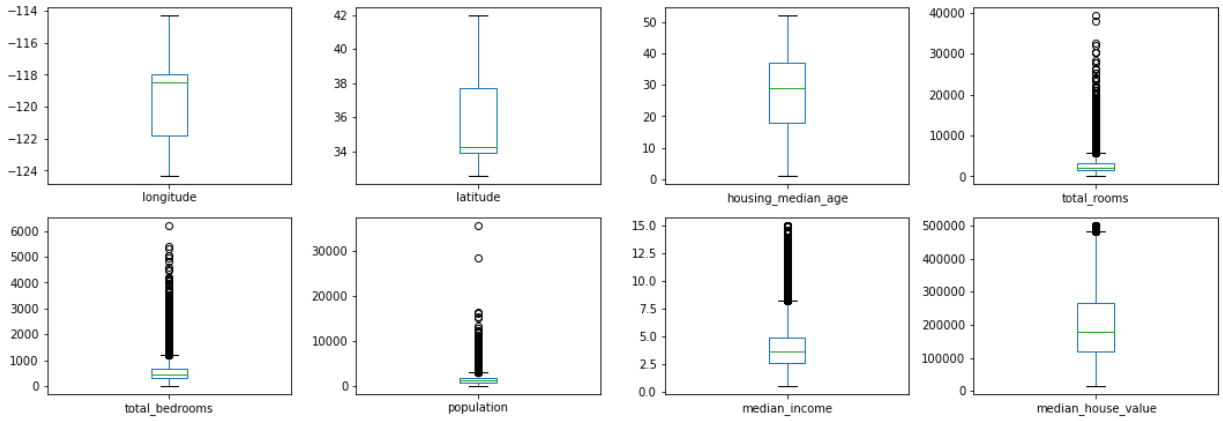
In [331]:
```
data.hist(bins=40, figsize=(15,20), color='green')
plt.show()
```



In [332]:
```
data.plot(kind='density',subplots=True,layout=(4,4),sharex=False,figsize=(15
plt.tight_layout()
```



In [333]:
```
data.plot(kind='box',subplots=True,layout=(4,4),sharex=False,figsize=(15,10)
plt.tight_layout()
```



In [334]:
```
data.drop('gender', axis=1 , inplace=True)
```

In [335]:
```
1 data.head()
```

Out[335]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households
0	-122.23	37.88	41.0	880	129.0	322.0	126
1	-122.22	37.86	21.0	7099	1106.0	2401.0	1138
2	-122.24	37.85	52.0	1467	190.0	496.0	177
3	-122.25	37.85	52.0	1274	235.0	558.0	219
4	-122.25	37.85	NaN	1627	280.0	NaN	259

In [336]:
```
1 data.dropna(thresh=9, inplace=True)
```

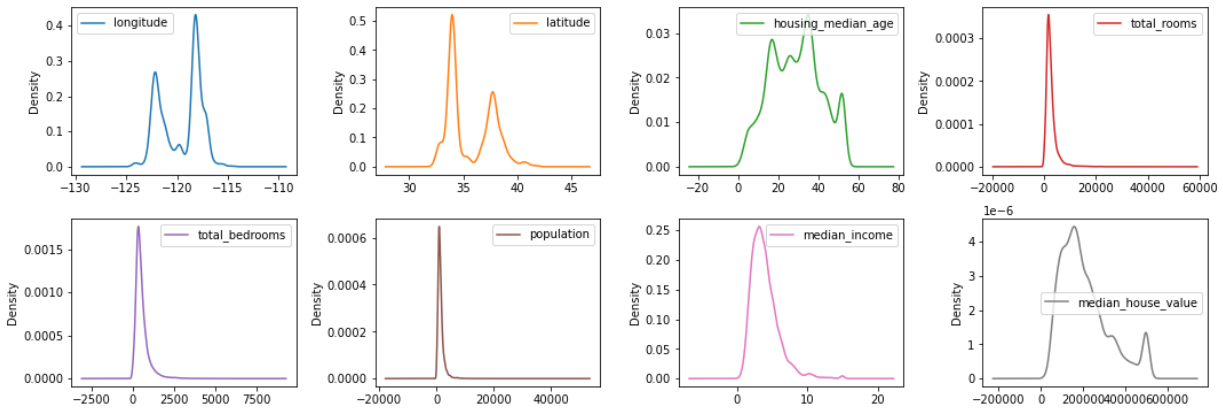
In [337]:
```
1 data
```

Out[337]:

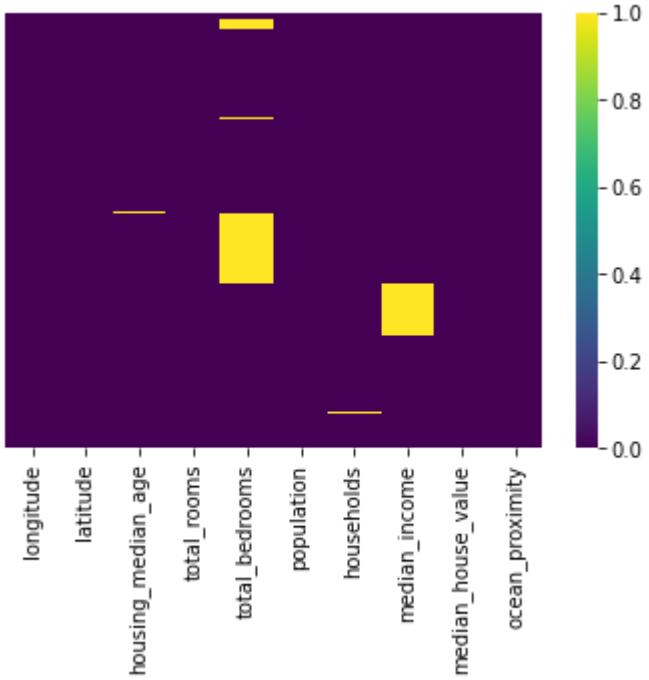
	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	househol
0	-122.23	37.88	41.0	880	129.0	322.0	1
1	-122.22	37.86	21.0	7099	1106.0	2401.0	11
2	-122.24	37.85	52.0	1467	190.0	496.0	1
3	-122.25	37.85	52.0	1274	235.0	558.0	2
13	-122.26	37.84	52.0	696	191.0	345.0	N
...	...	...	...	...	...	...	
20635	-121.09	39.48	25.0	1665	374.0	845.0	3
20636	-121.21	39.49	18.0	697	150.0	356.0	1
20637	-121.22	39.43	17.0	2254	485.0	1007.0	4
20638	-121.32	39.43	18.0	1860	409.0	741.0	3
20639	-121.24	39.37	16.0	2785	616.0	1387.0	5

19284 rows × 10 columns

In [338]:
```
data.plot(kind='density',subplots=True,layout=(4,4),sharex=False,figsize=(15
plt.tight_layout()
```



In [339]:
```
sns.heatmap(data.isnull(), cmap='viridis', cbar=True ,yticklabels =False)
plt.show()
```



In [340]:
```
data.isnull().sum()
```

Out[340]:
```
longitude             0
latitude              0
housing_median_age    36
total_rooms           0
total_bedrooms        3683
population            29
households            53
median_income         2323
median_house_value    0
ocean_proximity       0
dtype: int64
```

In [341]:
```
1 data_mean=data['median_income'].mean()
```

In [342]:
```
1 data_mean
```

Out[342]:
```
3.963598201757007
```

In [343]:
```
1 data_mode=data['median_income'].mode()
```

In [235]:
```
1 data_mode
```

Out[235]:
```
0    15.0001
dtype: float64
```

In [344]:
```
1 list=[3.9635,15.0001,4.1250,2.6250]
2 data['median_income']=data['median_income'].fillna(pd.Series(np.random.choic
3 data.head()
```

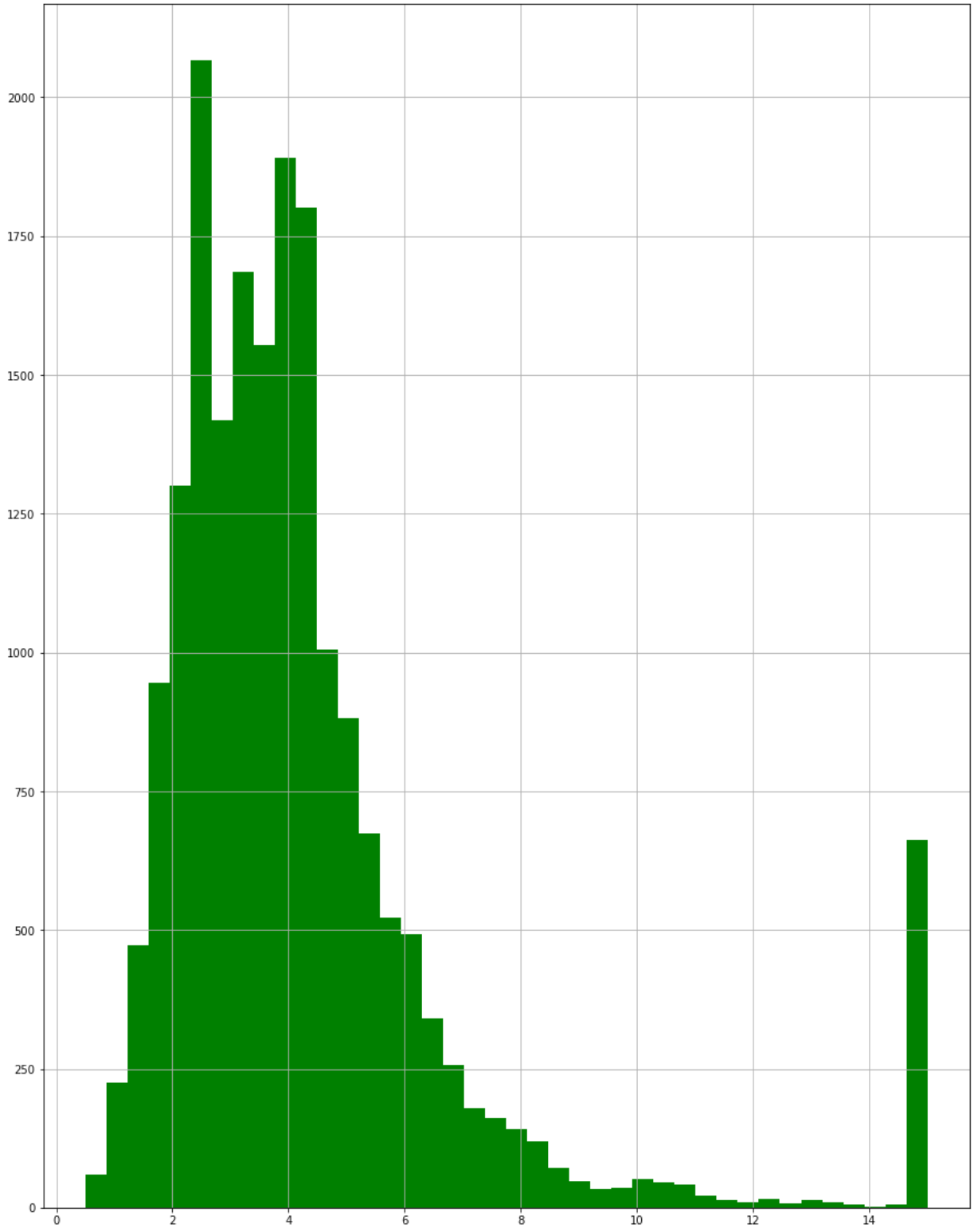
Out[344]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households
0	-122.23	37.88	41.0	880	129.0	322.0	126
1	-122.22	37.86	21.0	7099	1106.0	2401.0	1138
2	-122.24	37.85	52.0	1467	190.0	496.0	177
3	-122.25	37.85	52.0	1274	235.0	558.0	219
13	-122.26	37.84	52.0	696	191.0	345.0	NaN

In [345]:
```
1 data['median_income'].value_counts()
```

Out[345]:
```
15.0001    660
4.1250     605
2.6250     600
3.9635     578
2.8750      40
          ...
4.0452       1
7.7848       1
7.1621       1
6.6689       1
2.5577       1
Name: median_income, Length: 11115, dtype: int64
```

In [346]:
```
data['median_income'].hist(bins=40, figsize=(15,20), color='green')
plt.show()
```



In [347]:
```
1 data.isnull().sum()
```

Out[347]:
```
longitude               0
latitude                0
housing_median_age     36
total_rooms             0
total_bedrooms       3683
population             29
households             53
median_income           0
median_house_value      0
ocean_proximity         0
dtype: int64
```

In [348]:
```
1 data['housing_median_age'].value_counts()
```

Out[348]:
```
52.0    1223
36.0     821
35.0     784
16.0     707
34.0     655
17.0     641
33.0     578
26.0     563
32.0     519
25.0     516
18.0     515
37.0     513
15.0     470
19.0     462
27.0     442
24.0     441
30.0     440
28.0     439
31.0     434
29.0     427
20.0     418
21.0     406
23.0     405
14.0     380
38.0     369
22.0     355
42.0     354
44.0     341
39.0     341
43.0     336
40.0     294
45.0     286
41.0     282
13.0     279
10.0     240
46.0     239
11.0     226
5.0      225
12.0     217
8.0      196
47.0     192
9.0      187
4.0      186
48.0     173
7.0      163
6.0      143
49.0     132
50.0     128
3.0       59
2.0       56
51.0      46
1.0        4
Name: housing_median_age, dtype: int64
```

In [349]:
```
1 data.isnull().sum()
```

Out[349]:
```
longitude               0
latitude                0
housing_median_age      36
total_rooms             0
total_bedrooms          3683
population              29
households              53
median_income           0
median_house_value      0
ocean_proximity         0
dtype: int64
```

In [350]:
```
1 data['housing_median_age'].mean()
```

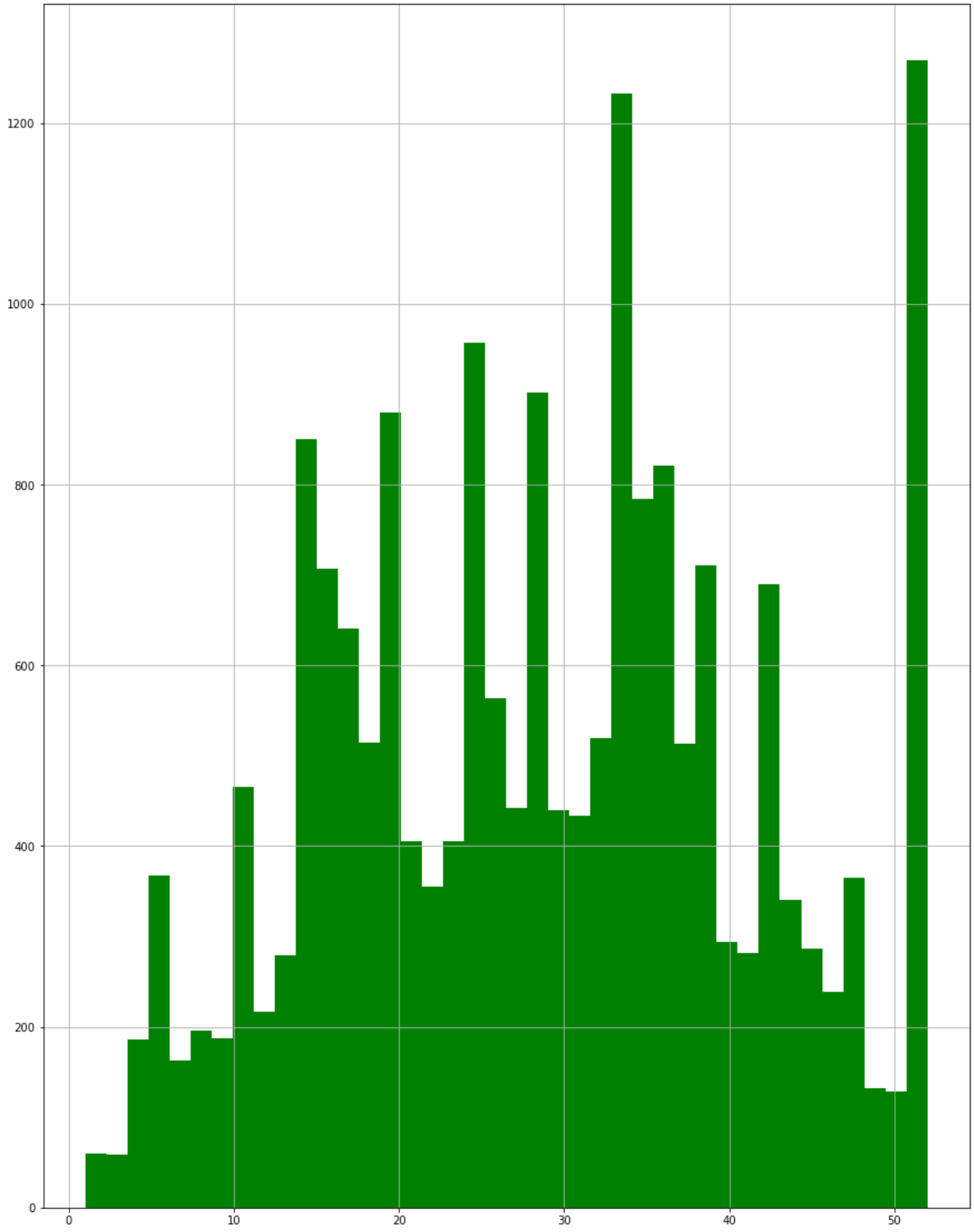
Out[350]:
```
28.857335827098918
```

In [351]:
```
1 data['housing_median_age'].replace(np.nan , data['housing_median_age'].mean(
```

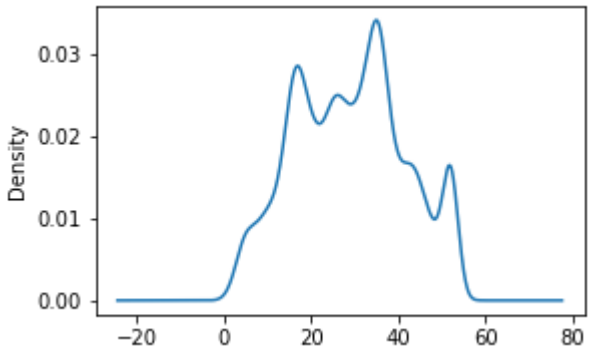
In [352]:
```
1 data.isnull().sum()
```

Out[352]:
```
longitude               0
latitude                0
housing_median_age      0
total_rooms             0
total_bedrooms          3683
population              29
households              53
median_income           0
median_house_value      0
ocean_proximity         0
dtype: int64
```

In [353]:
```
data['housing_median_age'].hist(bins=40, figsize=(15,20), color='green')
plt.show()
```



In [354]:
```
1 data['housing_median_age'].plot(kind='density',subplots=True,layout=(4,4),sh
2 plt.tight_layout()
```



In [355]:
```
1 data.isnull().sum()
```

Out[355]:
```
longitude               0
latitude                0
housing_median_age      0
total_rooms             0
total_bedrooms       3683
population             29
households             53
median_income           0
median_house_value      0
ocean_proximity         0
dtype: int64
```

In [356]:
```
1 data['households'].value_counts()
```

Out[356]:
```
no      3069
282       46
306       45
380       45
375       44
         ...
1125       1
1381       1
1907       1
1060       1
889        1
Name: households, Length: 1696, dtype: int64
```

In [357]:
```
1 data['households'] = data['households'].replace('no' , np.nan)
```

In [358]:
```
1 data['households'].value_counts()
```

Out[358]:
```
282     46
380     45
306     45
375     44
239     42
        ..
1126     1
1381     1
2100     1
2826     1
2125     1
Name: households, Length: 1695, dtype: int64
```

In [359]:
```
1 type('households')
```

Out[359]:
```
str
```

In [360]:
```
1 data['households'] = pd.to_numeric(data['households'])
```

In [361]:
```
1 type('households')
```

Out[361]:
```
str
```

In [254]:
```
1 data['households']=data['households'].astype(int)
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-254-bd3fb5bb726e> in <module>
----> 1 data['households']=data['households'].astype(int)

~\anaconda3\lib\site-packages\pandas\core\generic.py in astype(self, dtype, cop
y, errors)
   5544         else:
   5545             # else, only a single dtype is given
-> 5546             new_data = self._mgr.astype(dtype=dtype, copy=copy, errors=
errors,)
   5547             return self._constructor(new_data).__finalize__(self, metho
d="astype")
   5548

~\anaconda3\lib\site-packages\pandas\core\internals\managers.py in astype(self,
dtype, copy, errors)
    593         self, dtype, copy: bool = False, errors: str = "raise"
    594     ) -> "BlockManager":
--> 595         return self.apply("astype", dtype=dtype, copy=copy, errors=erro
rs)
    596
    597     def convert(

~\anaconda3\lib\site-packages\pandas\core\internals\managers.py in apply(self,
 f, align_keys, **kwargs)
    404                 applied = b.apply(f, **kwargs)
    405             else:
--> 406                 applied = getattr(b, f)(**kwargs)
    407             result_blocks = _extend_blocks(applied, result_blocks)
    408

~\anaconda3\lib\site-packages\pandas\core\internals\blocks.py in astype(self, d
type, copy, errors)
    593             vals1d = values.ravel()
    594             try:
--> 595                 values = astype_nansafe(vals1d, dtype, copy=True)
    596             except (ValueError, TypeError):
    597                 # e.g. astype_nansafe can fail on object-dtype of strin
gs

~\anaconda3\lib\site-packages\pandas\core\dtypes\cast.py in astype_nansafe(arr,
dtype, copy, skipna)
    964
    965         if not np.isfinite(arr).all():
--> 966             raise ValueError("Cannot convert non-finite values (NA or i
nf) to integer")
    967
    968     elif is_object_dtype(arr):

ValueError: Cannot convert non-finite values (NA or inf) to integer
```

In [362]:
```
1 data['households'].value_counts()
```

Out[362]:
```
282.0     46
380.0     45
306.0     45
375.0     44
239.0     42
          ..
1729.0     1
1858.0     1
1583.0     1
2723.0     1
1442.0     1
Name: households, Length: 1695, dtype: int64
```

In [363]:
```
data['households'] = data['households'].fillna(1)
data.head(50)
```

Out[363]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	household
0	-122.23	37.88	41.0	880	129.0	322.0	126
1	-122.22	37.86	21.0	7099	1106.0	2401.0	1138
2	-122.24	37.85	52.0	1467	190.0	496.0	177
3	-122.25	37.85	52.0	1274	235.0	558.0	219
13	-122.26	37.84	52.0	696	191.0	345.0	1
14	-122.26	37.85	52.0	2643	626.0	1212.0	1
15	-122.26	37.85	50.0	1120	283.0	697.0	1
16	-122.27	37.85	52.0	1966	347.0	793.0	1
17	-122.27	37.85	52.0	1228	293.0	648.0	303
18	-122.26	37.84	50.0	2239	455.0	990.0	419
19	-122.27	37.84	52.0	1503	298.0	690.0	275
20	-122.27	37.85	40.0	751	184.0	409.0	166
21	-122.27	37.85	42.0	1639	367.0	929.0	366
22	-122.27	37.84	52.0	2436	541.0	1015.0	478
23	-122.27	37.84	52.0	1688	337.0	853.0	325
24	-122.27	37.84	52.0	2224	437.0	1006.0	422
25	-122.28	37.85	41.0	535	123.0	317.0	119
26	-122.28	37.85	49.0	1130	244.0	607.0	239
27	-122.28	37.85	52.0	1898	421.0	1102.0	397
28	-122.28	37.84	50.0	2082	492.0	1131.0	473
29	-122.28	37.84	52.0	729	160.0	395.0	155
30	-122.28	37.84	49.0	1916	447.0	863.0	378
31	-122.28	37.84	52.0	2153	481.0	1168.0	441
32	-122.27	37.84	48.0	1922	409.0	1026.0	335
33	-122.27	37.83	49.0	1655	366.0	754.0	329
34	-122.27	37.83	51.0	2665	574.0	1258.0	536
35	-122.27	37.83	49.0	1215	282.0	570.0	264
36	-122.27	37.83	48.0	1798	432.0	987.0	374
37	-122.28	37.83	52.0	1511	390.0	901.0	403
38	-122.26	37.83	52.0	1470	330.0	689.0	309
39	-122.26	37.83	52.0	2432	715.0	1377.0	696
40	-122.26	37.83	52.0	1665	419.0	946.0	395
41	-122.26	37.83	51.0	936	311.0	517.0	249

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	household
42	-122.26	37.84	49.0	713	202.0	462.0	189
43	-122.26	37.84	52.0	950	202.0	467.0	198
44	-122.26	37.83	52.0	1443	311.0	660.0	292
45	-122.26	37.83	52.0	1656	420.0	718.0	382
46	-122.26	37.83	50.0	1125	322.0	616.0	304
47	-122.27	37.82	43.0	1007	312.0	558.0	253
48	-122.26	37.82	40.0	624	195.0	423.0	160
49	-122.27	37.82	40.0	946	375.0	700.0	352
50	-122.27	37.82	21.0	896	453.0	735.0	438
51	-122.27	37.82	43.0	1868	456.0	1061.0	407
52	-122.27	37.82	41.0	3221	853.0	1959.0	720
53	-122.27	37.82	52.0	1630	456.0	1162.0	400
54	-122.28	37.82	52.0	1170	235.0	701.0	233
55	-122.28	37.82	52.0	945	243.0	576.0	220
56	-122.28	37.82	52.0	1238	288.0	622.0	259
57	-122.28	37.82	52.0	1489	335.0	728.0	244
58	-122.28	37.82	52.0	1387	341.0	1074.0	304

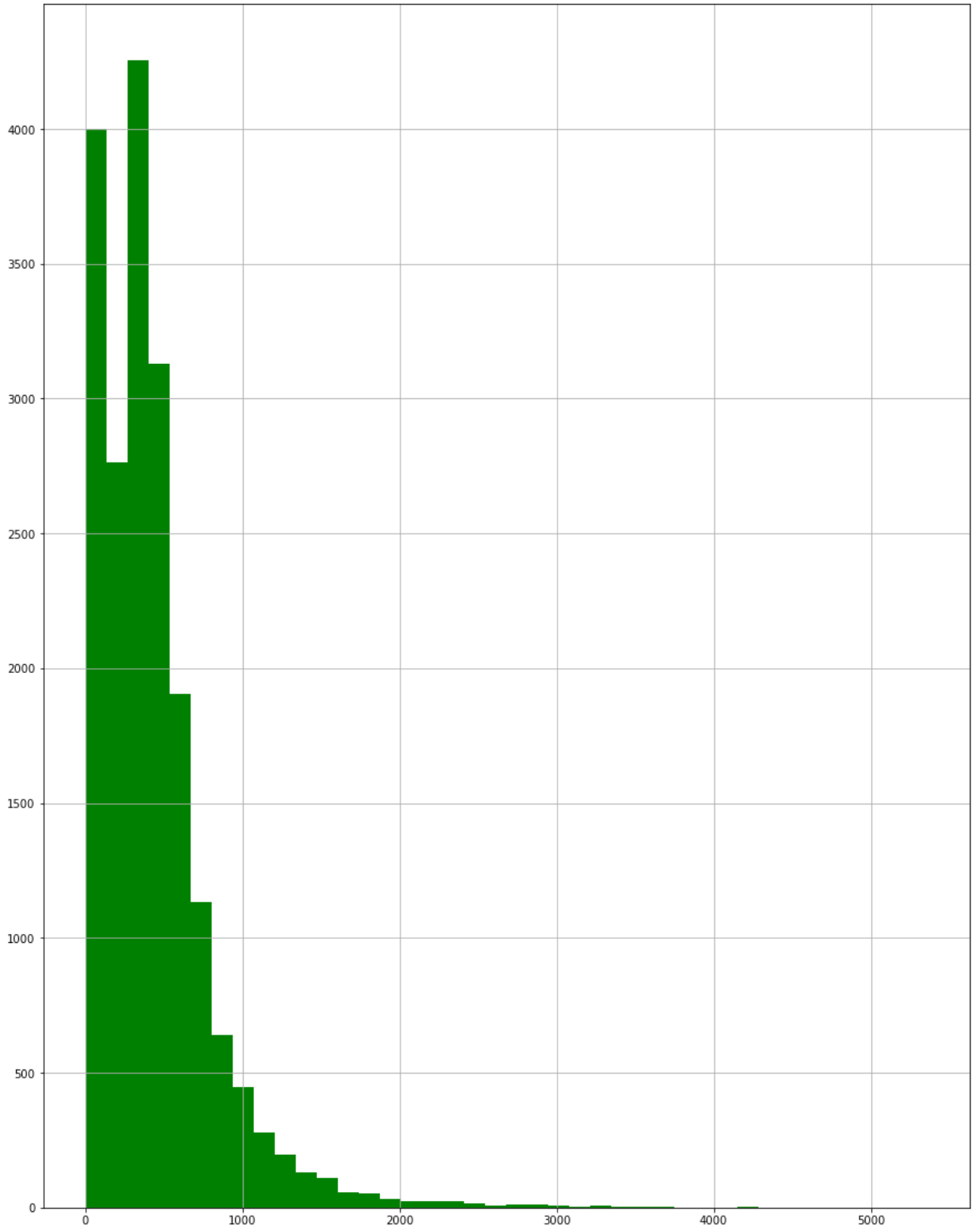
In [364]:
```
1 data['households'].value_counts()
```

Out[364]:
```
1.0       3123
282.0       46
306.0       45
380.0       45
375.0       44
          ...
1858.0       1
1583.0       1
2723.0       1
1060.0       1
1442.0       1
Name: households, Length: 1695, dtype: int64
```

In [365]:
```
1 data.isnull().sum()
```

Out[365]:
```
longitude               0
latitude                0
housing_median_age      0
total_rooms             0
total_bedrooms       3683
population             29
households              0
median_income           0
median_house_value      0
ocean_proximity         0
dtype: int64
```

In [366]:
```
data['households'].hist(bins=40, figsize=(15,20), color='green')
plt.show()
```



In [367]:
```
1 data['total_bedrooms'].max()
```

Out[367]:
```
6210.0
```

In [368]:
```
1 data['total_bedrooms'].min()
```

Out[368]:
```
1.0
```

In [369]:
```
1 data['total_bedrooms'].mean()
```

Out[369]:
```
540.5933594000385
```

In [370]:
```
1 data['total_bedrooms'].mode()
```

Out[370]:
```
0    280.0
dtype: float64
```

In [371]:
```
1 data['total_bedrooms'].median()
```

Out[371]:
```
435.0
```

In [372]:
```
1 data['total_bedrooms'].replace(np.nan , data['total_bedrooms'].mean() , inpl
```

In [373]:
```
1 data.head()
```

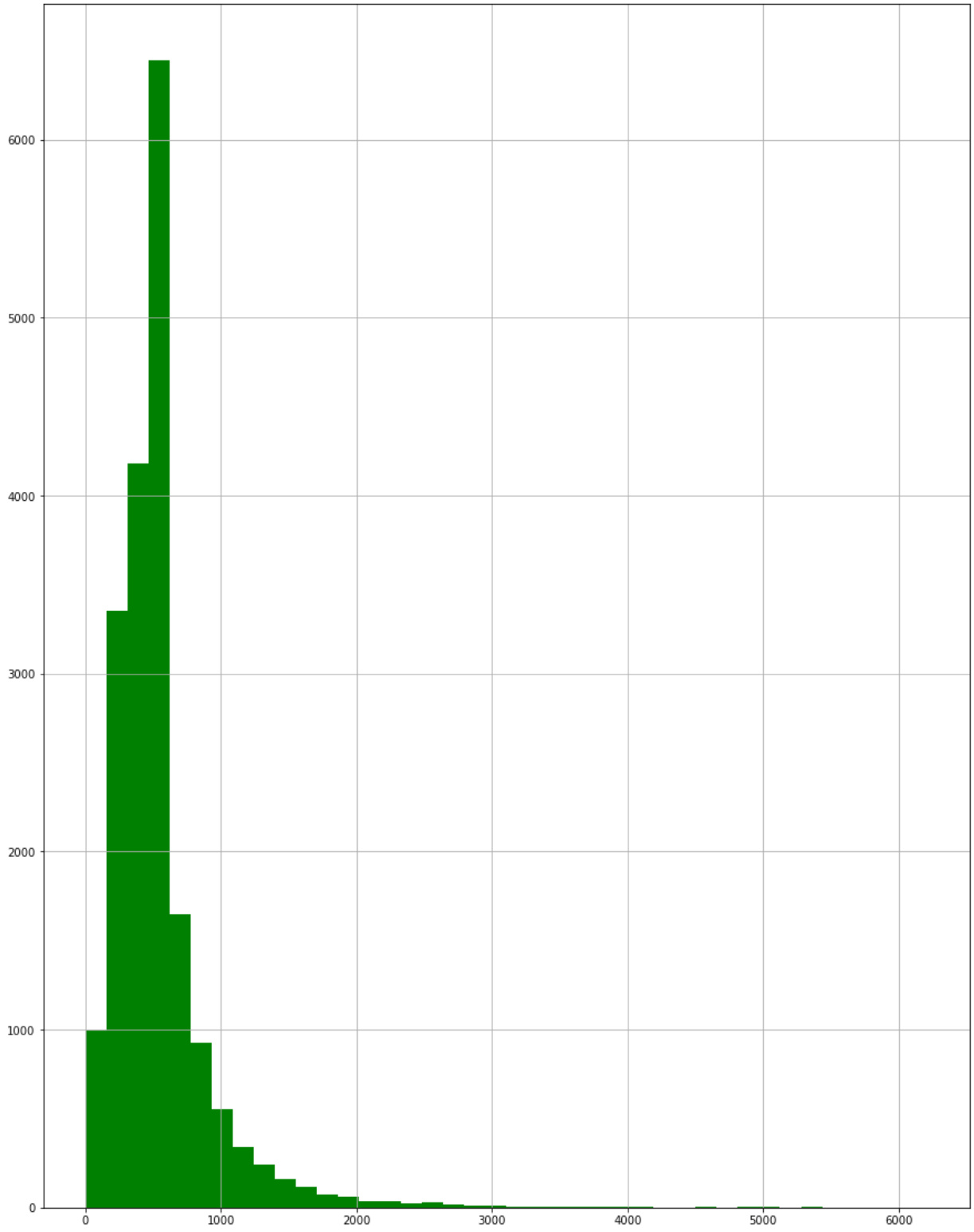
Out[373]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households
0	-122.23	37.88	41.0	880	129.0	322.0	126.0
1	-122.22	37.86	21.0	7099	1106.0	2401.0	1138.0
2	-122.24	37.85	52.0	1467	190.0	496.0	177.0
3	-122.25	37.85	52.0	1274	235.0	558.0	219.0
13	-122.26	37.84	52.0	696	191.0	345.0	1.0

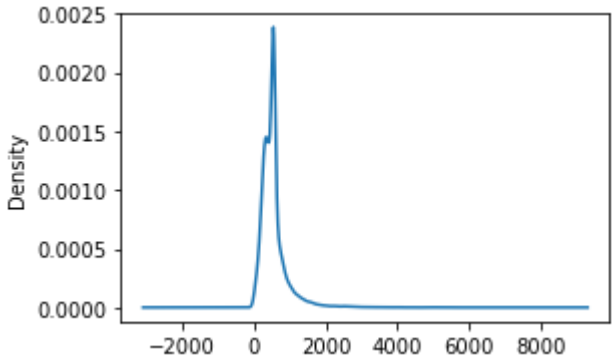
In [374]:
```
1 data.isnull().sum()
```

Out[374]:
```
longitude             0
latitude              0
housing_median_age    0
total_rooms           0
total_bedrooms        0
population           29
households            0
median_income         0
median_house_value    0
ocean_proximity       0
dtype: int64
```

In [375]:
```
data['total_bedrooms'].hist(bins=40, figsize=(15,20), color='green')
plt.show()
```



In [376]:
```
data['total_bedrooms'].plot(kind='density',subplots=True,layout=(4,4),sharex
plt.tight_layout()
```



In [377]:
```
data['population'].value_counts()
```

Out[377]:
```
850.0     23
825.0     23
761.0     22
1227.0    22
891.0     22
          ..
2379.0     1
3191.0     1
123.0      1
2992.0     1
3591.0     1
Name: population, Length: 3819, dtype: int64
```

In [378]:
```
data['population'].max()
```

Out[378]:
```
35682.0
```

In [379]:
```
1 data['population'].min()
```

Out[379]:
```
3.0
```

In [380]:
```
1 data['population'].mean()
```

Out[380]:
```
1427.6867307192938
```

In [381]:
```
1 data['population'].mode()
```

Out[381]:
```
0    825.0
1    850.0
dtype: float64
```

In [382]:
```
1 data['population'].median()
```

Out[382]:
```
1167.0
```

In [383]:
```
1 data['population'].replace(np.nan , data['population'].mean() , inplace = Tr
```

In [384]:
```
1 data.isnull().sum()
```

Out[384]:
```
longitude             0
latitude              0
housing_median_age    0
total_rooms           0
total_bedrooms        0
population            0
households            0
median_income         0
median_house_value    0
ocean_proximity       0
dtype: int64
```

In [385]:
```
1 data.head(5)
```

Out[385]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households
0	-122.23	37.88	41.0	880	129.0	322.0	126.0
1	-122.22	37.86	21.0	7099	1106.0	2401.0	1138.0
2	-122.24	37.85	52.0	1467	190.0	496.0	177.0
3	-122.25	37.85	52.0	1274	235.0	558.0	219.0
13	-122.26	37.84	52.0	696	191.0	345.0	1.0

In [386]:
```
1 from sklearn.preprocessing import LabelEncoder
```

In [387]:

```
labelEncoder = LabelEncoder()
print(data["ocean_proximity"].value_counts())
data["ocean_proximity"] = labelEncoder.fit_transform(data["ocean_proximity"]
data["ocean_proximity"].value_counts()
data.describe()
```

```
<1H OCEAN     8916
INLAND        5950
NEAR BAY      2281
NEAR OCEAN    2132
ISLAND           5
Name: ocean_proximity, dtype: int64
```

Out[387]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	populati
count	19284.000000	19284.000000	19284.000000	19284.000000	19284.000000	19284.0000
mean	-119.610450	35.664153	28.857336	2641.018409	540.593359	1427.6867
std	2.021006	2.121135	12.632397	2184.404567	378.850493	1134.5499
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.0000
25%	-121.840000	33.940000	18.000000	1452.000000	328.000000	790.0000
50%	-118.460000	34.250000	29.000000	2128.000000	519.000000	1167.5000
75%	-118.020000	37.720000	37.000000	3149.000000	586.000000	1725.0000
max	-114.310000	41.950000	52.000000	39320.000000	6210.000000	35682.0000

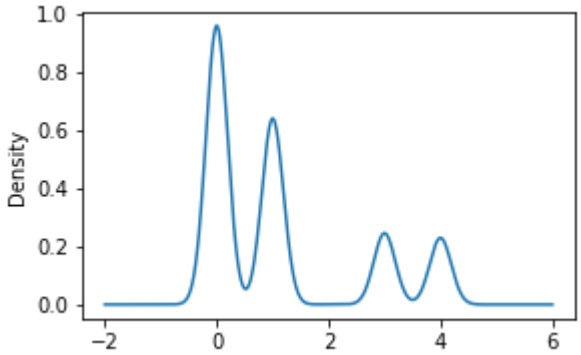
In [388]:
```
1 data["ocean_proximity"].head(50)
```

Out[388]:
```
0     3
1     3
2     3
3     3
13    3
14    3
15    3
16    3
17    3
18    3
19    3
20    3
21    3
22    3
23    3
24    3
25    3
26    3
27    3
28    3
29    3
30    3
31    3
32    3
33    3
34    3
35    3
36    3
37    3
38    3
39    3
40    3
41    3
42    3
43    3
44    3
45    3
46    3
47    3
48    3
49    3
50    3
51    3
52    3
53    3
54    3
55    3
56    3
57    3
58    3
Name: ocean_proximity, dtype: int32
```

In [389]:
```
1 data["ocean_proximity"].value_counts()
```

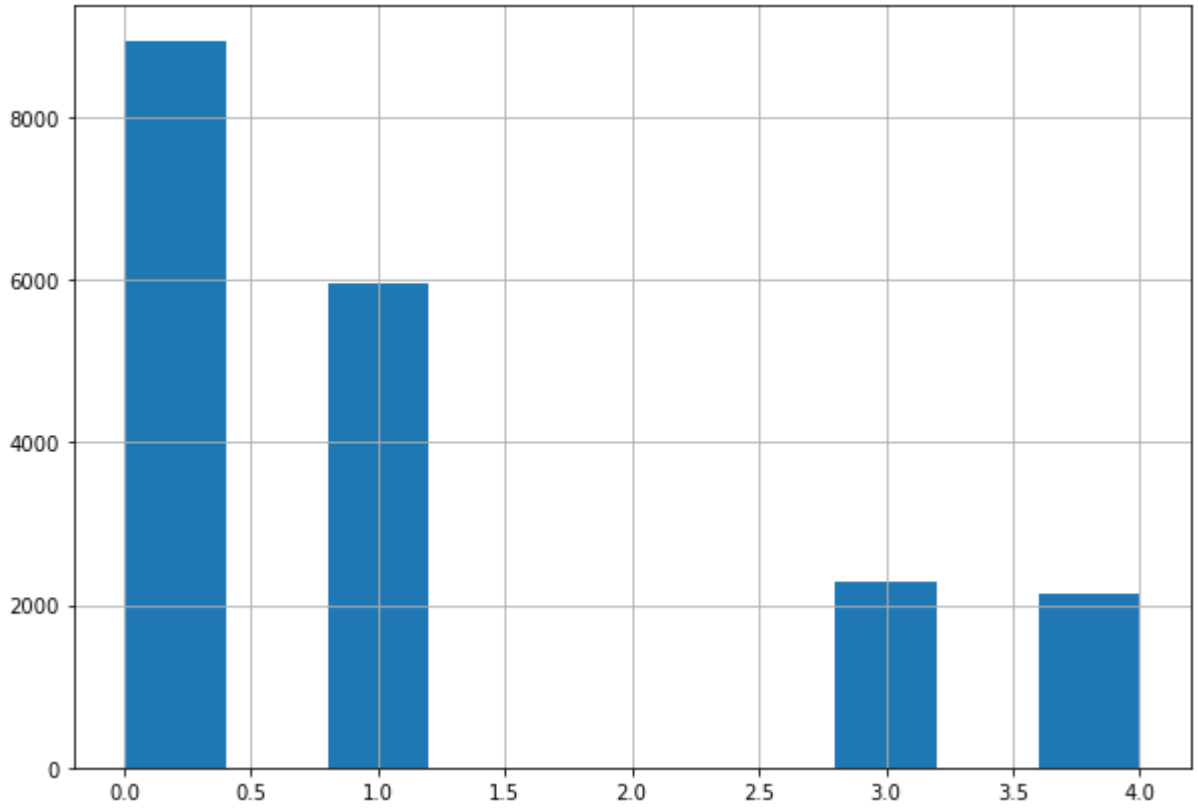
Out[389]:
```
0    8916
1    5950
3    2281
4    2132
2       5
Name: ocean_proximity, dtype: int64
```

In [390]:
```
1 data["ocean_proximity"].plot(kind='density',subplots=True,layout=(4,4),share
2 plt.tight_layout()
```

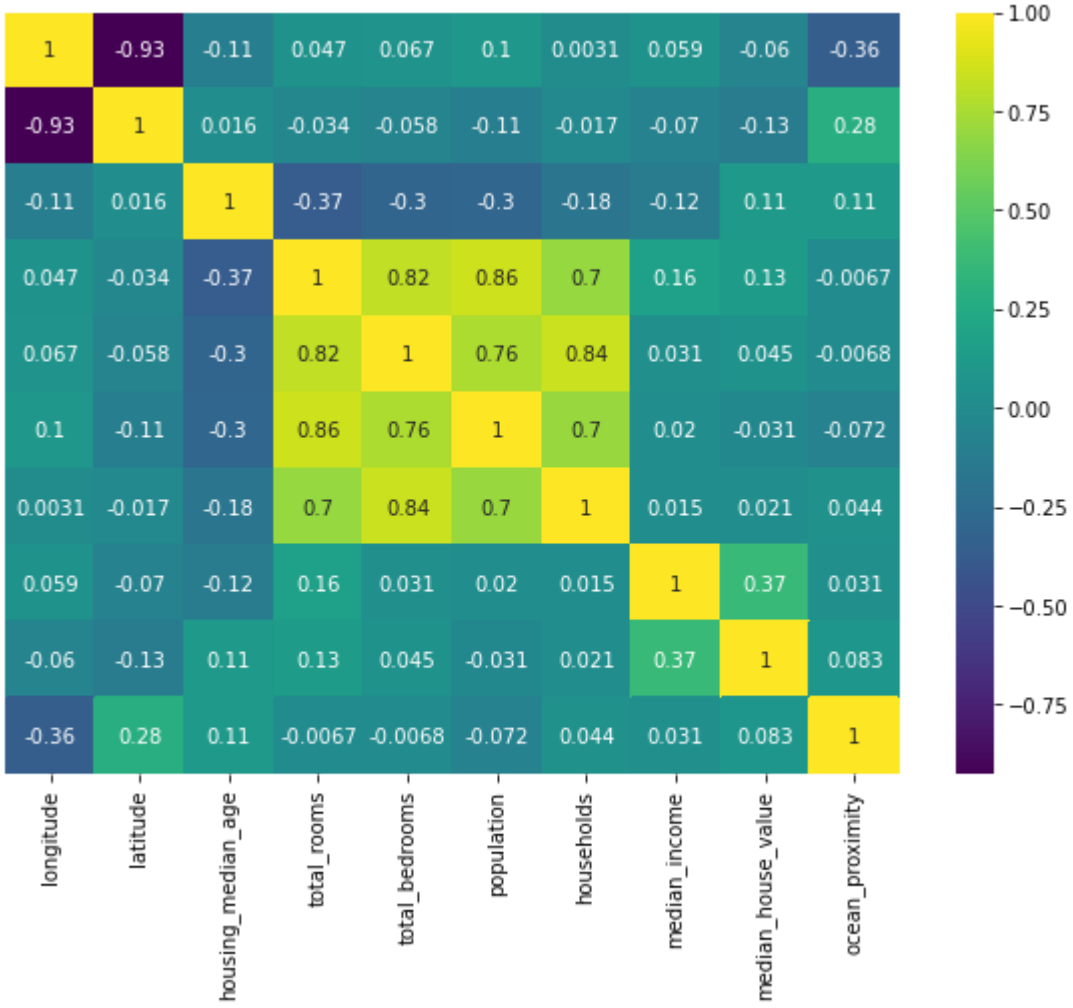


In [391]:
```
1 data["ocean_proximity"].hist(bins=10 , figsize=(10,7))
```

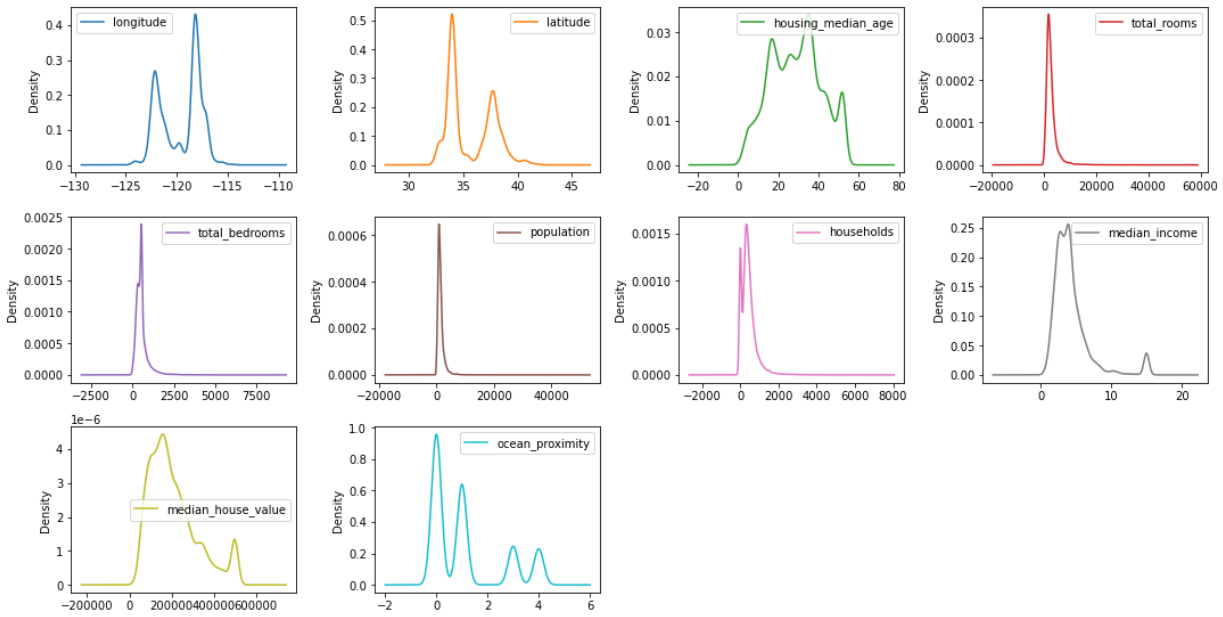
Out[391]: <AxesSubplot:>



In [392]:
```
plt.figure(figsize=(10,7))
sns.heatmap(data.corr(), cmap='viridis', cbar=True , annot=True, yticklabels
plt.show()
```



In [393]:
```
data.plot(kind='density',subplots=True,layout=(4,4),sharex=False,figsize=(15
plt.tight_layout()
```

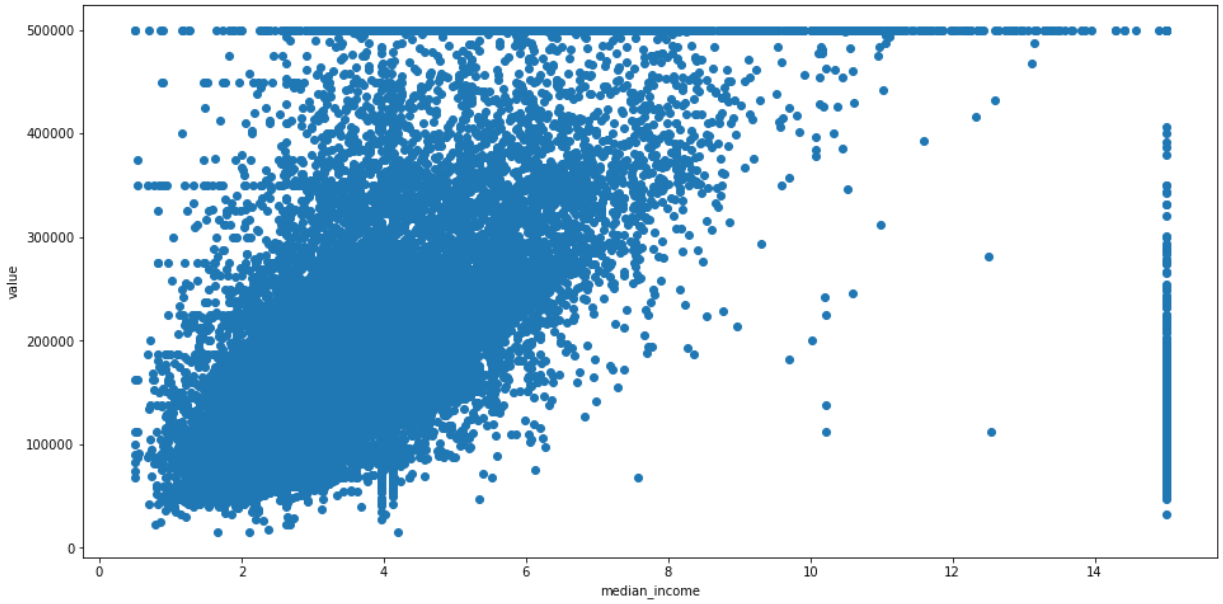


In [394]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 19284 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   longitude           19284 non-null  float64
 1   latitude            19284 non-null  float64
 2   housing_median_age  19284 non-null  float64
 3   total_rooms         19284 non-null  int64  
 4   total_bedrooms      19284 non-null  float64
 5   population          19284 non-null  float64
 6   households          19284 non-null  float64
 7   median_income       19284 non-null  float64
 8   median_house_value  19284 non-null  int64  
 9   ocean_proximity     19284 non-null  int32  
dtypes: float64(7), int32(1), int64(2)
memory usage: 1.5 MB
```

In [426]:
```
fig , ax = plt.subplots(figsize=(16,8))
ax.scatter (data['median_income'],data['median_house_value'])

ax.set_xlabel ('median_income')
ax.set_ylabel('value')
plt.show()
```



In [424]:
```
print("outliers:", data[  (data['median_income']>=8 ) ] ['median_income'].co
```

outliers: 1249

In [425]:
```
print("outliers:", data[  (data['median_income'].iloc[:,4:5  ] ['median_inco
```

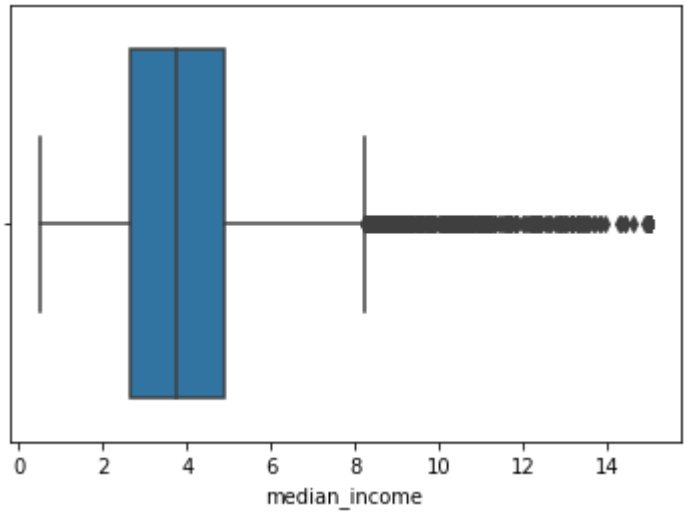
outliers: 668

In [ ]:
```
print("outliers:", data[  (data['median_income']>=8 ) ] ['median_income'].co
```

In [397]:
```
1 sns.boxplot(data['median_income'])
```

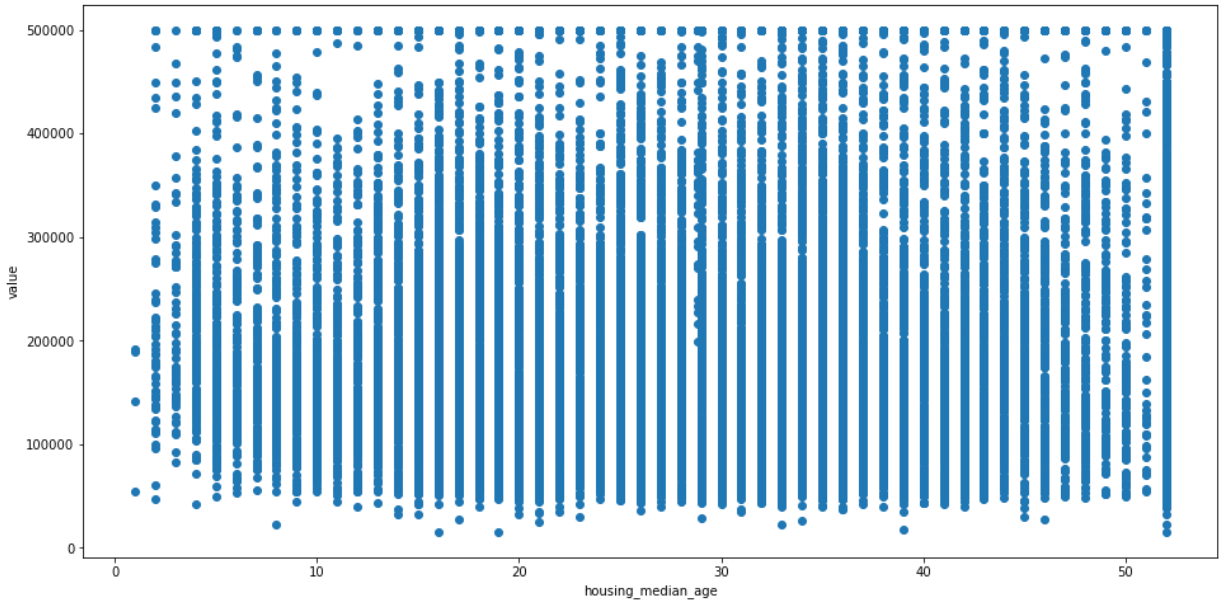
```
C:\Users\Qebaa\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWar
ning: Pass the following variable as a keyword arg: x. From version 0.12, the o
nly valid positional argument will be `data`, and passing other arguments witho
ut an explicit keyword will result in an error or misinterpretation.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

Out[397]:
```
<AxesSubplot:xlabel='median_income'>
```



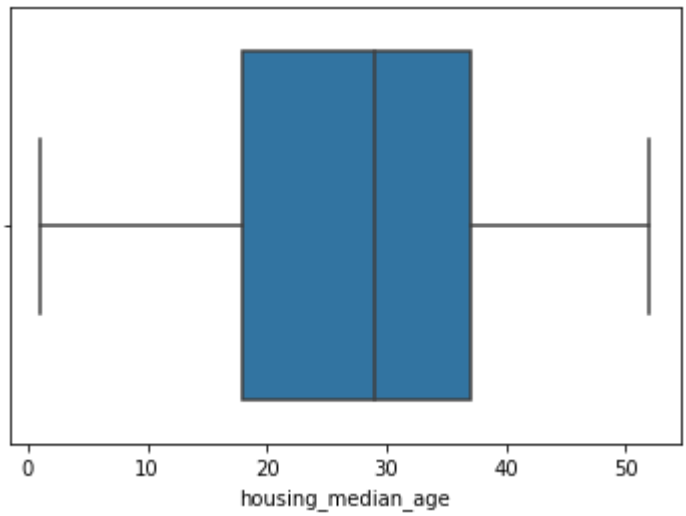
In [427]:
```
fig , ax = plt.subplots(figsize=(16,8))
ax.scatter (data['housing_median_age'],data['median_house_value'])

ax.set_xlabel ('housing_median_age')
ax.set_ylabel('value')
plt.show()
```



In [399]:
```
sns.boxplot(data['housing_median_age'])
plt.show()
```

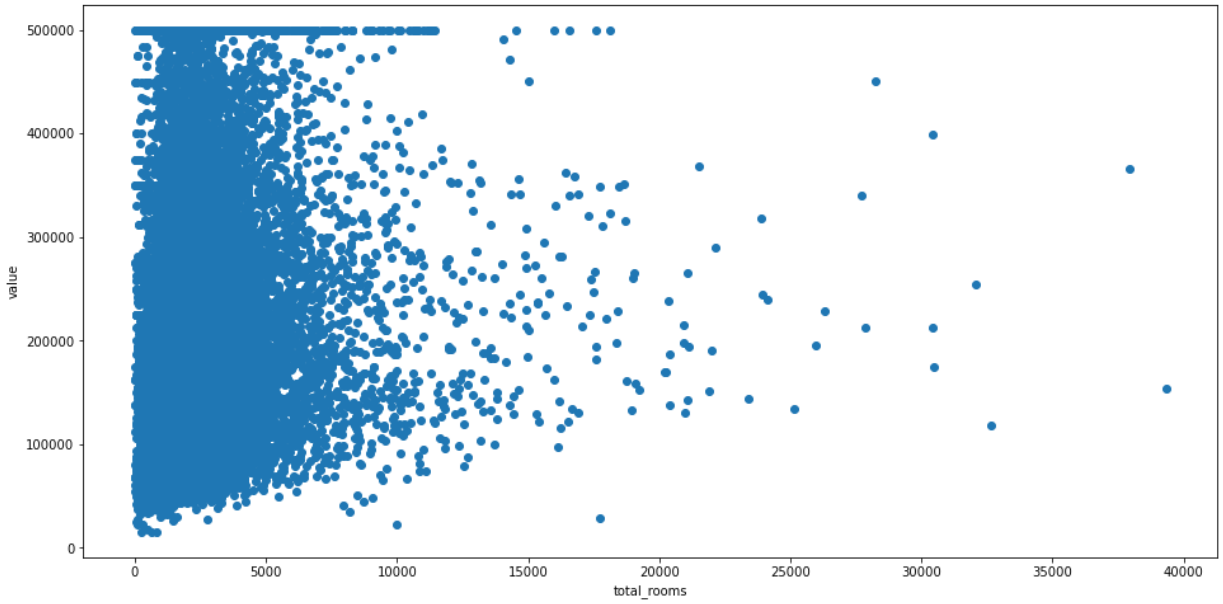
```
C:\Users\Qebaa\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWar
ning: Pass the following variable as a keyword arg: x. From version 0.12, the o
nly valid positional argument will be `data`, and passing other arguments witho
ut an explicit keyword will result in an error or misinterpretation.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



In [428]:

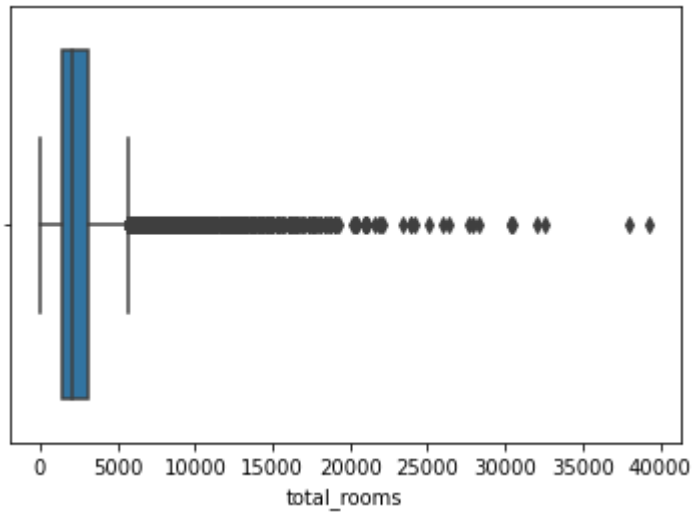
```
fig , ax = plt.subplots(figsize=(16,8))
ax.scatter (data['total_rooms'],data['median_house_value'])

ax.set_xlabel ('total_rooms')
ax.set_ylabel('value')
plt.show()
```



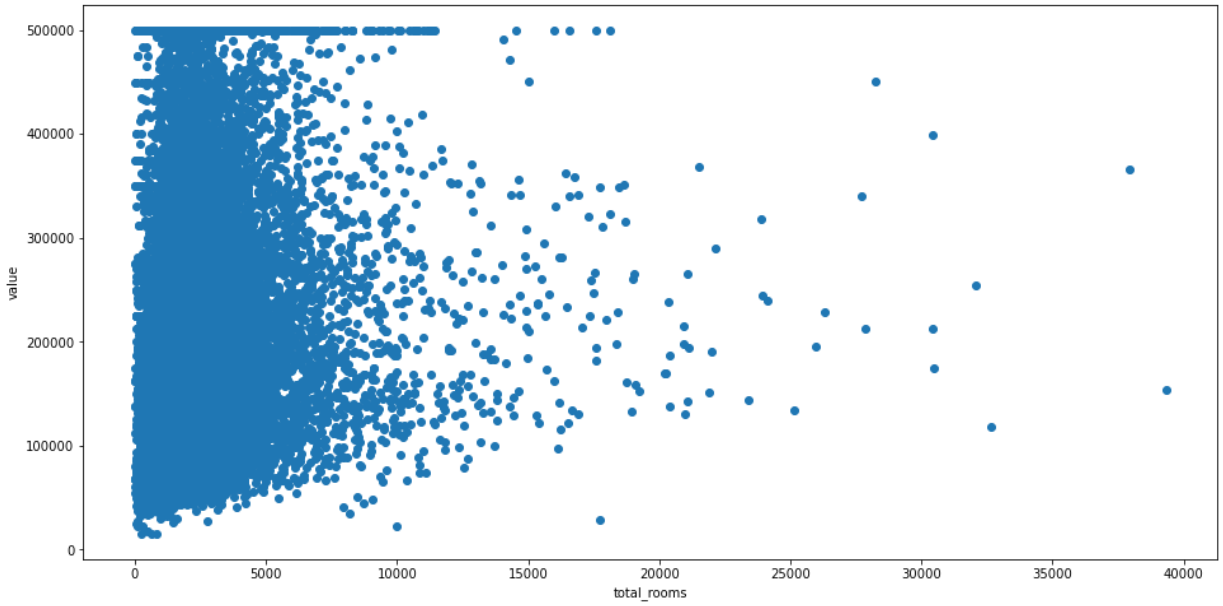
In [401]:
```
sns.boxplot(data['total_rooms'])
plt.show()
```

```
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(
```



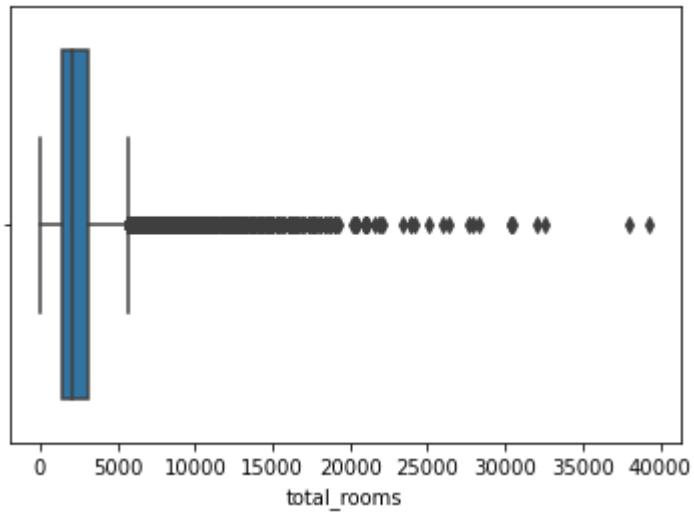
In [429]:
```
fig , ax = plt.subplots(figsize=(16,8))
ax.scatter (data['total_rooms'],data['median_house_value'])

ax.set_xlabel ('total_rooms')
ax.set_ylabel('value')
plt.show()
```



In [403]:
```
1 sns.boxplot(data['total_rooms'])
2 plt.show()
```

```
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(
```



In [404]:
```
1 data['bedroom_per_totalroom']=data['total_bedrooms']/data['total_rooms']
```

In [405]:
```
1 data.head()
```

Out[405]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households
0	-122.23	37.88	41.0	880	129.0	322.0	126.0
1	-122.22	37.86	21.0	7099	1106.0	2401.0	1138.0
2	-122.24	37.85	52.0	1467	190.0	496.0	177.0
3	-122.25	37.85	52.0	1274	235.0	558.0	219.0
13	-122.26	37.84	52.0	696	191.0	345.0	1.0

In [406]:
```
1 data['room_per_households']=data['total_rooms']/data['households']
```

In [407]:
```
1 data.head()
```

Out[407]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households
0	-122.23	37.88	41.0	880	129.0	322.0	126.0
1	-122.22	37.86	21.0	7099	1106.0	2401.0	1138.0
2	-122.24	37.85	52.0	1467	190.0	496.0	177.0
3	-122.25	37.85	52.0	1274	235.0	558.0	219.0
13	-122.26	37.84	52.0	696	191.0	345.0	1.0

In [408]:
```
1 data['population_per_households']=data['population']/data['households']
```

In [409]:
```
1 data.head()
```

Out[409]:

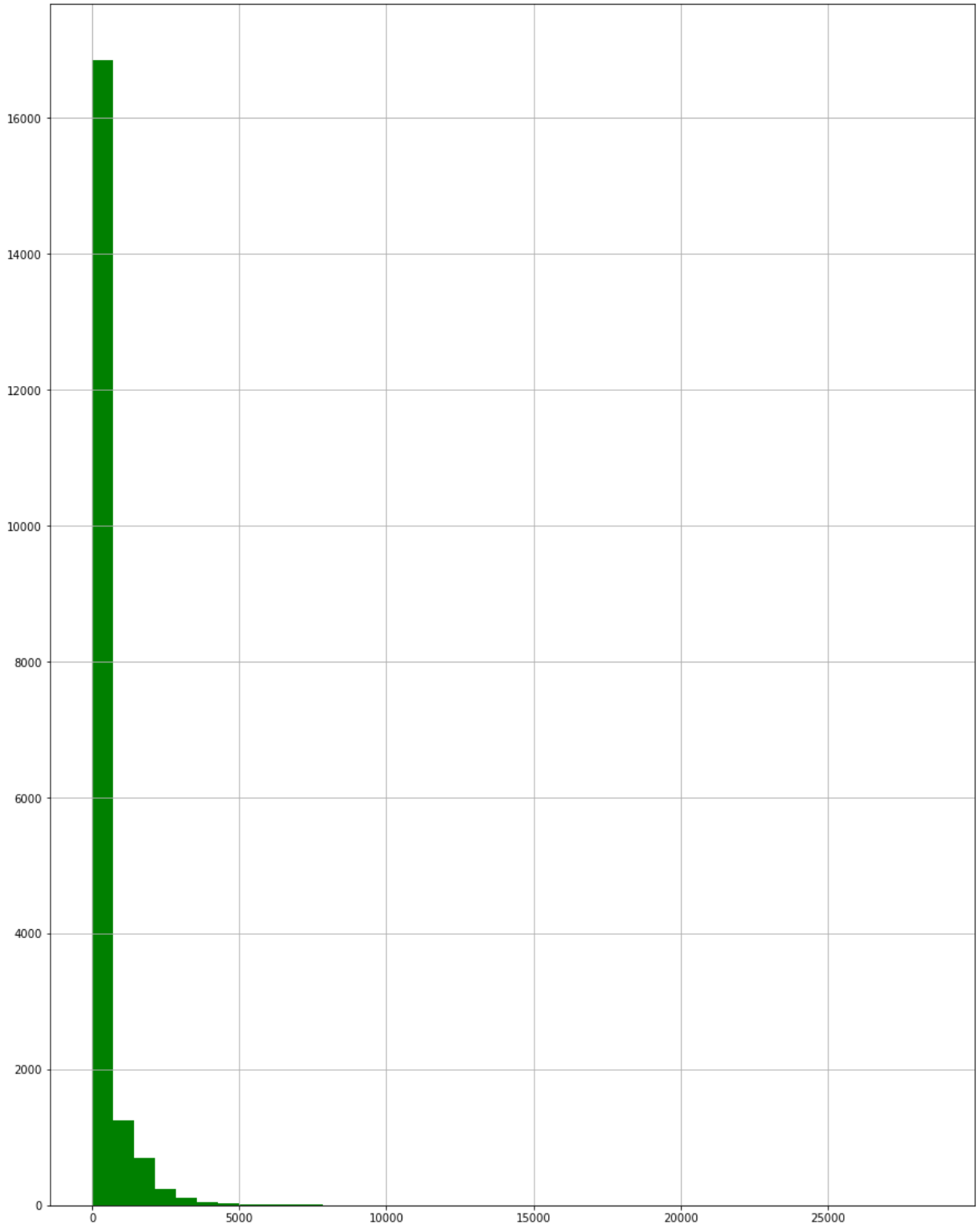
	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households
0	-122.23	37.88	41.0	880	129.0	322.0	126.0
1	-122.22	37.86	21.0	7099	1106.0	2401.0	1138.0
2	-122.24	37.85	52.0	1467	190.0	496.0	177.0
3	-122.25	37.85	52.0	1274	235.0	558.0	219.0
13	-122.26	37.84	52.0	696	191.0	345.0	1.0

In [410]:
```
1 corr_matrix=data.corr()
2 corr_matrix['median_house_value'].sort_values(ascending=True)
```

Out[410]:
```
latitude                   -0.125047
longitude                  -0.060249
population                 -0.030614
bedroom_per_totalroom      -0.028820
households                  0.020986
population_per_households   0.039277
total_bedrooms              0.045374
ocean_proximity             0.083447
room_per_households         0.087382
housing_median_age          0.107534
total_rooms                 0.127434
median_income               0.368293
median_house_value          1.000000
Name: median_house_value, dtype: float64
```

In [411]:
```
data['bedroom_per_totalroom'].hist(bins=40, figsize=(15,20), color='green')
plt.show()
```

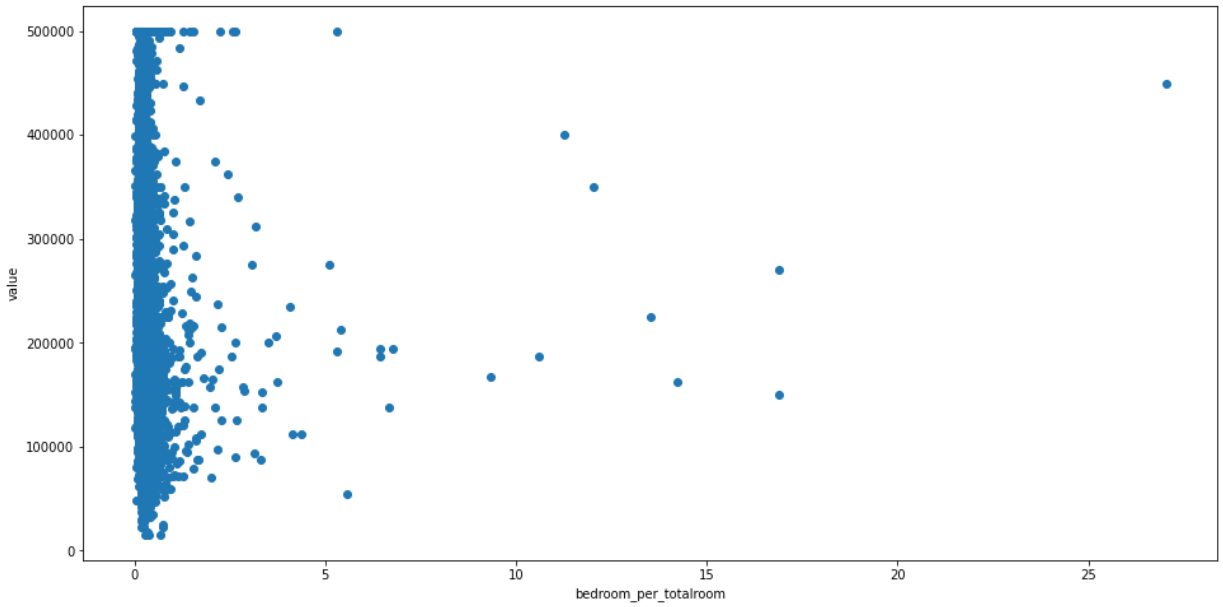
In [412]:
```
data['population_per_households'].hist(bins=40, figsize=(15,20), color='gree
plt.show()
```



In [413]:
```
1 data.isnull().sum()
```

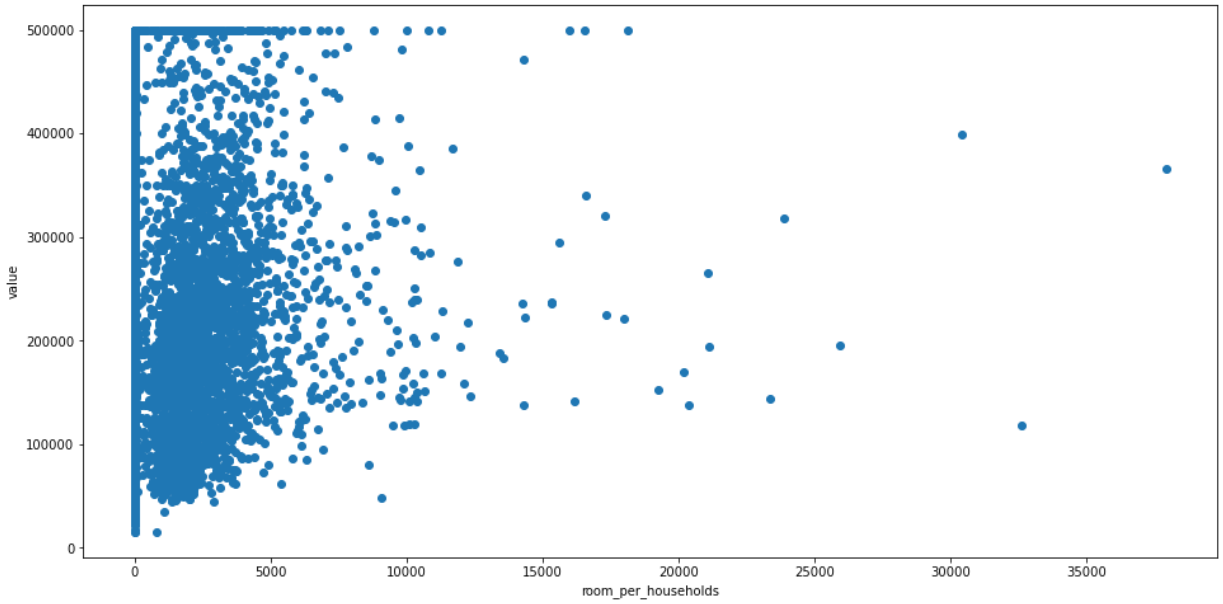
Out[413]:
```
longitude                   0
latitude                    0
housing_median_age          0
total_rooms                 0
total_bedrooms              0
population                  0
households                  0
median_income               0
median_house_value          0
ocean_proximity             0
bedroom_per_totalroom       0
room_per_households         0
population_per_households   0
dtype: int64
```

In [430]:
```
1 fig , ax = plt.subplots(figsize=(16,8))
2 ax.scatter (data['bedroom_per_totalroom'],data['median_house_value'])
3
4 ax.set_xlabel ('bedroom_per_totalroom')
5 ax.set_ylabel('value')
6 plt.show()
```



In [431]:
```
fig , ax = plt.subplots(figsize=(16,8))
ax.scatter (data['room_per_households'],data['median_house_value'])

ax.set_xlabel ('room_per_households')
ax.set_ylabel('value')
plt.show()
```



In [416]:
```
data1=data[['longitude','latitude','housing_median_age','median_income','med
```

In [315]:
```
1 data1
```

Out[315]:

	longitude	latitude	housing_median_age	median_income	median_house_value	ocean_proxir
0	-122.23	37.88	41.0	8.3252	452600	
1	-122.22	37.86	21.0	8.3014	358500	
2	-122.24	37.85	52.0	7.2574	352100	
3	-122.25	37.85	52.0	5.6431	341300	
13	-122.26	37.84	52.0	2.6736	191300	
...	...	...	...	...	...	
20635	-121.09	39.48	25.0	1.5603	78100	
20636	-121.21	39.49	18.0	2.5568	77100	
20637	-121.22	39.43	17.0	1.7000	92300	
20638	-121.32	39.43	18.0	1.8672	84700	
20639	-121.24	39.37	16.0	2.3886	89400	

19284 rows × 9 columns

In [417]:
```
1 data['room_per_households'].value_counts()
```

Out[417]:
```
5.000000       22
4.000000       15
6.000000       15
4.500000       14
5.333333        9
               ..
1367.000000     1
8527.000000     1
5.544118        1
5.900560        1
4.982500        1
Name: room_per_households, Length: 17690, dtype: int64
```

In [418]:
```
1 data['population_per_households'].value_counts()
```

Out[418]:
```
3.000000    26
2.000000    15
2.500000    14
2.666667    12
2.600000    10
            ..
1.897744     1
2.518900     1
2.807281     1
2.577869     1
5.500000     1
Name: population_per_households, Length: 16848, dtype: int64
```

In [419]:
```
1 data['bedroom_per_totalroom'].value_counts()
```

Out[419]:
```
0.250000    26
0.200000    22
0.166667    17
0.181818     9
0.222222     9
            ..
0.205181     1
0.275000     1
0.308119     1
0.160889     1
0.312500     1
Name: bedroom_per_totalroom, Length: 17451, dtype: int64
```

In [420]:
```
1 data1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 19284 entries, 0 to 20639
Data columns (total 9 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   longitude                  19284 non-null  float64
 1   latitude                   19284 non-null  float64
 2   housing_median_age         19284 non-null  float64
 3   median_income              19284 non-null  float64
 4   median_house_value         19284 non-null  int64
 5   ocean_proximity            19284 non-null  int32
 6   bedroom_per_totalroom      19284 non-null  float64
 7   room_per_households        19284 non-null  float64
 8   population_per_households  19284 non-null  float64
dtypes: float64(7), int32(1), int64(1)
memory usage: 1.4 MB
```

In [6]:
```
def find_outliers(x):
    q1 = x.quantile(.25)
    q3 = x.quantile(.75)
    iqr = q3 - q1
    floor = q1 - 1.5*iqr
    ceiling = q3 + 1.5*iqr
    outlier_indices = list(x.index[(x < floor) | (x > ceiling)])
    outlier_values = list(x[outlier_indices])
    return outlier_indices, outlier_values
```

In [2]:
```
median_house_value_indices, median_house_value_values = find_outliers(data['
print("Outliers for median_house_value")
print(np.sort(median_house_value_values))



print("Outliers for median_income")
median_income_indices, median_income_values = find_outliers(data['median_inc
print(np.sort(median_income_values))
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-2-5b08abcb60ce> in <module>
----> 1 median_house_value_indices, median_house_value_values = find_outliers(d
ata['median_house_value'])
      2 print("Outliers for median_house_value")
      3 print(np.sort(median_house_value_values))
      4
      5

NameError: name 'data' is not defined
```

In [ ]:
```
1
```