In [1]:

```
import pandas as pd
data = pd.read_csv("C:\\Users\\kmit\\Desktop\\housing.csv",",")
print(data.head(5))
```

```
   longitude  latitude  housing_median_age  total_rooms  total_bedrooms  \
0    -122.23     37.88                41.0        880.0           129.0   
1    -122.22     37.86                21.0       7099.0          1106.0   
2    -122.24     37.85                52.0       1467.0           190.0   
3    -122.25     37.85                52.0       1274.0           235.0   
4    -122.25     37.85                52.0       1627.0           280.0   

   population  households  median_income  median_house_value ocean_proximity  

0       322.0       126.0         8.3252            452600.0        NEAR BAY  

1      2401.0      1138.0         8.3014            358500.0        NEAR BAY  

2       496.0       177.0         7.2574            352100.0        NEAR BAY  

3       558.0       219.0         5.6431            341300.0        NEAR BAY  

4       565.0       259.0         3.8462            342200.0        NEAR BAY  
```

In [2]:

```
print("total samples....\n",data.size)
print("Null Values....\n", data.isnull().sum())
#print(data.isnull().count())
```

```
total samples....
 206400
Null Values....
 longitude               0
latitude              0
housing_median_age    0
total_rooms           0
total_bedrooms      207
population            0
households            0
median_income         0
median_house_value    0
ocean_proximity       0
dtype: int64
```

In [3]:

```
d1 = data.dropna(subset=['total_bedrooms'])
print(data.shape,d1.shape)
```

```
(20640, 10) (20433, 10)
```

In [4]:

```
d1.cov()
```

Out[4]:

	longitude	latitude	housing_median_age	total_rooms	total_b
longitude	4.014324	-3.957670	-2.758919	1.991284e+02	5.876
latitude	-3.957670	4.563981	0.320091	-1.711788e+02	-6.029
housing_median_age	-2.758919	0.320091	158.553558	-9.923225e+03	-1.700
total_rooms	199.128445	-171.178818	-9923.224538	4.775403e+06	8.567
total_bedrooms	58.768508	-60.299623	-1700.312817	8.567306e+05	1.775
population	227.660858	-263.874646	-4220.630517	2.122942e+06	4.191
households	43.286878	-58.619704	-1457.475788	7.677502e+05	1.578
median_income	-0.059174	-0.323087	-2.828672	8.213000e+02	-6.180
median_house_value	-10499.897668	-35669.333210	154703.602850	3.362452e+07	2.416

In [5]:

```
d1.corr()
```

Out[5]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	p
longitude	1.000000	-0.924616	-0.109357	0.045480	0.069608	
latitude	-0.924616	1.000000	0.011899	-0.036667	-0.066983	
housing_median_age	-0.109357	0.011899	1.000000	-0.360628	-0.320451	
total_rooms	0.045480	-0.036667	-0.360628	1.000000	0.930380	
total_bedrooms	0.069608	-0.066983	-0.320451	0.930380	1.000000	
population	0.100270	-0.108997	-0.295787	0.857281	0.877747	
households	0.056513	-0.071774	-0.302768	0.918992	0.979728	
median_income	-0.015550	-0.079626	-0.118278	0.197882	-0.007723	
median_house_value	-0.045398	-0.144638	0.106432	0.133294	0.049686	

In [6]:

```
d1.corr()['median_house_value']
```

Out[6]:

```
longitude            -0.045398
latitude             -0.144638
housing_median_age    0.106432
total_rooms           0.133294
total_bedrooms        0.049686
population           -0.025300
households            0.064894
median_income         0.688355
median_house_value    1.000000
Name: median_house_value, dtype: float64
```

In [7]:

```
d1.corr()['median_house_value'].sort_values()[::-1]
```

Out[7]:

```
median_house_value    1.000000
median_income         0.688355
total_rooms           0.133294
housing_median_age    0.106432
households            0.064894
total_bedrooms        0.049686
population           -0.025300
longitude            -0.045398
latitude             -0.144638
Name: median_house_value, dtype: float64
```

In [8]:

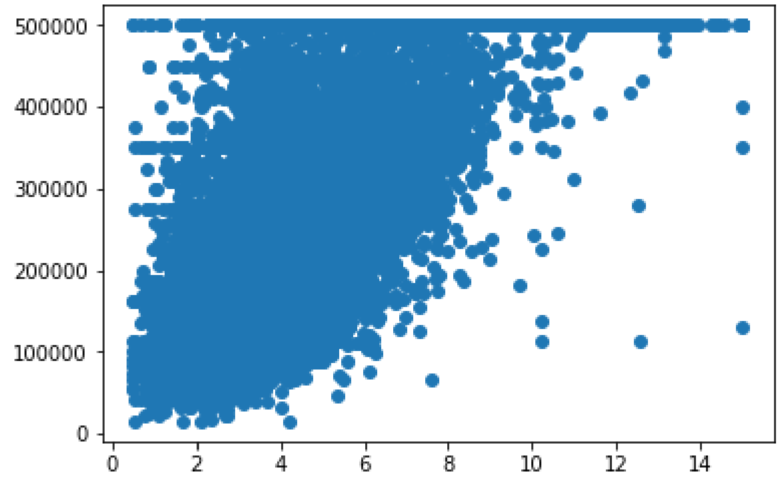
```
corr_cols=d1.corr()['median_house_value'].sort_values()[::-1]
corr_cols[1:4]
corr_cols.index
```

Out[8]:

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

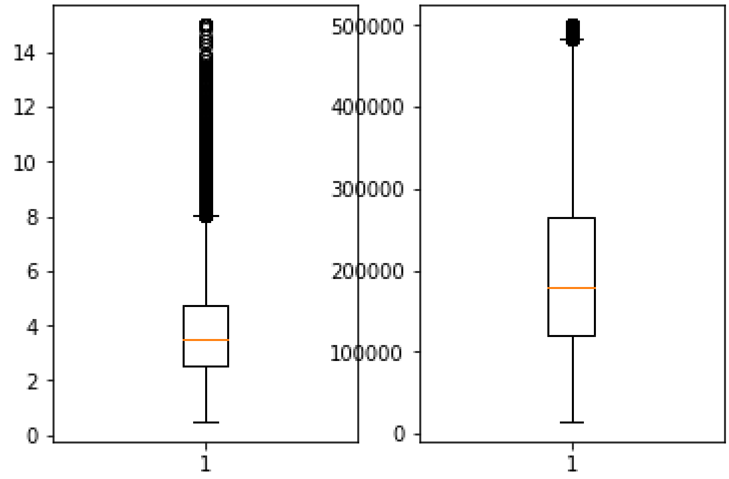
In [9]:

```
import matplotlib.pyplot as plt
plt.scatter(d1.median_income,d1.median_house_value)
plt.show()
```



In [10]:

```
plt.subplot(121)
plt.boxplot(d1.median_income)
plt.subplot(122)
plt.boxplot(d1.median_house_value)
plt.show()
```



In [11]:

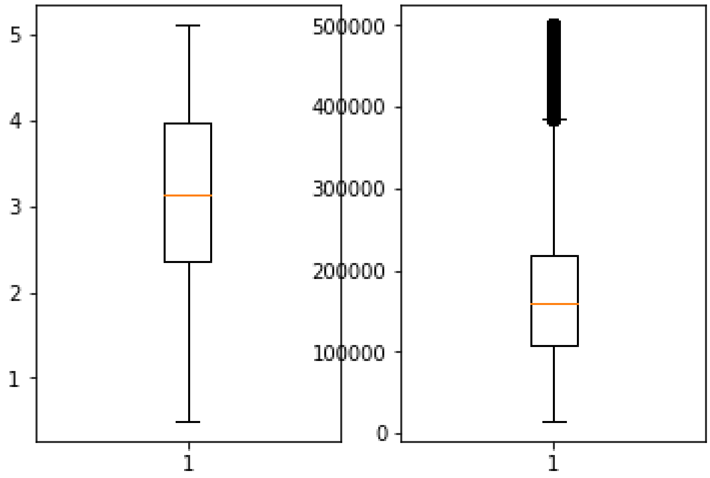
```
d2=d1[d1.median_income<d1.median_income.quantile(0.8)]
d2.shape
```

Out[11]:

```
(16346, 10)
```

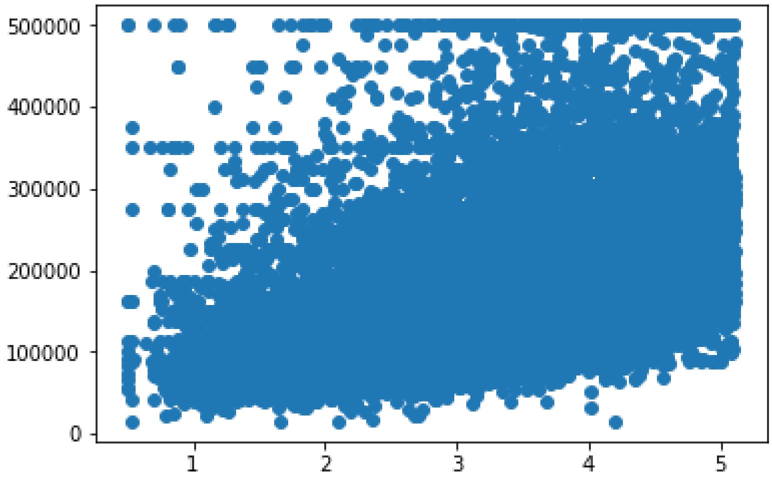
In [12]:

```
plt.subplot(121)
plt.boxplot(d2.median_income)
plt.subplot(122)
plt.boxplot(d2.median_house_value)
plt.show()
```



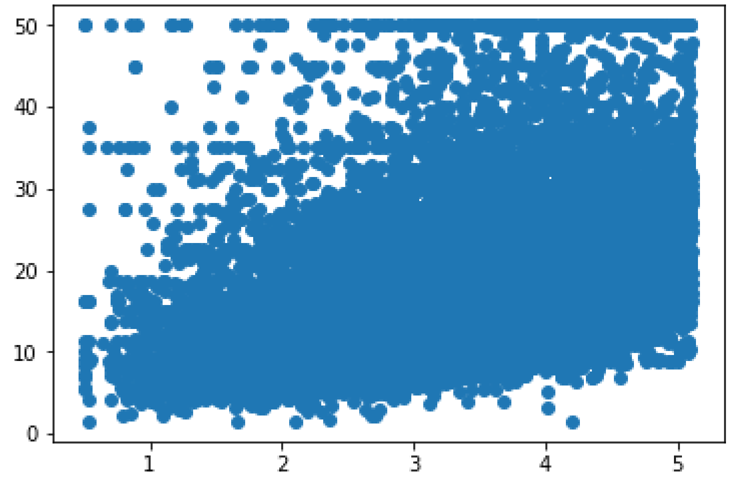
In [13]:

```
import matplotlib.pyplot as plt
plt.scatter(d2.median_income,d2.median_house_value)
plt.show()
```



In [14]:

```
import matplotlib.pyplot as plt
plt.scatter(d2.median_income,d2.median_house_value/10000)
plt.show()
```



In [15]:

```
d3=d2.drop(corr_cols.index[2:],axis=1)
print(d3)
d3=d3.drop(['ocean_proximity'],axis=1)
print(d3.shape)
#d3
d3.median_house_value = d3.median_house_value/100000
```

```
       median_income  median_house_value ocean_proximity
4             3.8462            342200.0        NEAR BAY
5             4.0368            269700.0        NEAR BAY
6             3.6591            299200.0        NEAR BAY
7             3.1200            241400.0        NEAR BAY
8             2.0804            226700.0        NEAR BAY
9             3.6912            261100.0        NEAR BAY
10            3.2031            281500.0        NEAR BAY
11            3.2705            241800.0        NEAR BAY
12            3.0750            213500.0        NEAR BAY
13            2.6736            191300.0        NEAR BAY
14            1.9167            159200.0        NEAR BAY
15            2.1250            140000.0        NEAR BAY
16            2.7750            152500.0        NEAR BAY
17            2.1202            155500.0        NEAR BAY
18            1.9911            158700.0        NEAR BAY
19            2.6033            162900.0        NEAR BAY
20            1.3578            147500.0        NEAR BAY
21            1.7135            159800.0        NEAR BAY
22            1.7250            113900.0        NEAR BAY
23            2.1806             99700.0        NEAR BAY
24            2.6000            132600.0        NEAR BAY
25            2.4038            107500.0        NEAR BAY
26            2.4597             93800.0        NEAR BAY
27            1.8080            105500.0        NEAR BAY
28            1.6424            108900.0        NEAR BAY
29            1.6875            132000.0        NEAR BAY
30            1.9274            122300.0        NEAR BAY
31            1.9615            115200.0        NEAR BAY
32            1.7969            110400.0        NEAR BAY
33            1.3750            104900.0        NEAR BAY
...              ...                 ...             ...
20610         1.3631             45500.0          INLAND
20611         1.2857             47000.0          INLAND
20612         1.4934             48300.0          INLAND
20613         1.4958             53400.0          INLAND
20614         2.4695             58000.0          INLAND
20615         2.3598             57500.0          INLAND
20616         2.0469             55100.0          INLAND
20617         3.3021             70800.0          INLAND
20618         2.2500             63400.0          INLAND
20619         2.7303             99100.0          INLAND
20620         4.5625            100000.0          INLAND
20621         2.3661             77500.0          INLAND
20622         2.4167             67000.0          INLAND
20623         2.8235             65500.0          INLAND
20624         3.0739             87200.0          INLAND
20625         4.1250             72000.0          INLAND
20626         2.1667             93800.0          INLAND
20627         3.0000            162500.0          INLAND
20628         2.5952             92400.0          INLAND
20629         2.0943            108300.0          INLAND
```

```
20630         3.5673           112000.0          INLAND
20631         3.5179           107200.0          INLAND
20632         3.1250           115600.0          INLAND
20633         2.5495            98300.0          INLAND
20634         3.7125           116800.0          INLAND
20635         1.5603            78100.0          INLAND
20636         2.5568            77100.0          INLAND
20637         1.7000            92300.0          INLAND
20638         1.8672            84700.0          INLAND
20639         2.3886            89400.0          INLAND

[16346 rows x 3 columns]
(16346, 2)
```

In [16]:

```
testsize=(int)(d3.shape[0]*0.30)
```

In [17]:

```
train=d3[:-testsize]
print(train.shape)
test=d3[-testsize:]
print(test.shape)
```

```
(11443, 2)
(4903, 2)
```

In [18]:

```
import numpy as np
train_x=train['median_income']
train_x=train_x[:,np.newaxis]
train_y=train['median_house_value']
train_y=train_y[:,np.newaxis]
test_x = test['median_income']
test_x = test_x[:,np.newaxis]
test_y=test['median_house_value']
test_y=test_y[:,np.newaxis]
train_x.shape
```

Out[18]:

```
(11443, 1)
```

In [19]:

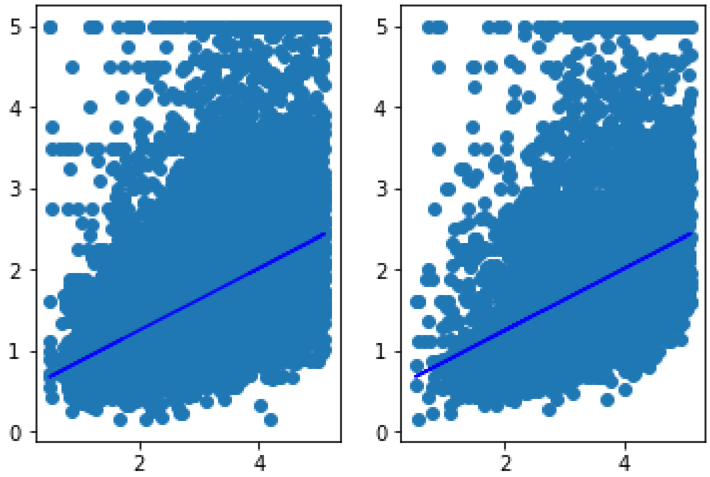
```
from sklearn import linear_model
lm=linear_model.LinearRegression()
lm.fit(train_x,train_y)
train_pred=lm.predict(train_x)
#train_y-train_pred
```

In [20]:

```
#plt.scatter(train.median_income,train.median_house_value)
plt.subplot(121)
plt.scatter(train_x,train_y)
plt.plot(train_x,train_pred,'b')

test_pred=lm.predict(test_x)
plt.subplot(122)
plt.scatter(test_x,test_y)
plt.plot(test_x,test_pred,'b')

plt.show()
```



In [22]:

```
print(d2.median_income.size,train.median_income.size)
```

```
16346 11443
```

In [23]:

```
import matplotlib.pyplot as plt
plt.boxplot(d3.total_bedrooms)
plt.show()
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
<ipython-input-23-e1b635c37c1d> in <module>()
      1 import matplotlib.pyplot as plt
----> 2 plt.boxplot(d3.total_bedrooms)
      3 plt.show()

~\Anaconda3\lib\site-packages\pandas\core\generic.py in __getattr__(self, na
me)
   3079             if name in self._info_axis:
   3080                 return self[name]
-> 3081             return object.__getattribute__(self, name)
   3082 
   3083     def __setattr__(self, name, value):

AttributeError: 'DataFrame' object has no attribute 'total_bedrooms'
```

In [ ]: