In [1]:
```
#Exploring data and correlation of the variables
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
housing_data = pd.read_csv('datasets/housing.csv')
```

In [3]:
```
#print first 5 sample of the dataset
#median_house_value for target in regression
housing_data.head()
```

Out[3]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_incc
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8

In [4]:
```
#shape of datasets (number of entries, variables)
housing_data.shape
```

Out[4]:
```
(20640, 10)
```

In [6]:
```
#remove missing entry
housing_data = housing_data.dropna()
```

In [7]:
```
housing_data.shape
```

Out[7]:
```
(20433, 10)
```

In [8]:
```
#view variable statistics
housing_data.describe()
```

Out[8]:

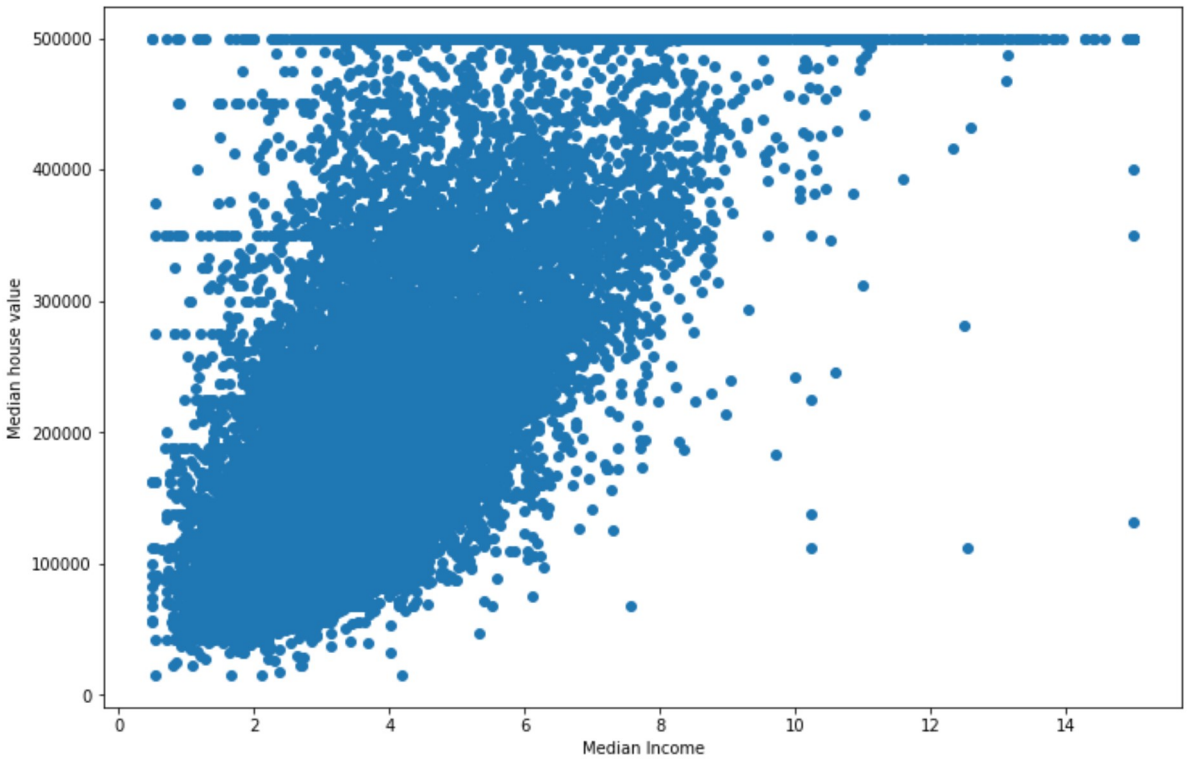
	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	house
count	20433.000000	20433.000000	20433.000000	20433.000000	20433.000000	20433.000000	20433.0
mean	-119.570689	35.633221	28.633094	2636.504233	537.870553	1424.946949	499.4
std	2.003578	2.136348	12.591805	2185.269567	421.385070	1133.208490	382.2
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	1.0
25%	-121.800000	33.930000	18.000000	1450.000000	296.000000	787.000000	280.0
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.000000	409.0
75%	-118.010000	37.720000	37.000000	3143.000000	647.000000	1722.000000	604.0
max	-114.310000	41.950000	52.000000	39320.000000	6445.000000	35682.000000	6082.0

In [9]:
```
#this is unique because the value is string
housing_data['ocean_proximity'].unique()
```

Out[9]:
```
array(['NEAR BAY', '<1H OCEAN', 'INLAND', 'NEAR OCEAN', 'ISLAND'],
      dtype=object)
```

In [13]:
```
fig, ax = plt.subplots(figsize=(12,8))
plt.scatter(housing_data['median_income'], housing_data['median_house_value'])
plt.xlabel('Median Income')
plt.ylabel('Median house value')
```

Out[13]:
```
Text(0, 0.5, 'Median house value')
```



In [14]:
```
#to view correlation between every correlation. as you can see, in median_house_val
ue, median_income variable is the most correlated to the house price. range from -1
to 1.
housing_data_corr = housing_data.corr()
housing_data_corr
```

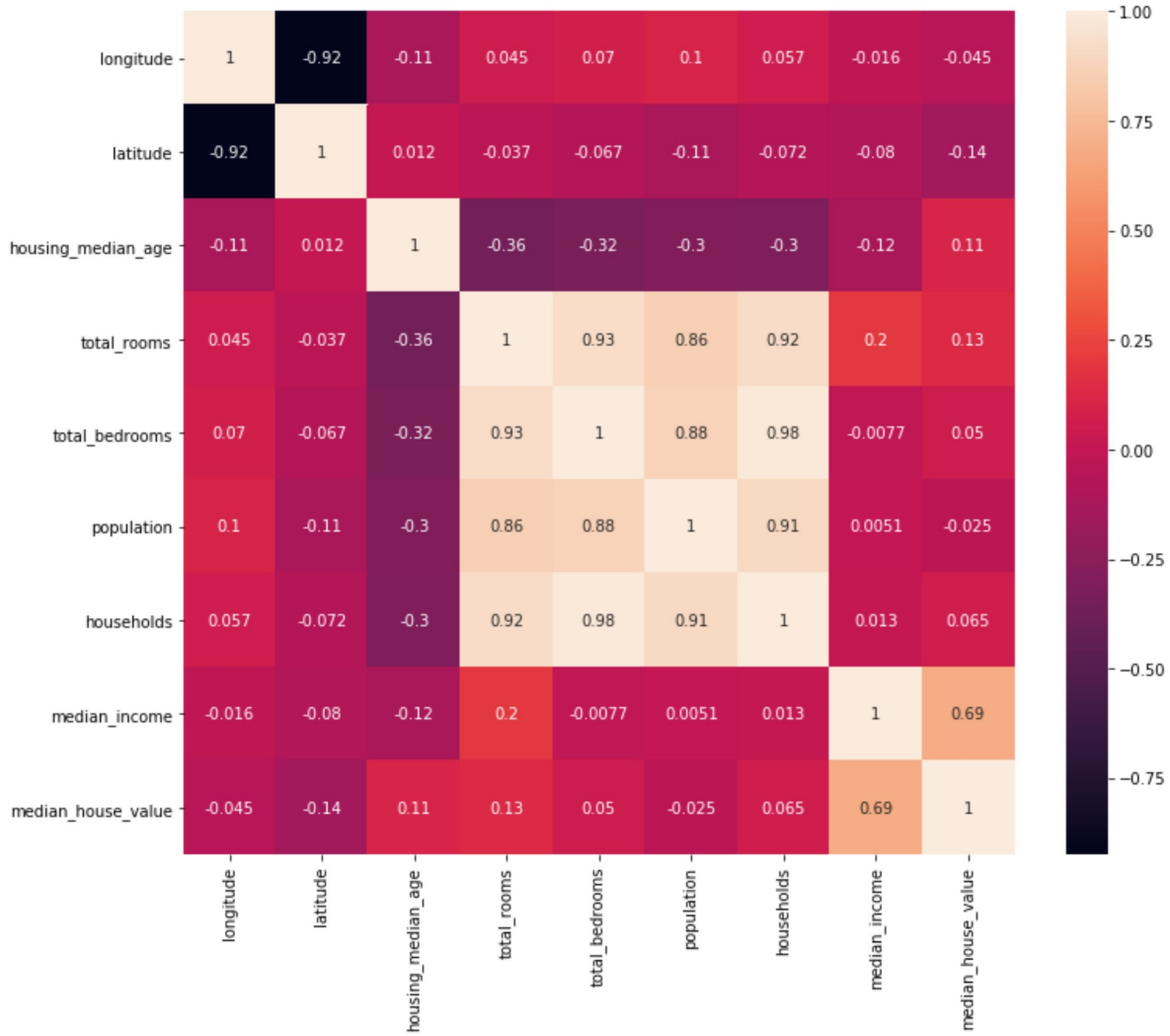
Out[14]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	hou
longitude	1.000000	-0.924616	-0.109357	0.045480	0.069608	0.100270	0
latitude	-0.924616	1.000000	0.011899	-0.036667	-0.066983	-0.108997	-0
housing_median_age	-0.109357	0.011899	1.000000	-0.360628	-0.320451	-0.295787	-0
total_rooms	0.045480	-0.036667	-0.360628	1.000000	0.930380	0.857281	0
total_bedrooms	0.069608	-0.066983	-0.320451	0.930380	1.000000	0.877747	0
population	0.100270	-0.108997	-0.295787	0.857281	0.877747	1.000000	0
households	0.056513	-0.071774	-0.302768	0.918992	0.979728	0.907186	1
median_income	-0.015550	-0.079626	-0.118278	0.197882	-0.007723	0.005087	0
median_house_value	-0.045398	-0.144638	0.106432	0.133294	0.049686	-0.025300	0

In [15]:
```
#heatmat from seaborn view better visualization of correlation, input params is cor
relattion data

fig, ax = plt.subplots(figsize=(12,10))
sns.heatmap(housing_data_corr, annot=True)
```

Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0x1c454783c48>



In [ ]: