# Data Analysis

¶

In [1]:

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```python
data= pd.read_csv("D:\A.S\Working\Material\Machinfy\Sessions\Session 10\Assignments\hou
data.head()
```

Out[2]:

| | longitude | latitude | housing_median_age | total_rooms | total_bedrooms | population | householc |
|---|---|---|---|---|---|---|---|
| **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 [3]:

```
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 [4]:

```
data.describe()
```

Out[4]:

| | longitude | latitude | housing_median_age | total_rooms | total_bedrooms | popu |
|---|---|---|---|---|---|---|
| **count** | 20640.000000 | 20640.000000 | 20382.000000 | 20640.000000 | 15758.000000 | 20596.0 |
| **mean** | -119.569704 | 35.631861 | 28.676283 | 2635.763081 | 539.920104 | 1424.9 |
| **std** | 2.003532 | 2.135952 | 12.589284 | 2181.615252 | 419.834171 | 1132.2 |
| **min** | -124.350000 | 32.540000 | 1.000000 | 2.000000 | 1.000000 | 3.0 |
| **25%** | -121.800000 | 33.930000 | 18.000000 | 1447.750000 | 296.000000 | 787.0 |
| **50%** | -118.490000 | 34.260000 | 29.000000 | 2127.000000 | 435.000000 | 1166.0 |
| **75%** | -118.010000 | 37.710000 | 37.000000 | 3148.000000 | 652.000000 | 1725.0 |
| **max** | -114.310000 | 41.950000 | 52.000000 | 39320.000000 | 6210.000000 | 35682.0 |

In [5]:

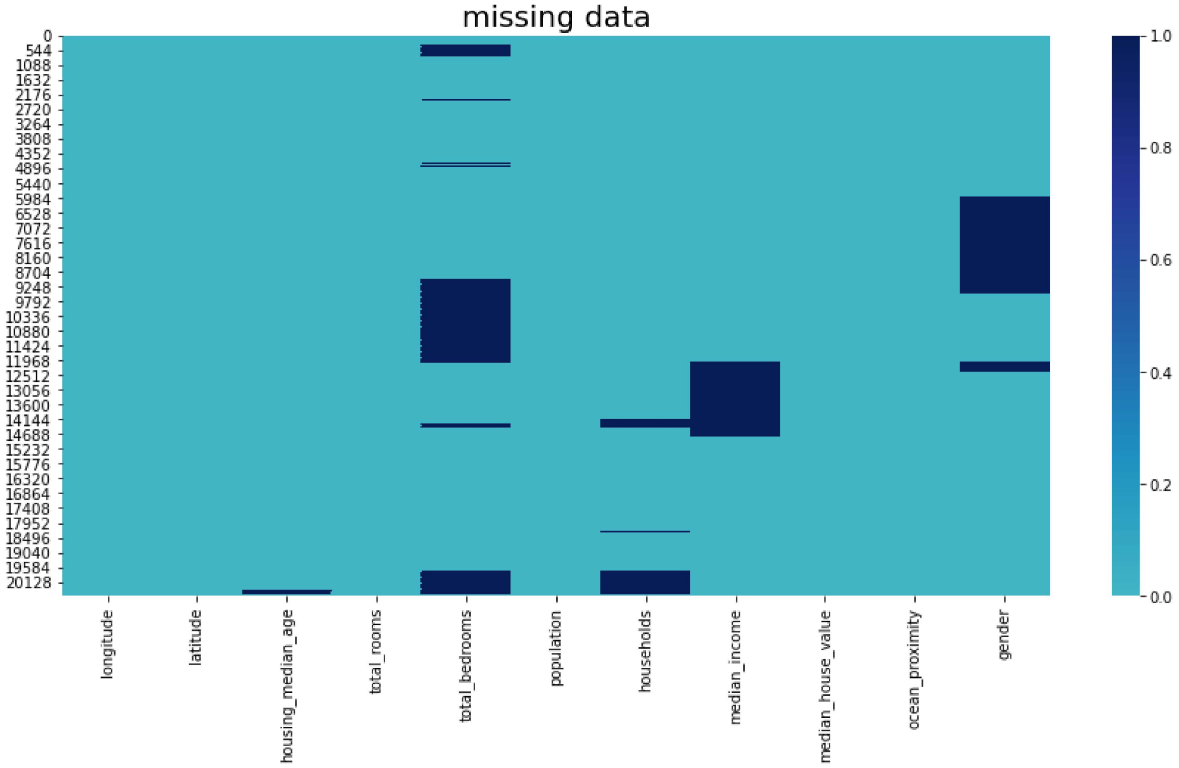
```
data.isnull().sum()
```

Out[5]:

```
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 [4]:

```
plt.figure(figsize=(15,7))
sns.heatmap(data.isnull(),cmap='YlGnBu',center=0)
font1={'size':20}
plt.title('missing data',fontdict=font1)
plt.show()
```



In [11]:

```
data["ocean_proximity"].value_counts()
```

Out[11]:

```
0    9136
1    6551
2    2658
3    2290
4       5
Name: ocean_proximity, dtype: int64
```

In [14]:

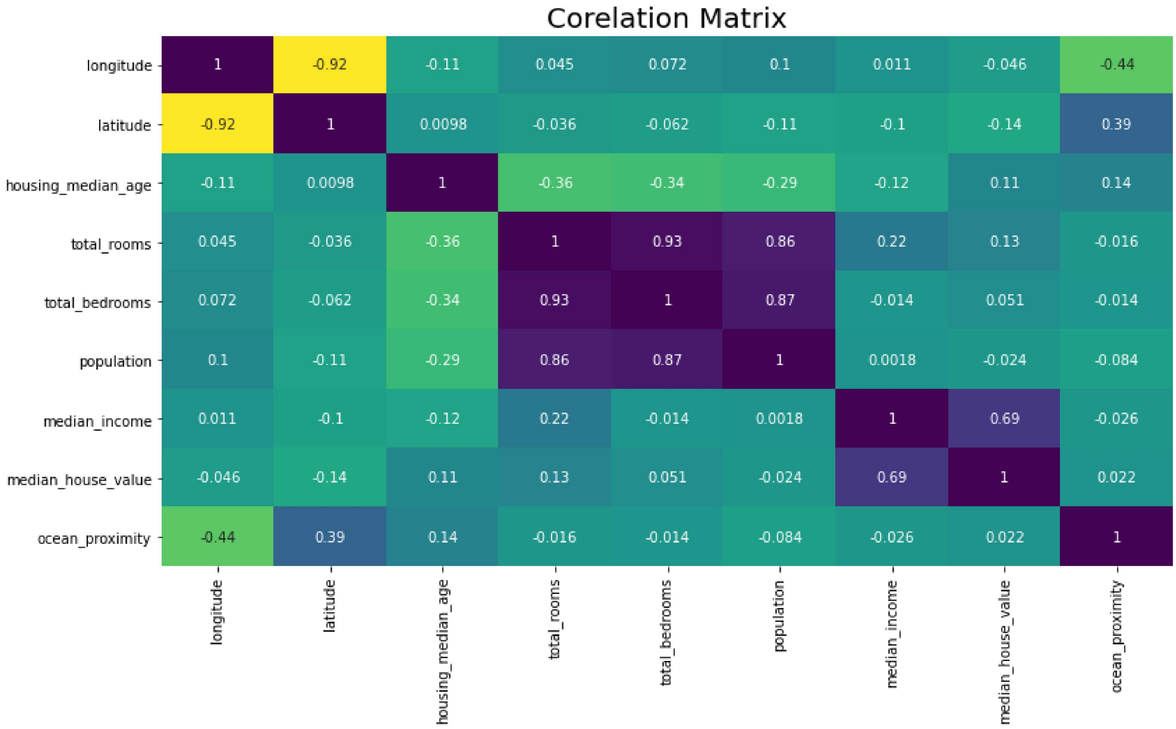
```
data["ocean_proximity"].replace('<1H OCEAN',0,inplace=True)
data["ocean_proximity"].replace('INLAND',1,inplace=True)
data["ocean_proximity"].replace('NEAR OCEAN',2,inplace=True)
data["ocean_proximity"].replace('NEAR BAY',3,inplace=True)
data["ocean_proximity"].replace('ISLAND',4,inplace=True)
data.head()
```

Out[14]:

| | longitude | latitude | housing_median_age | total_rooms | total_bedrooms | population | househol |
|---|---|---|---|---|---|---|---|
| **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 [16]:

```
plt.figure(figsize=(13,7))
sns.heatmap(cbar=False,annot=True,data=data.corr(),cmap='viridis_r')
plt.title('Corelation Matrix',fontdict=font1)
plt.show()
```



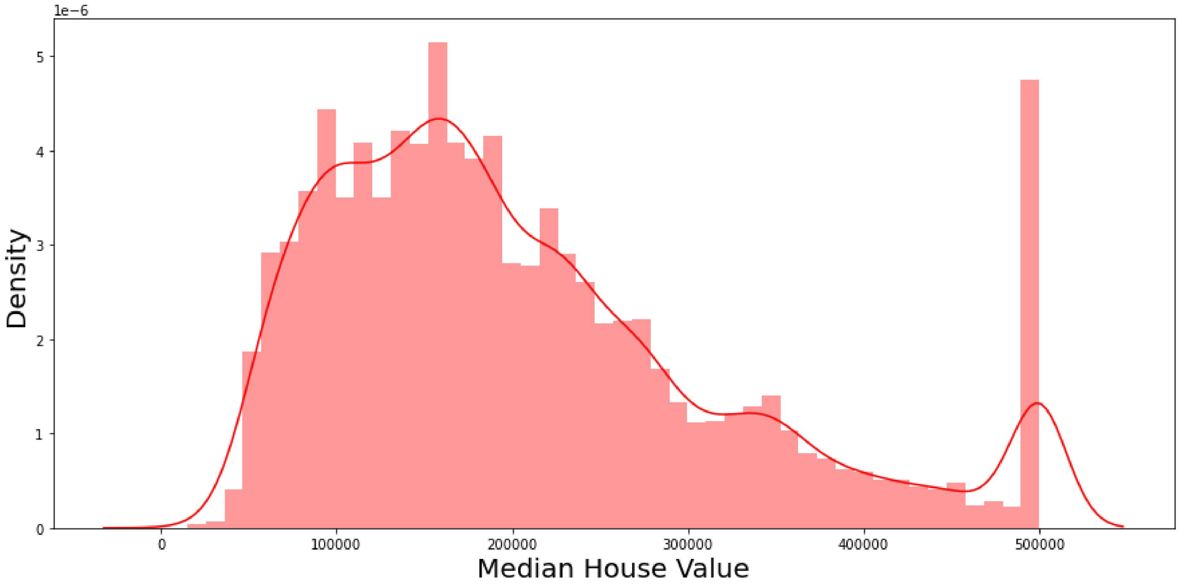
Corelation Matrix

| | longitude | latitude | housing_median_age | total_rooms | total_bedrooms | population | median_income | median_house_value | ocean_proximity |
|---|---|---|---|---|---|---|---|---|---|
| longitude | 1 | -0.92 | -0.11 | 0.045 | 0.072 | 0.1 | 0.011 | -0.046 | -0.44 |
| latitude | -0.92 | 1 | 0.0098 | -0.036 | -0.062 | -0.11 | -0.1 | -0.14 | 0.39 |
| housing_median_age | -0.11 | 0.0098 | 1 | -0.36 | -0.34 | -0.29 | -0.12 | 0.11 | 0.14 |
| total_rooms | 0.045 | -0.036 | -0.36 | 1 | 0.93 | 0.86 | 0.22 | 0.13 | -0.016 |
| total_bedrooms | 0.072 | -0.062 | -0.34 | 0.93 | 1 | 0.87 | -0.014 | 0.051 | -0.014 |
| population | 0.1 | -0.11 | -0.29 | 0.86 | 0.87 | 1 | 0.0018 | -0.024 | -0.084 |
| median_income | 0.011 | -0.1 | -0.12 | 0.22 | -0.014 | 0.0018 | 1 | 0.69 | -0.026 |
| median_house_value | -0.046 | -0.14 | 0.11 | 0.13 | 0.051 | -0.024 | 0.69 | 1 | 0.022 |
| ocean_proximity | -0.44 | 0.39 | 0.14 | -0.016 | -0.014 | -0.084 | -0.026 | 0.022 | 1 |

In [9]:

```
plt.figure(figsize=(15,7))
sns.distplot(data["median_house_value"],color='r')
plt.xlabel("Median House Value",fontdict=font1)
plt.ylabel("Density",fontdict=font1)
plt.show()
```

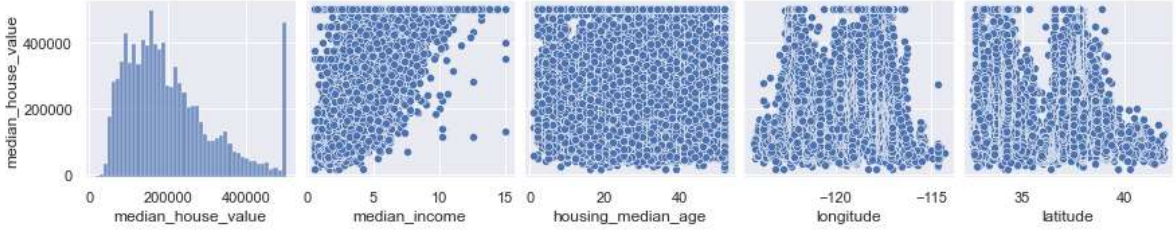
```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\distributions.py:2551: Fu
tureWarning: `distplot` is a deprecated function and will be removed in a fu
ture version. Please adapt your code to use either `displot` (a figure-level
function with similar flexibility) or `histplot` (an axes-level function for
histograms).
  warnings.warn(msg, FutureWarning)
```



In [119]:

```
sns.pairplot(data,y_vars=["median_house_value"],x_vars=["median_house_value","median_i
plt.show()
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\axisgrid.py:1912: UserWar
ning: The `size` parameter has been renamed to `height`; please update your
code.
  warnings.warn(msg, UserWarning)
```

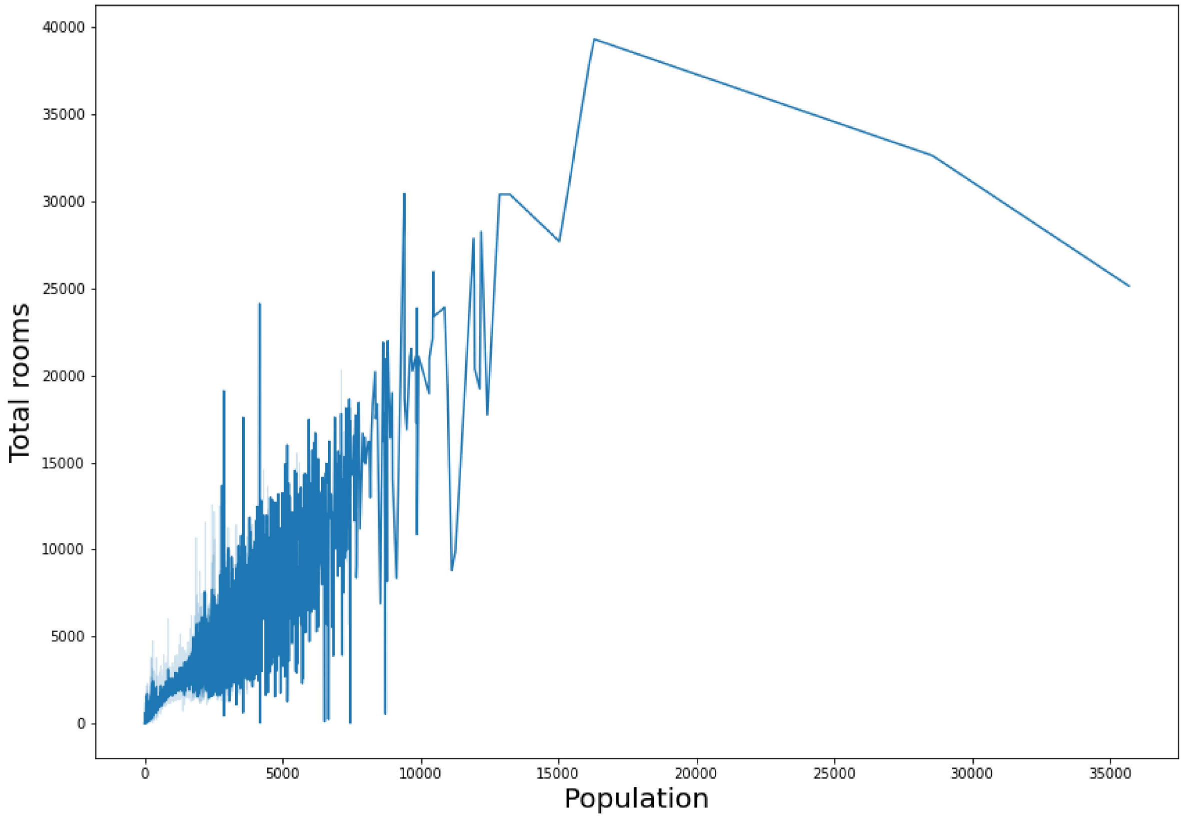


# population & total rooms

In [19]:

```
plt.figure(figsize=(14,10))
x4=data['population'].fillna(data['population'].mode())
y4=data["total_rooms"]
sns.lineplot(x4,y4,palette="cividis_r")
plt.xlabel('Population',fontdict=font1)
plt.ylabel('Total rooms',fontdict=font1)
plt.show()
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\_decorators.py:36: Future
Warning: Pass the following variables as keyword args: x, y. From version 0.
12, the only valid positional argument will be `data`, and passing other arg
uments without an explicit keyword will result in an error or misinterpretat
ion.
  warnings.warn(
```

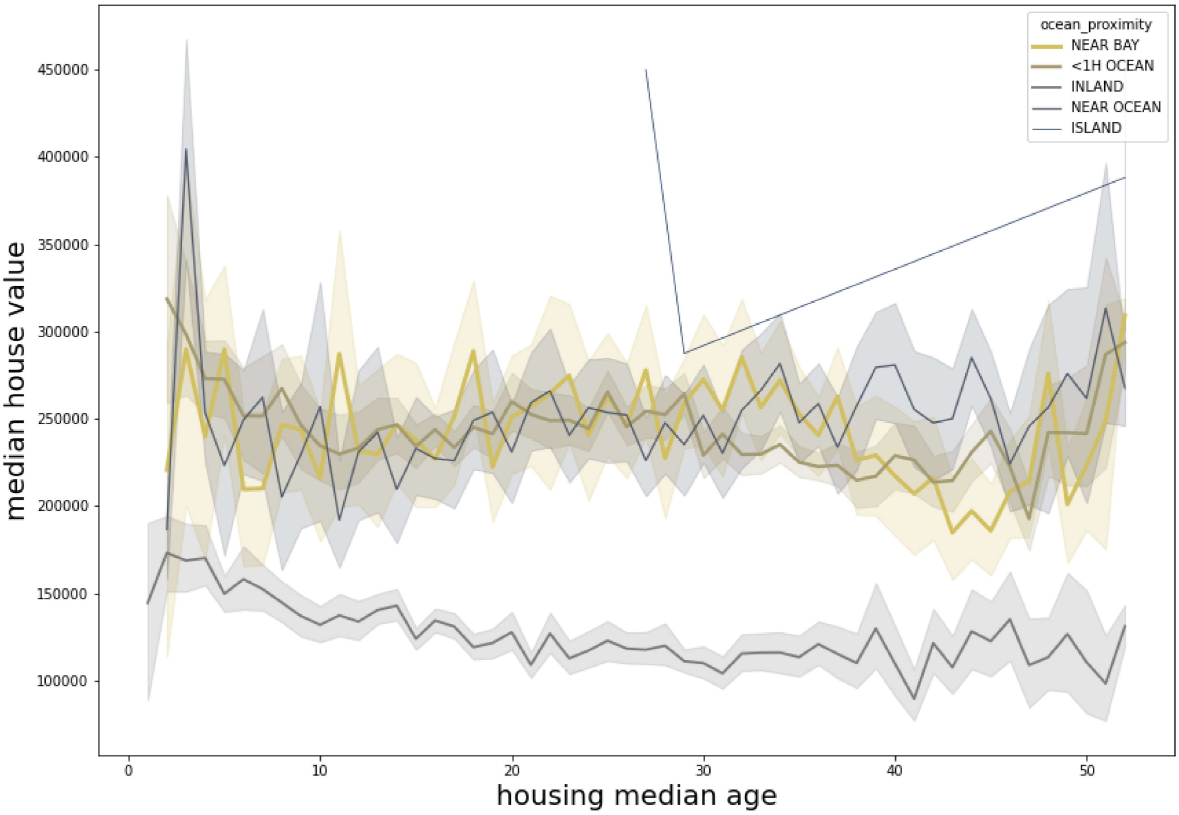


# housing median age & median house value

In [5]:

```
plt.figure(figsize=(14,10))
x3=data['housing_median_age'].fillna(data['housing_median_age'].median())
y3=data["median_house_value"]
sns.lineplot(x3,y3,hue='ocean_proximity',size='ocean_proximity',data=data,palette="civi
plt.xlabel('housing median age',fontdict=font1)
plt.ylabel('median house value',fontdict=font1)
plt.show()
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\_decorators.py:36: Future
Warning: Pass the following variables as keyword args: x, y. From version 0.
12, the only valid positional argument will be `data`, and passing other arg
uments without an explicit keyword will result in an error or misinterpretat
ion.
  warnings.warn(
```

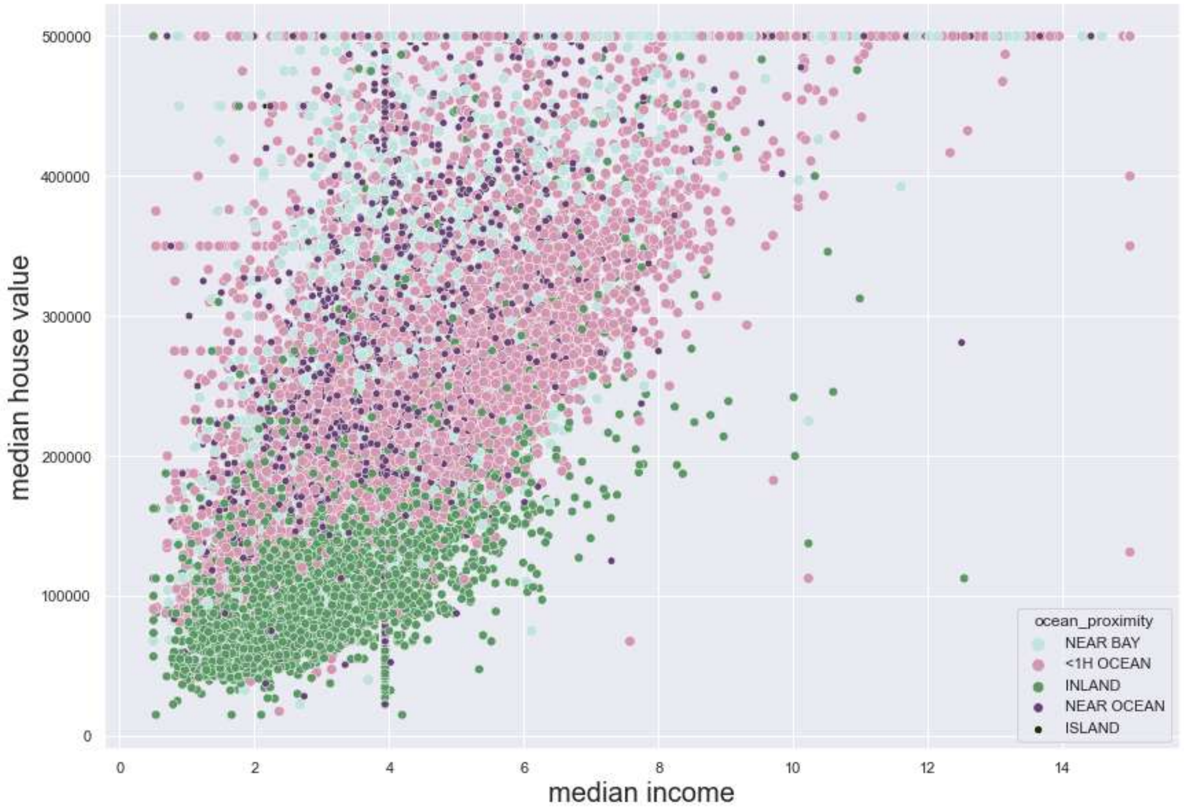


## median income & median house value

In [53]:

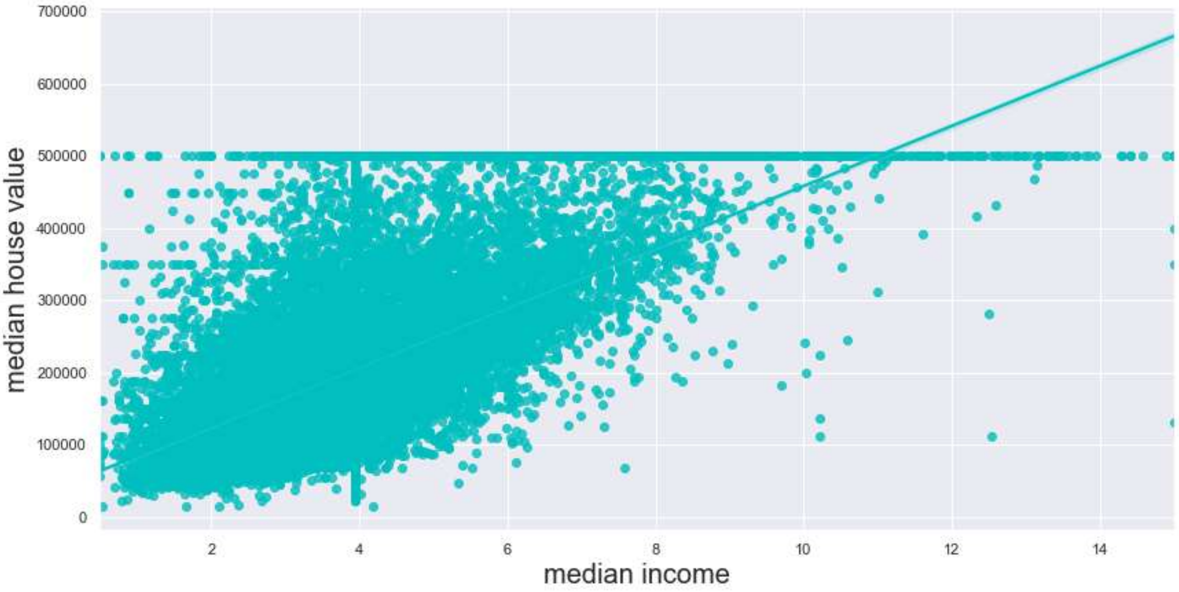
```
sns.set_theme(color_codes=False)
plt.figure(figsize=(14,10))
x2=data['median_income'].fillna(data['median_income'].mean())
y2=data["median_house_value"]
sns.scatterplot(x2,y2,hue='ocean_proximity',size='ocean_proximity',data=data,palette="
plt.xlabel('median income',fontdict=font1)
plt.ylabel('median house value',fontdict=font1)
plt.show()
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\_decorators.py:36: Future
Warning: Pass the following variables as keyword args: x, y. From version 0.
12, the only valid positional argument will be `data`, and passing other arg
uments without an explicit keyword will result in an error or misinterpretat
ion.
  warnings.warn(
```



In [23]:

```
sns.set_theme(color_codes=False)
plt.figure(figsize=(14,7))
x2=data['median_income'].fillna(data['median_income'].mean())
y2=data["median_house_value"]
s=sns.regplot(x2,y2,color="c")
plt.xlabel('median income',fontdict=font1)
plt.ylabel('median house value',fontdict=font1)
plt.show()
```


median house value
median income