Compte rendu TP N°1

Data preparation

Bedhief Safwa

3DNI 1

1. Chargement des données : "Caliornia Housing" dataset

In [32]:

In [33]:

```
housing's shape :  (20640, 10)
```

Out[33]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_valı
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200

2. Exploration et visualisation des données

In [34]:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  float64
 3   total_rooms         20640 non-null  float64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  float64
 6   households          20640 non-null  float64
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  float64
 9   ocean_proximity     20640 non-null  object 
dtypes: float64(9), object(1)
memory usage: 1.6+ MB
```

In [35]:

```
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 [36]:

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

Out[36]:

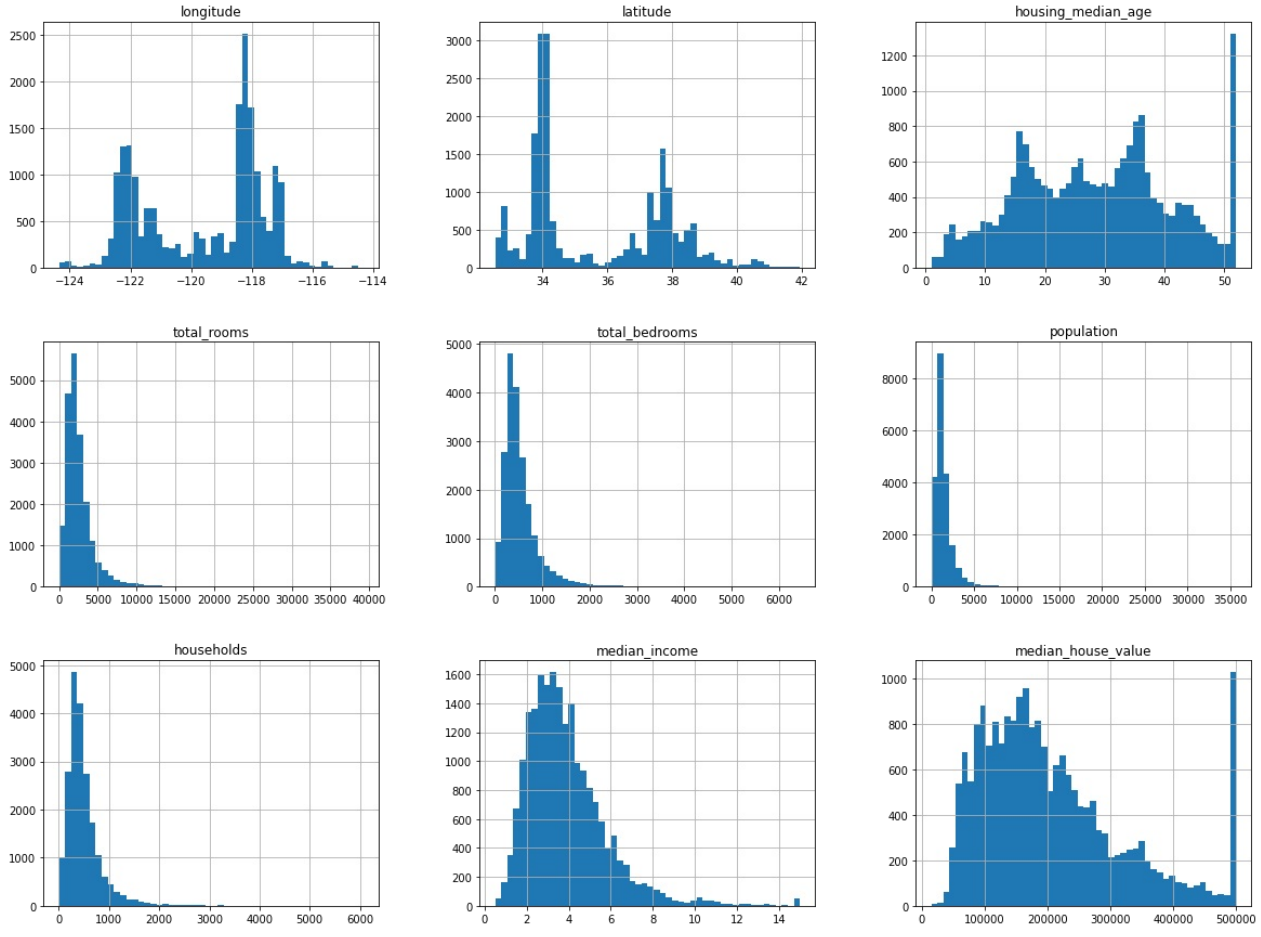
```
<1H OCEAN     9136
INLAND        6551
NEAR OCEAN    2658
NEAR BAY      2290
ISLAND           5
Name: ocean_proximity, dtype: int64
```

In [37]:

Out[37]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income
count	20640.000000	20640.000000	20640.000000	20640.000000	20433.000000	20640.000000	20640.000000	20640.000000
mean	-119.569704	35.631861	28.639486	2635.763081	537.870553	1425.476744	499.539680	3.870671
std	2.003532	2.135952	12.585558	2181.615252	421.385070	1132.462122	382.329753	1.899822
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	1.000000	0.499900
25%	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.000000	280.000000	2.563400
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.000000	409.000000	3.534800
75%	-118.010000	37.710000	37.000000	3148.000000	647.000000	1725.000000	605.000000	4.743250
max	-114.310000	41.950000	52.000000	39320.000000	6445.000000	35682.000000	6082.000000	15.000100

In [38]:



In [39]:

```
longitude            -0.297801
latitude              0.465953
housing_median_age    0.060331
total_rooms           4.147343
total_bedrooms        3.459546
population            4.935858
households            3.410438
median_income         1.646657
median_house_value    0.977763
dtype: float64
```

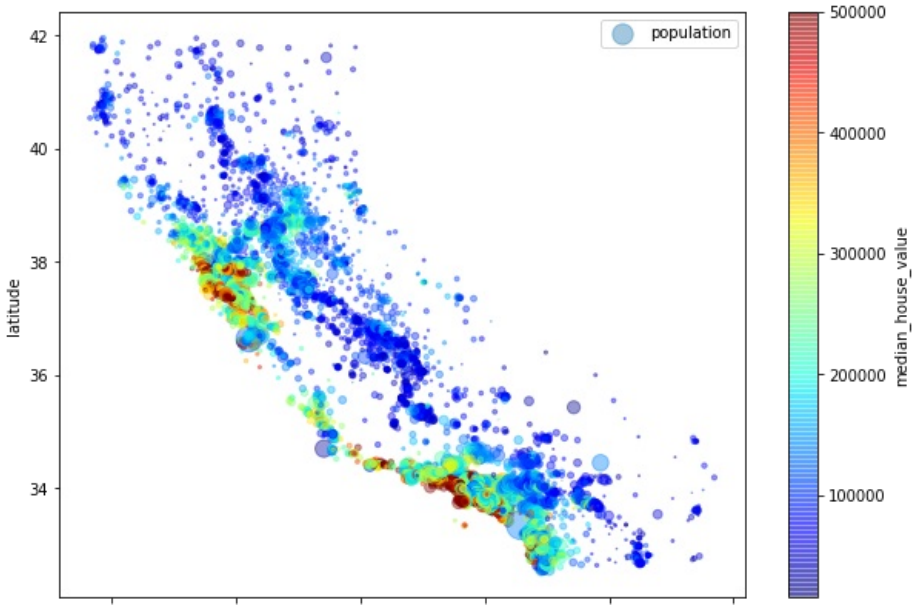
In [40]:

```
longitude            -1.330152
latitude             -1.117760
housing_median_age   -0.800629
total_rooms          32.630927
total_bedrooms       21.985575
population           73.553116
households           22.057988
median_income         4.952524
median_house_value    0.327870
dtype: float64
```

In [41]:

Out[41]:

```
<matplotlib.legend.Legend at 0x1f190c22e88>
```



In [42]:

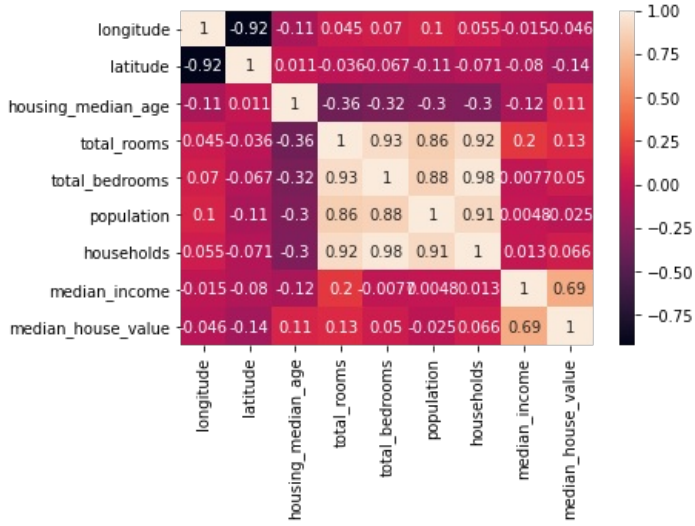
Out[42]:

	population	median_house_value
population	1.00000	-0.02465
median_house_value	-0.02465	1.00000

In [43]:

Out[43]:

<AxesSubplot:>

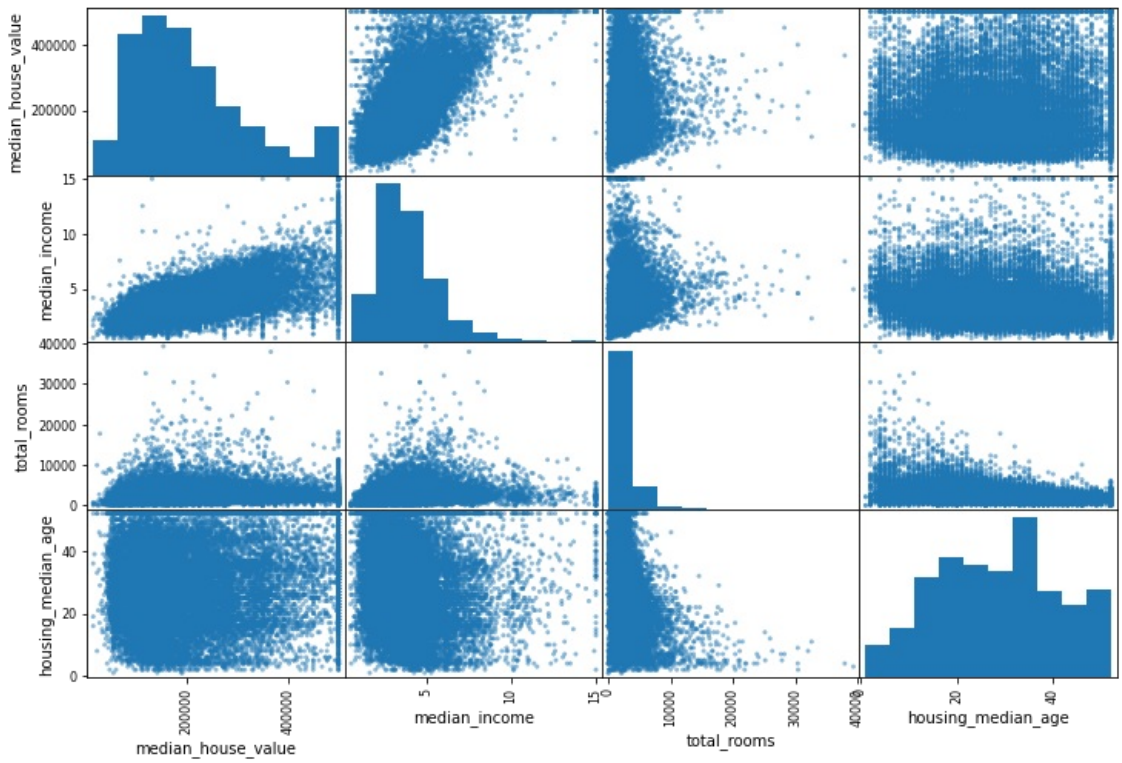


In [44]:

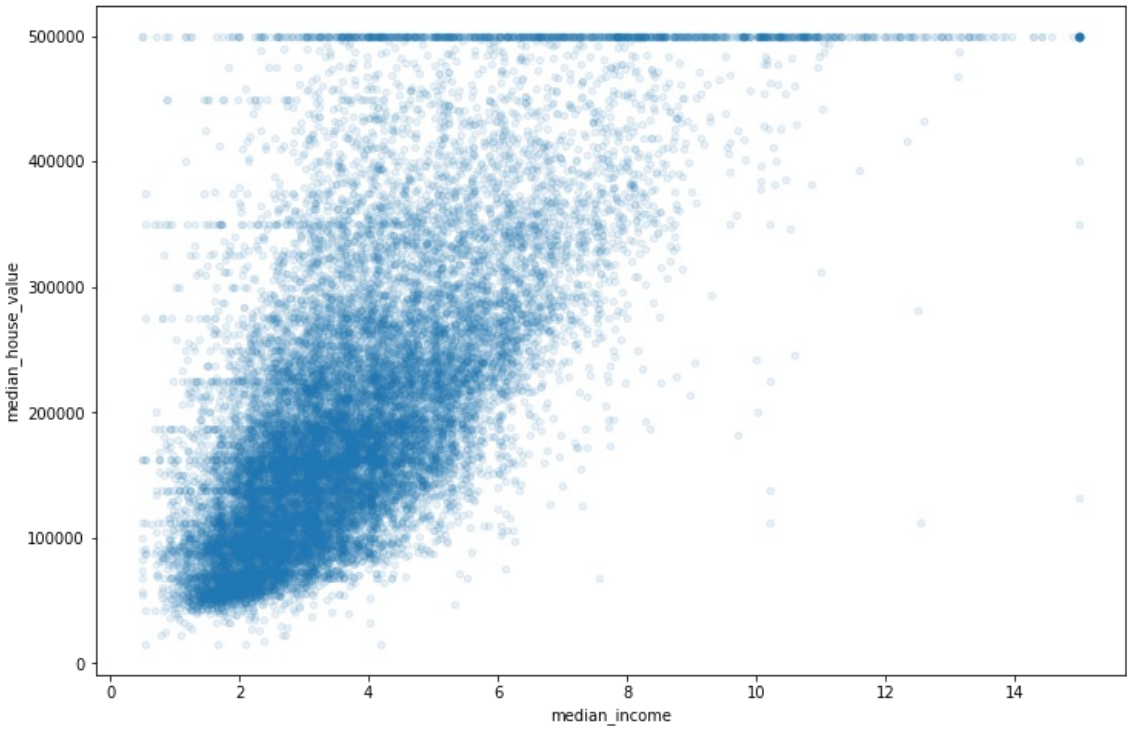
Out[44]:

```
median_house_value    1.000000
median_income         0.688075
total_rooms           0.134153
housing_median_age    0.105623
households            0.065843
total_bedrooms        0.049686
population           -0.024650
longitude            -0.045967
latitude             -0.144160
Name: median_house_value, dtype: float64
```

In [45]:



In [46]:



3. Préparation des données

In [47]:

Out[47]:

```
median_house_value          1.000000
median_income               0.688075
rooms_per_household         0.151948
total_rooms                 0.134153
housing_median_age          0.105623
households                  0.065843
population_per_household   -0.023737
population                 -0.024650
longitude                  -0.045967
latitude                   -0.144160
bedrooms_per_room          -0.255880
Name: median_house_value, dtype: float64
```

Division des données : Training & Testing Datatsets

Echontiollonnage aléatoire

80% données d'entrainement (training dataset).

20% doonées de test (testing dataset).

In [48]:

```
X_train: (16512, 11) X_test: (4128, 11)
```

In [49]:

Out[49]:

```
array([[ 7. ,  2. ,  3. ],
       [ 4. ,  3.5,  6. ],
       [10. ,  5. ,  9. ]])
```

In [50]:

```
[[7 9 5]
 [2 8 2]
 [6 9 6]]

mean [5.         8.66666667 4.33333333]
[[ 0.9258201   0.70710678  0.39223227]
 [-1.38873015 -1.41421356 -1.37281295]
 [ 0.46291005  0.70710678  0.98058068]]

Moy: 4.934324553889585e-16 Std: 1.0
```

In [51]:

```
[[0. 1. 0.]
 [0. 1. 0.]
 [1. 0. 0.]
 [0. 0. 1.]
 [1. 0. 0.]]
[array(['JAUNE', 'ROUGE', 'VERT'], dtype=object)]
```

In [52]:

```
Index(['longitude', 'latitude', 'housing_median_age', 'total_rooms',
       'population', 'households', 'median_income', 'rooms_per_household',
       'bedrooms_per_room', 'population_per_household'],
      dtype='object')
Index(['ocean_proximity'], dtype='object')
```

In [53]:

```
 X_train: (16512, 15) X_test: (4128, 15)
```

In [23]:

```
[-1.30857243  1.05876908  0.82883368 -1.18819343 -1.40850033 -0.8230231
 -0.24550082 -1.26183905  1.32210962 -2.15552052  0.          0.
  0.          1.          0.        ]
['longitude' 'latitude' 'housing_median_age' 'total_rooms' 'population'
 'households' 'median_income' 'rooms_per_household' 'bedrooms_per_room'
 'population_per_household' 'ocean_1' 'ocean_2' 'ocean_3' 'ocean_4'
 'ocean_5' 'median_house_value']
```

In [24]:

```
(16512, 15)
(16512,)
```

Out[24]:

	longitude	latitude	housing_median_age	total_rooms	population	households	median_income	rooms_per_household	bedrooms
0	-1.308572	1.058769	0.828834	-1.188193	-1.408500	-0.823023	-0.245501	-1.261839	
1	0.567871	-0.635401	1.048089	0.058527	-0.136146	0.127886	-0.193363	-0.146672	
2	1.382643	-1.582079	-0.736547	1.393413	1.388602	1.565610	-0.393989	-0.216869	
3	-1.185122	0.890334	-0.243567	0.579701	0.367176	0.547581	0.833637	0.152921	
4	-0.172830	0.660399	1.406965	-0.671689	-0.045317	-0.356069	-1.907723	-0.961184	

In [25]:

```
(4128, 15)
(4128,)
```

Out[25]:

	longitude	latitude	housing_median_age	total_rooms	population	households	median_income	rooms_per_household	bedrooms
0	-0.172830	0.684426	-0.736547	-0.482004	-0.302572	-0.120655	-1.726629	-1.045449	
1	-0.123450	-0.488583	-0.084457	0.830651	0.568531	0.908358	0.315593	-0.080205	
2	0.222210	0.035660	-0.084457	1.146837	0.953889	1.206860	-0.211090	-0.006154	
3	0.617251	-0.591933	1.120496	-0.566029	-0.487707	-0.459440	0.586153	-0.324347	
4	0.543181	-0.756144	-0.993750	-3.363673	-3.809589	-3.937811	1.732088	3.803353	

In [54]:

Out[54]:

```
['Data_prepration.pkl']
```

In [ ]: