Housing dataset

Dataset Description

1. longitude: A measure of how far west a house is; a higher value is farther west
2. latitude: A measure of how far north a house is; a higher value is farther north
3. housingMedianAge: Median age of a house within a block; a lower number is a newer building
4. totalRooms: Total number of rooms within a block
5. totalBedrooms: Total number of bedrooms within a block
6. population: Total number of people residing within a block
7. households: Total number of households, a group of people residing within a home unit, for a block
8. medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
9. medianHouseValue: Median house value for households within a block (measured in US Dollars)
10. oceanProximity: Location of the house w.r.t ocean/sea

Import Libraries

In [1]:
```
import pandas as pd
```

In [2]:
```
df=pd.read_csv('housing.csv')
```

In [3]:
```
df.head()
```

Out[3]:

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

In [4]:
```
df.info()
```

```
<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 [5]:
```
df.describe()
```

Out[5]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	
count	20640.000000	20640.000000	20640.000000	20640.000000	20433.000000	20640.000000	20
mean	-119.569704	35.631861	28.639486	2635.763081	537.870553	1425.476744	
std	2.003532	2.135952	12.585558	2181.615252	421.385070	1132.462122	
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	
25%	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.000000	
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.000000	
75%	-118.010000	37.710000	37.000000	3148.000000	647.000000	1725.000000	
max	-114.310000	41.950000	52.000000	39320.000000	6445.000000	35682.000000	6

In [6]:
```
df.isnull().sum()
```

Out[6]:
```
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 [7]:
```
df.dropna(inplace=True)
```

In [8]:
```
df.isnull().sum()
```

Out[8]:
```
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 [9]:
```
df.duplicated().sum()
```

Out[9]: 0

to remove duplicates df.drop_duplicates(inplace=True)

In [10]:
```
df.drop(['households'],axis=1,inplace=True) #axis=1 to drop column ---- axis=0 to drop
```

In [11]:
```
df.head(10)
```

Out[11]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	median_income	r
0	-122.23	37.88	41.0	880.0	129.0	322.0	8.3252	
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	8.3014	
2	-122.24	37.85	52.0	1467.0	190.0	496.0	7.2574	
3	-122.25	37.85	52.0	1274.0	235.0	558.0	5.6431	
4	-122.25	37.85	52.0	1627.0	280.0	565.0	3.8462	
5	-122.25	37.85	52.0	919.0	213.0	413.0	4.0368	
6	-122.25	37.84	52.0	2535.0	489.0	1094.0	3.6591	
7	-122.25	37.84	52.0	3104.0	687.0	1157.0	3.1200	
8	-122.26	37.84	42.0	2555.0	665.0	1206.0	2.0804	
9	-122.25	37.84	52.0	3549.0	707.0	1551.0	3.6912	

In [12]:
```
df.sort_values("median_house_value",inplace=True)
```

In [13]:
```
df
```

Out[13]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	median_incon
9188	-117.86	34.24	52.0	803.0	267.0	628.0	4.19
2521	-122.74	39.71	16.0	255.0	73.0	85.0	1.66
2799	-117.02	36.40	19.0	619.0	239.0	490.0	2.10
19802	-123.17	40.31	36.0	98.0	28.0	18.0	0.53
5887	-118.33	34.15	39.0	493.0	168.0	259.0	2.36

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	median_incon
...	...	...	...	...	...	...	
8268	-118.17	33.74	36.0	2006.0	453.0	807.0	3.78
18046	-122.00	37.23	36.0	3191.0	430.0	1234.0	9.07
8189	-118.13	33.78	31.0	3039.0	739.0	1199.0	3.72
8304	-118.12	33.75	47.0	3330.0	569.0	1220.0	7.36
16137	-122.49	37.79	52.0	2488.0	281.0	805.0	10.70

20433 rows × 9 columns

In [ ]: