```
import pandas as pd
import numpy as np
from io import StringIO


from google.colab import files
uploaded = files.upload()
```

```
Choose Files california h… dataset.csv
• california housing dataset.csv(text/csv) - 1444170 bytes, last modified: 5/10/2023 - 100% done
Saving california housing dataset.csv to california housing dataset.csv
```

```
import io
df_21BAI1380 = pd.read_csv(io.BytesIO(uploaded['california housing dataset.csv']))
print(df_21BAI1380)
```

```
       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
...          ...       ...                 ...          ...             ...
20635    -121.09     39.48                25.0       1665.0           374.0
20636    -121.21     39.49                18.0        697.0           150.0
20637    -121.22     39.43                17.0       2254.0           485.0
20638    -121.32     39.43                18.0       1860.0           409.0
20639    -121.24     39.37                16.0       2785.0           616.0

       population  households  median_income  median_house_value  \
0           322.0       126.0         8.3252            452600.0
1          2401.0      1138.0         8.3014            358500.0
2           496.0       177.0         7.2574            352100.0
3           558.0       219.0         5.6431            341300.0
4           565.0       259.0         3.8462            342200.0
...           ...         ...            ...                 ...
20635       845.0       330.0         1.5603             78100.0
20636       356.0       114.0         2.5568             77100.0
20637      1007.0       433.0         1.7000             92300.0
20638       741.0       349.0         1.8672             84700.0
20639      1387.0       530.0         2.3886             89400.0

      ocean_proximity
0            NEAR BAY
1            NEAR BAY
2            NEAR BAY
3            NEAR BAY
4            NEAR BAY
...               ...
20635          INLAND
20636          INLAND
20637          INLAND
20638          INLAND
20639          INLAND

[20640 rows x 10 columns]
```

```
print(df_21BAI1380.shape)
```

```
(20640, 10)
```

```
print(type(df_21BAI1380))
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
df_21BAI1380.head(3)
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	populatio
0	-122.23	37.88	41.0	880.0	129.0	322.
1	-122.22	37.86	21.0	7099.0	1106.0	2401.
2	-122.24	37.85	52.0	1467.0	190.0	496.

```
df_21BAI1380.keys()
```

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

```
print(len(df_21BAI1380))
```

```
20640
```

```
df_21BAI1380.columns
```

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

```
df_21BAI1380.items()
```

```
<generator object DataFrame.items at 0x7f9068356260>
```

```
df_21BAI1380.describe()
```

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

```
len(df_21BAI1380.columns)
```

```
10
```

```
df_21BAI1380.size
```

```
206400
```

```
len(df_21BAI1380.index)
```

```
20640
```

```
df_21BAI1380.isnull()
```

```
             longitude  latitude  housing median age  total rooms  total bedrooms  popul
df_21BAI1380.isnull().sum(axis=0)
```

```
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
```

```
 20636      False     False                False         False           False
```

```
df_21BAI1380.isnull().sum(axis=1)
```

```
0        0
1        0
2        0
3        0
4        0
        ..
20635    0
20636    0
20637    0
20638    0
20639    0
Length: 20640, dtype: int64
```

```
import numpy as np
arr = np.eye(5)
print(arr)
```

```
[[1. 0. 0. 0. 0.]
 [0. 1. 0. 0. 0.]
 [0. 0. 1. 0. 0.]
 [0. 0. 0. 1. 0.]
 [0. 0. 0. 0. 1.]]
```

```
from scipy.sparse import csr_matrix

sparse_arr = csr_matrix(arr)
print(sparse_arr)
```

```
  (0, 0)        1.0
  (1, 1)        1.0
  (2, 2)        1.0
  (3, 3)        1.0
  (4, 4)        1.0
```

```
np.eye(3)
```

```
array([[1., 0., 0.],
       [0., 1., 0.],
       [0., 0., 1.]])
```

```
print(df_21BAI1380.describe())
```

```
          longitude      latitude  housing_median_age   total_rooms  \
count  20640.000000  20640.000000        20640.000000  20640.000000
mean    -119.569704     35.631861           28.639486   2635.763081
std        2.003532      2.135952           12.585558   2181.615252
min     -124.350000     32.540000            1.000000      2.000000
25%     -121.800000     33.930000           18.000000   1447.750000
50%     -118.490000     34.260000           29.000000   2127.000000
75%     -118.010000     37.710000           37.000000   3148.000000
max     -114.310000     41.950000           52.000000  39320.000000

       total_bedrooms    population    households  median_income  \
count    20433.000000  20640.000000  20640.000000   20640.000000
mean       537.870553   1425.476744    499.539680       3.870671
std        421.385070   1132.462122    382.329753       1.899822
min          1.000000      3.000000      1.000000       0.499900
25%        296.000000    787.000000    280.000000       2.563400
50%        435.000000   1166.000000    409.000000       3.534800
75%        647.000000   1725.000000    605.000000       4.743250
max       6445.000000  35682.000000   6082.000000      15.000100
```

```
       median_house_value
count        20640.000000
mean        206855.816909
std         115395.615874
min          14999.000000
25%         119600.000000
50%         179700.000000
75%         264725.000000
max         500001.000000
```

```
df_21BAI1380["population"].mean()
```

```
1425.4767441860465
```

```
df_21BAI1380["population"].median()
```

```
1166.0
```

```
df_21BAI1380["population"].mode()
```

```
0    891.0
Name: population, dtype: float64
```

```
df_21BAI1380.ocean_proximity.unique()
```

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

```
print(df_21BAI1380['ocean_proximity'].value_counts())
```

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

```
mean_val=df_21BAI1380['median_income'].mean()
df_21BAI1380['category_rooms']=df_21BAI1380['median_income'].apply(lambda x:"can afford" if x<mean_val else "cannot afford")
```

```
df_21BAI1380['category_rooms']
```

```
0        cannot afford
1        cannot afford
2        cannot afford
3        cannot afford
4           can afford
              ...
20635       can afford
20636       can afford
20637       can afford
20638       can afford
20639       can afford
Name: category_rooms, Length: 20640, dtype: object
```

```
subset = df_21BAI1380.iloc[0:3, [1,2,3,4]]
print(subset)
```

```
   latitude  housing_median_age  total_rooms  total_bedrooms
0     37.88                41.0        880.0           129.0
1     37.86                21.0       7099.0          1106.0
2     37.85                52.0       1467.0           190.0
```

```
df_21BAI1380['Price'] = df_21BAI1380['median_house_value'] * 1000
print(df_21BAI1380.head())
```

```
   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  \
```

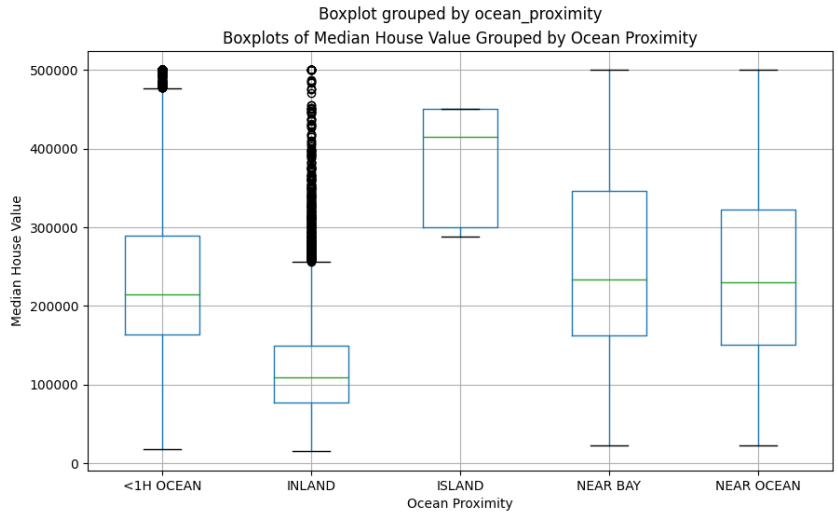
```
0       322.0       126.0         8.3252           452600.0        NEAR BAY
1      2401.0      1138.0         8.3014           358500.0        NEAR BAY
2       496.0       177.0         7.2574           352100.0        NEAR BAY
3       558.0       219.0         5.6431           341300.0        NEAR BAY
4       565.0       259.0         3.8462           342200.0        NEAR BAY

  category_rooms        Price
0  cannot afford  452600000.0
1  cannot afford  358500000.0
2  cannot afford  352100000.0
3  cannot afford  341300000.0
4     can afford  342200000.0
```

```
for col in df_21BAI1380.columns:
  if df_21BAI1380[col].dtype =='object':
    print(df_21BAI1380[col].unique())
```

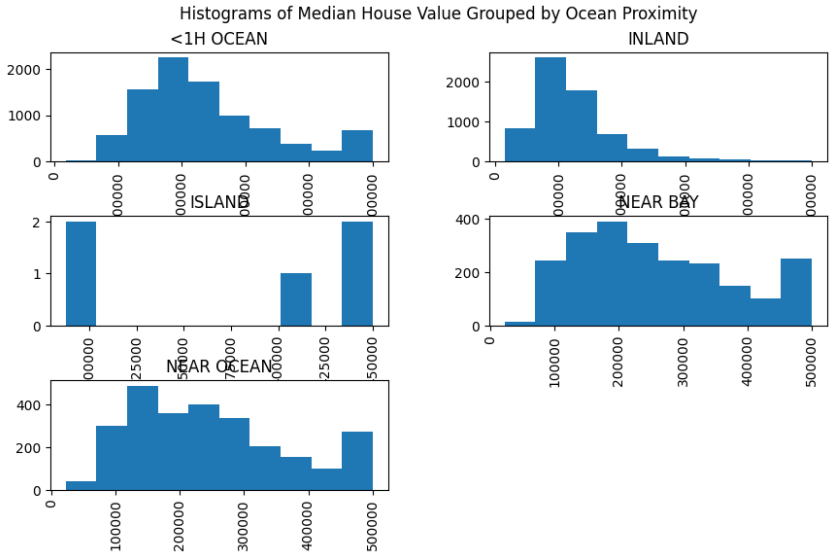
```
['NEAR BAY' '<1H OCEAN' 'INLAND' 'NEAR OCEAN' 'ISLAND']
['cannot afford' 'can afford']
```

```
import seaborn as sns
```

```
import pandas as pd
import matplotlib.pyplot as plt
df_21BAI1380.boxplot(column='median_house_value', by='ocean_proximity', figsize=(10, 6))
plt.title('Boxplots of Median House Value Grouped by Ocean Proximity')
plt.ylabel('Median House Value')
plt.xlabel('Ocean Proximity')
plt.show()
```



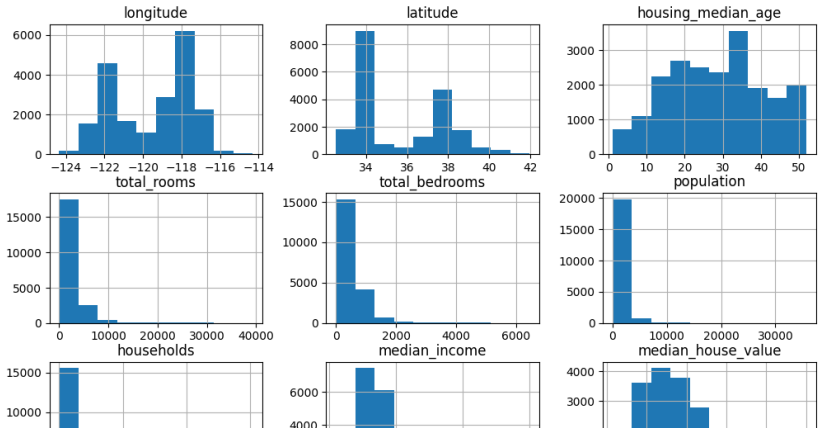

```
import pandas as pd
import matplotlib.pyplot as plt
df_21BAI1380.hist(column='median_house_value', by='ocean_proximity', figsize=(10, 6))
plt.suptitle('Histograms of Median House Value Grouped by Ocean Proximity')
plt.show()
```



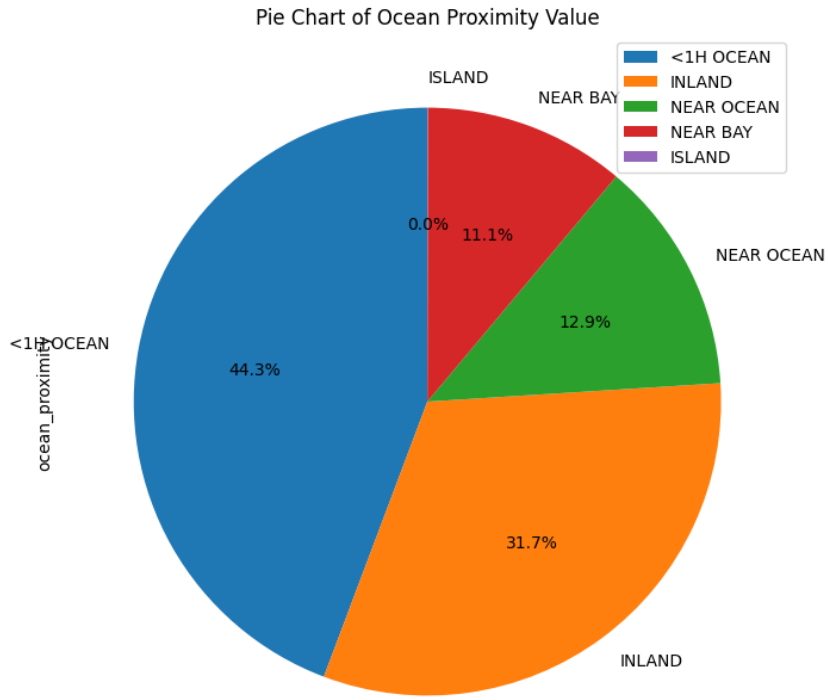

```
print(df_21BAI1380.tail())
```

```
       longitude  latitude  housing_median_age  total_rooms  total_bedrooms  \
20635    -121.09     39.48                25.0       1665.0           374.0
20636    -121.21     39.49                18.0        697.0           150.0
20637    -121.22     39.43                17.0       2254.0           485.0
20638    -121.32     39.43                18.0       1860.0           409.0
20639    -121.24     39.37                16.0       2785.0           616.0

       population  households  median_income  median_house_value  \
20635       845.0       330.0         1.5603             78100.0
20636       356.0       114.0         2.5568             77100.0
20637      1007.0       433.0         1.7000             92300.0
20638       741.0       349.0         1.8672             84700.0
20639      1387.0       530.0         2.3886             89400.0

      ocean_proximity category_rooms       Price
20635          INLAND     can afford  78100000.0
20636          INLAND     can afford  77100000.0
20637          INLAND     can afford  92300000.0
20638          INLAND     can afford  84700000.0
20639          INLAND     can afford  89400000.0
```

```
df_21BAI1380.hist(figsize=(12, 10))
plt.show()
```




```
ocean_proximity_count = df_21BAI1380['ocean_proximity'].value_counts()

ocean_proximity_count.plot(kind='pie', figsize=(8, 8), autopct='%1.1f%%', startangle=90)
plt.title('Pie Chart of Ocean Proximity Value')
plt.legend()
plt.show()
```

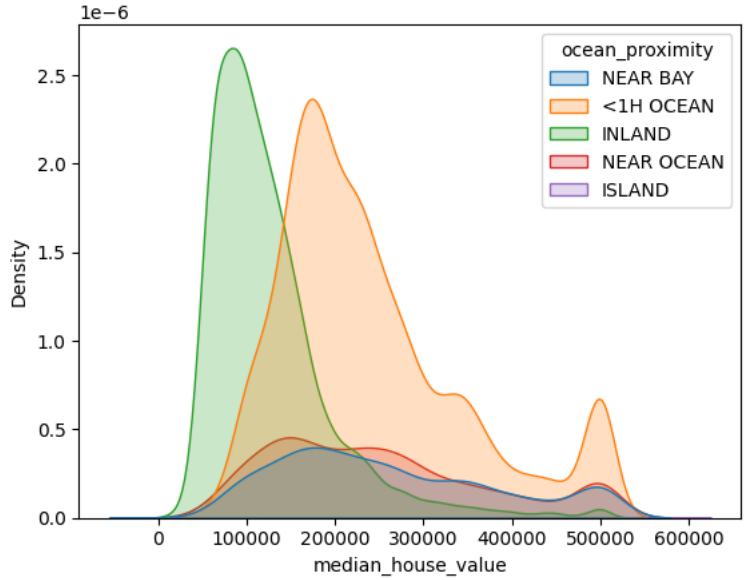



```
ocean_proximity_amount = df_21BAI1380.groupby('ocean_proximity')['median_house_value'].sum()
print(ocean_proximity_amount)
```

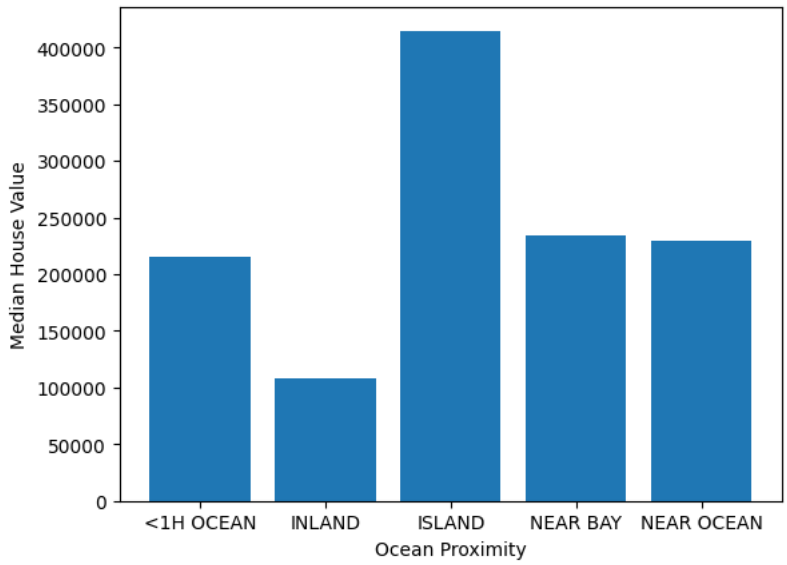
```
ocean_proximity
<1H OCEAN     2.193410e+09
INLAND        8.176001e+08
ISLAND        1.902200e+06
NEAR BAY      5.935962e+08
NEAR OCEAN    6.629955e+08
Name: median_house_value, dtype: float64
```

```
sns.kdeplot(data=df_21BAI1380, x='median_house_value', hue='ocean_proximity', fill=True)
```

```
<Axes: xlabel='median_house_value', ylabel='Density'>
```

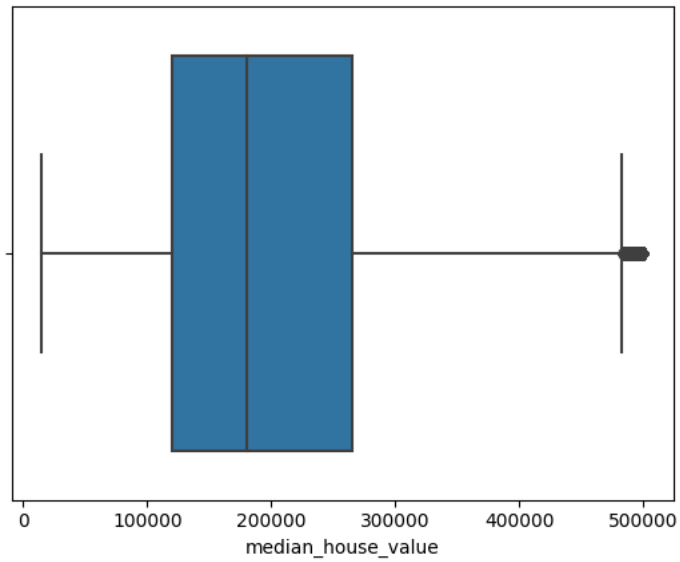



```
median_values = df_21BAI1380.groupby('ocean_proximity')['median_house_value'].median()

# Create a bar chart of median house value for each ocean proximity category
plt.bar(median_values.index, median_values.values)
plt.xlabel('Ocean Proximity')
plt.ylabel('Median House Value')
plt.show()
```




```
sns.boxplot(data=df_21BAI1380, x='median_house_value')
```

```
<Axes: xlabel='median_house_value'>
```



✓ 0s completed at 11:35 AM