DWFinal

March 8, 2024

[1]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
```

[2]:

```
url = r"C:\Users\MUBASHIR KHAN\Desktop\jupyter\Internship Project\California␣
  ↪Housing Prices Dataset.csv"
housing_data = pd.read_csv(url)
```

[3]:

```
numeric_columns = housing_data.select_dtypes(include=['number']).columns
numeric_data = housing_data[numeric_columns]
```

[4]:

```
print("Shape of the dataset:", housing_data.shape)
print("\nInfo about the dataset:")
print(housing_data.info())
print("\nSummary statistics of numerical features:")
print(housing_data.describe())
print("\nUnique values in 'ocean_proximity' column:")
print(housing_data['ocean_proximity'].unique())
```

```
Shape of the dataset: (20640, 10)

Info about the dataset:
<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
None

Summary statistics of numerical features:
          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

Unique values in 'ocean_proximity' column:
['NEAR BAY' '<1H OCEAN' 'INLAND' 'NEAR OCEAN' 'ISLAND']
```

[5]:

```python
print("\nCheck for missing values:")
print(housing_data.isnull().sum())
```

```
Check for missing values:
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
```

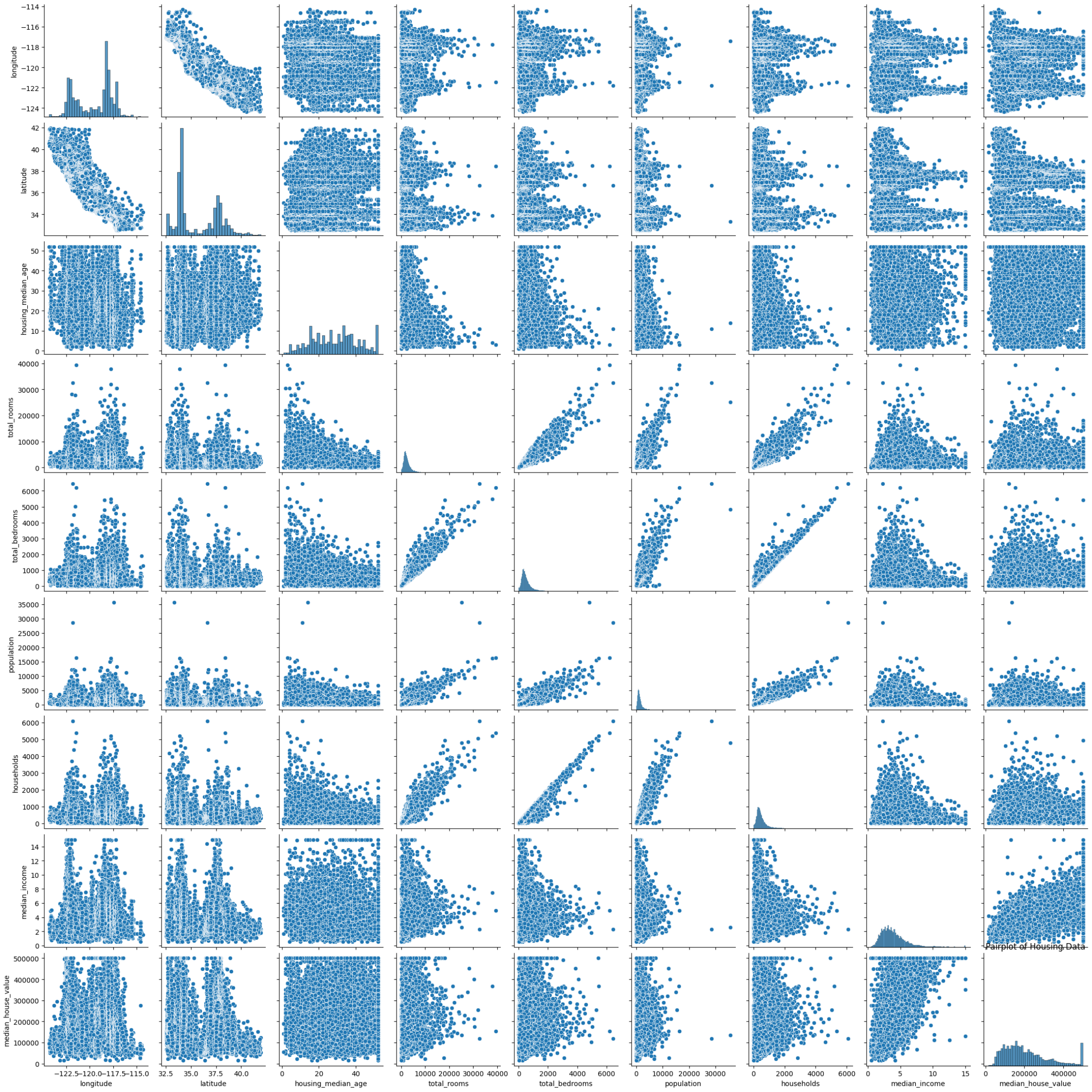
[6]:

```python
housing_data.dropna(inplace=True)
```

[9]:

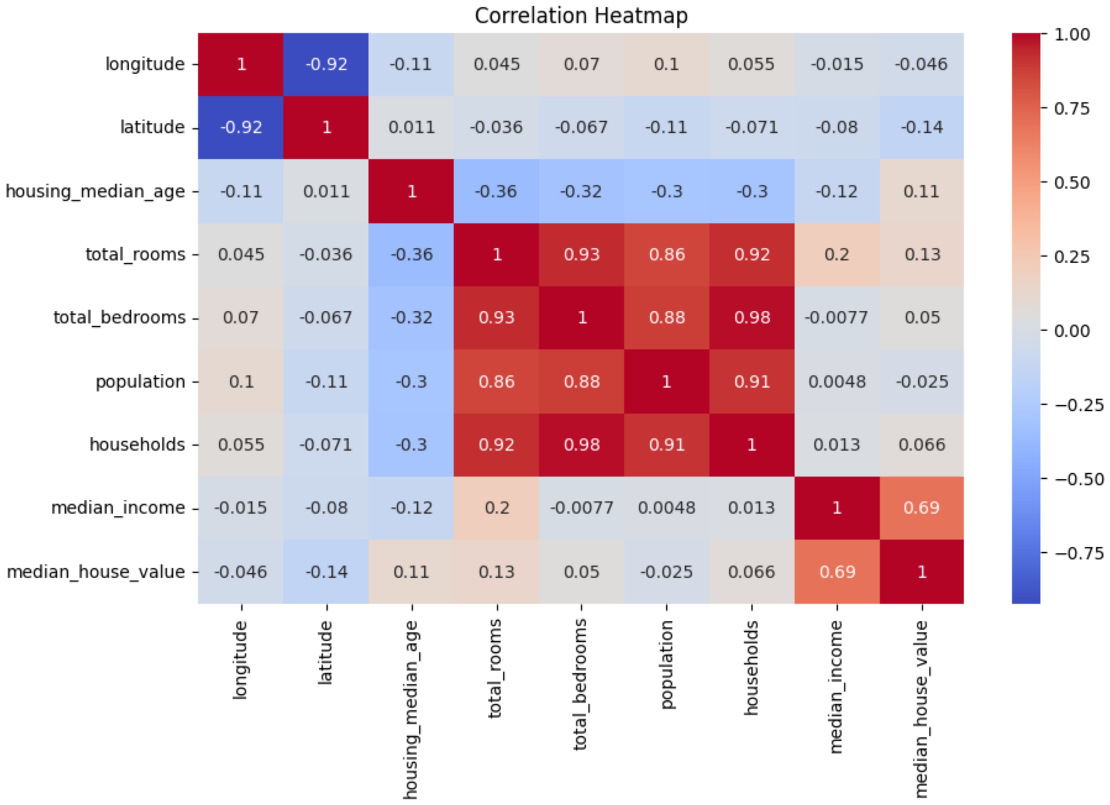
```python
plt.figure(figsize=(10, 6))
sns.pairplot(housing_data)
plt.title('Pairplot of Housing Data')
plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
```



[7]:

```python
plt.figure(figsize=(10, 6))
sns.heatmap(numeric_data.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```



[9]:

```python
california_data = housing_data[housing_data['ocean_proximity'] == 'NEAR OCEAN']
X = california_data[['population']]  # independent variable
y = california_data['median_house_value']  # dependent variable
```

[23]:

```python
model = LinearRegression()
model.fit(X, y)
```

[23]:

```
LinearRegression()
```

[24]:

```python
y_pred = model.predict(X)
```

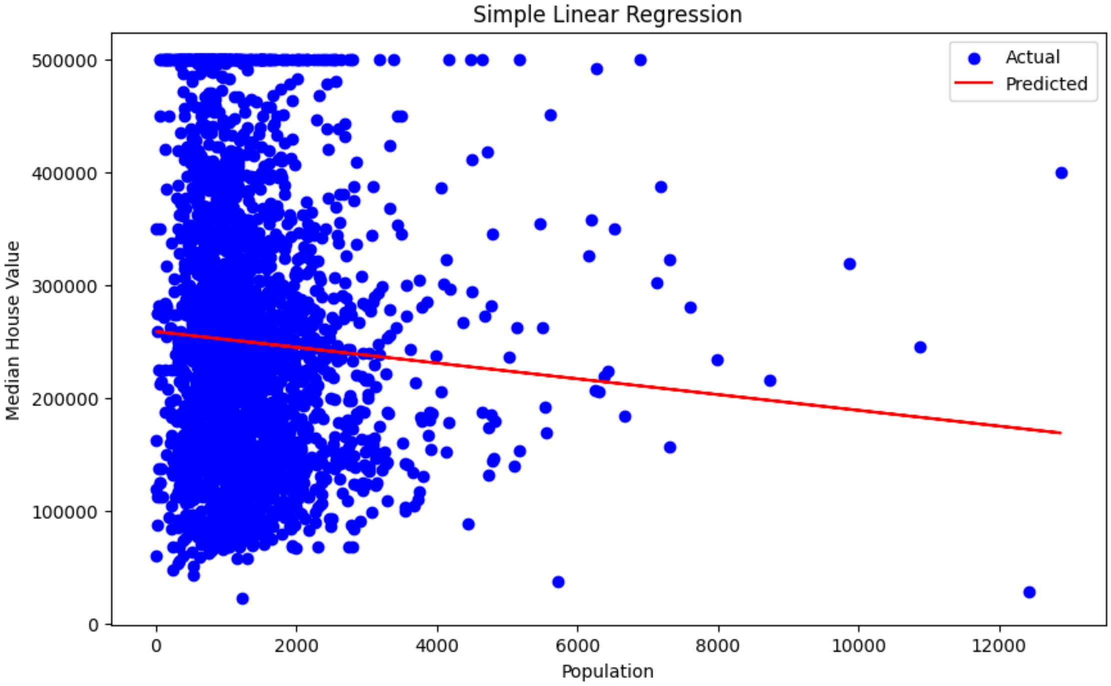
[11]:

```python
mse = mean_squared_error(y, y_pred)
r2 = r2_score(y, y_pred)
```

[12]:
```
print("\nModel Evaluation:")
print("Mean Squared Error (MSE):", mse)
print("R-squared (R2) Score:", r2)
```

```
Model Evaluation:
Mean Squared Error (MSE): 14962715711.248127
R-squared (R2) Score: 0.003302934451935413
```

[13]:
```
plt.figure(figsize=(10, 6))
plt.scatter(X, y, color='blue', label='Actual')
plt.plot(X, y_pred, color='red', label='Predicted')
plt.xlabel('Population')
plt.ylabel('Median House Value')
plt.title('Simple Linear Regression')
plt.legend()
plt.show()
```



[14]:
```
print("Intercept:", model.intercept_)
print("Coefficient:", model.coef_)
```

```
Intercept: 258513.11395519556
Coefficient: [-6.98620381]
```

[15]:

```
population_median = california_data['population'].median()
population_mean = california_data['population'].mean()
population_std = california_data['population'].std()
```

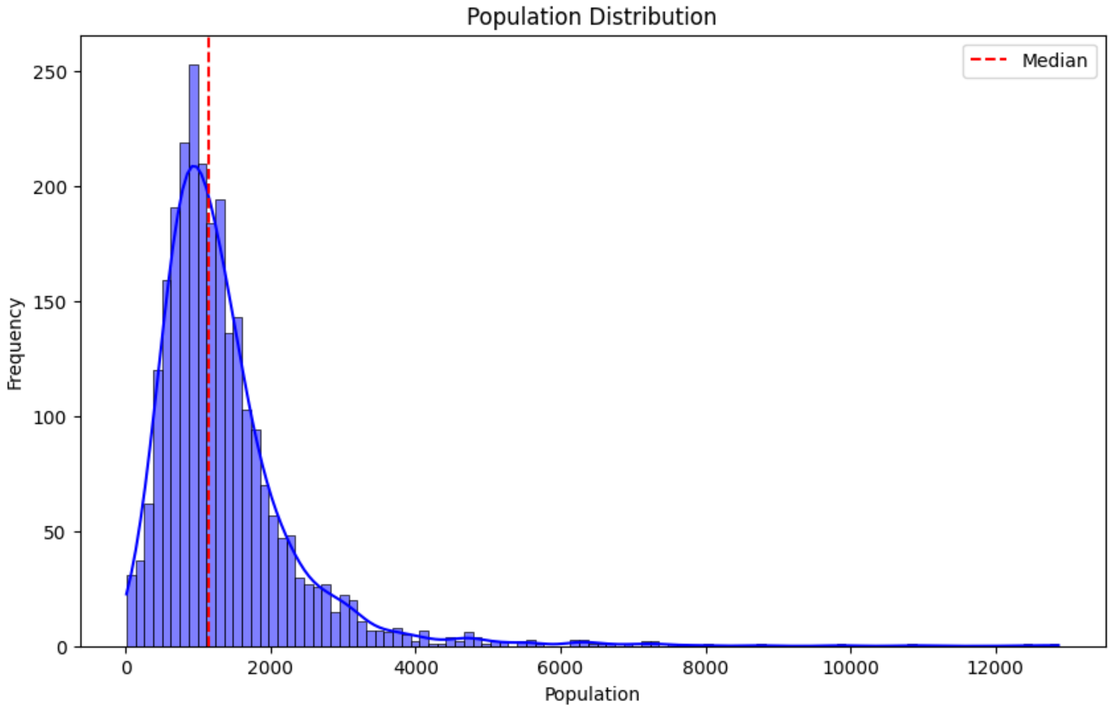
[16]:

```
print("\nPopulation Statistics:")
print("Median:", population_median)
print("Mean:", population_mean)
print("Standard Deviation:", population_std)
```

```

Population Statistics:
Median: 1137.5
Mean: 1355.6373668188737
Standard Deviation: 1008.1264118931317
```

[17]:

```
plt.figure(figsize=(10, 6))
sns.histplot(california_data['population'], kde=True, color='blue')
plt.xlabel('Population')
plt.ylabel('Frequency')
plt.title('Population Distribution')
plt.axvline(population_median, color='red', linestyle='--', label='Median')
plt.legend()
plt.show()
```



[18]:

```
median_house_value_median = california_data['median_house_value'].median()
median_house_value_mean = california_data['median_house_value'].mean()
median_house_value_std = california_data['median_house_value'].std()
```

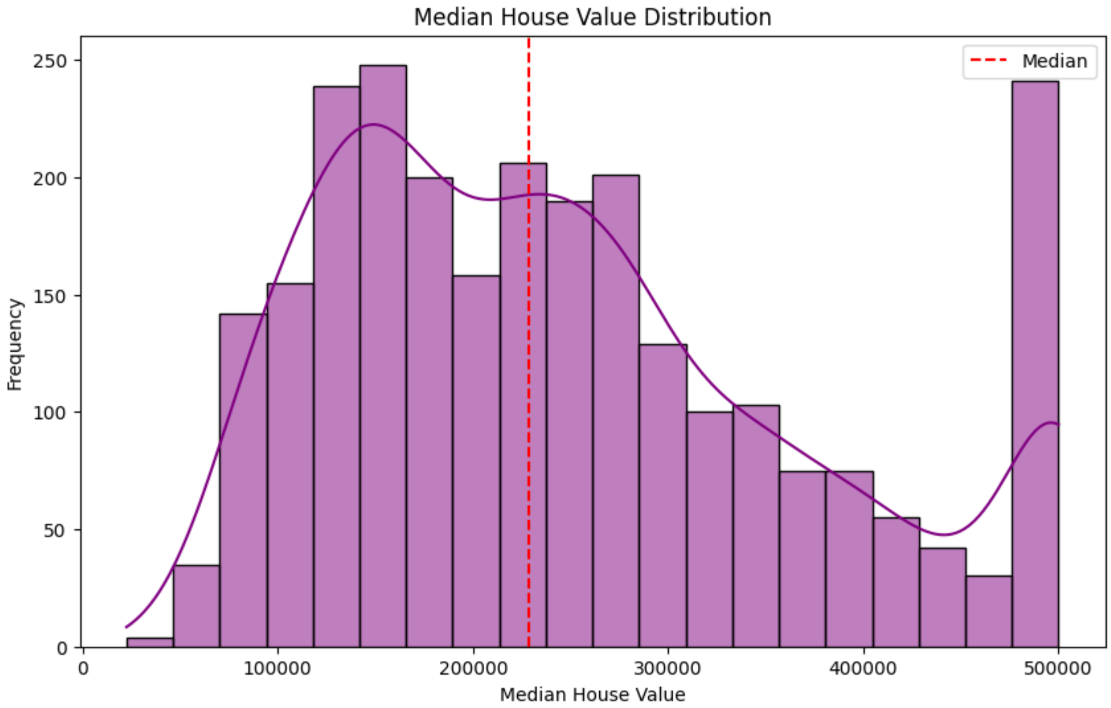
[19]:

```
print("\nMedian House Value Statistics:")
print("Median:", median_house_value_median)
print("Mean:", median_house_value_mean)
print("Standard Deviation:", median_house_value_std)
```

```
Median House Value Statistics:
Median: 228750.0
Mean: 249042.35502283106
Standard Deviation: 122548.0108890553
```

[20]:

```
plt.figure(figsize=(10, 6))
sns.histplot(california_data['median_house_value'], kde=True, color='purple')
plt.xlabel('Median House Value')
plt.ylabel('Frequency')
plt.title('Median House Value Distribution')
plt.axvline(median_house_value_median, color='red', linestyle='--',␣
 ↪label='Median')
plt.legend()
plt.show()
```



[21]:

```
# Median by ocean proximity
median_by_proximity = housing_data.groupby('ocean_proximity').median()
print("Median for each ocean proximity:")
print(median_by_proximity)
```

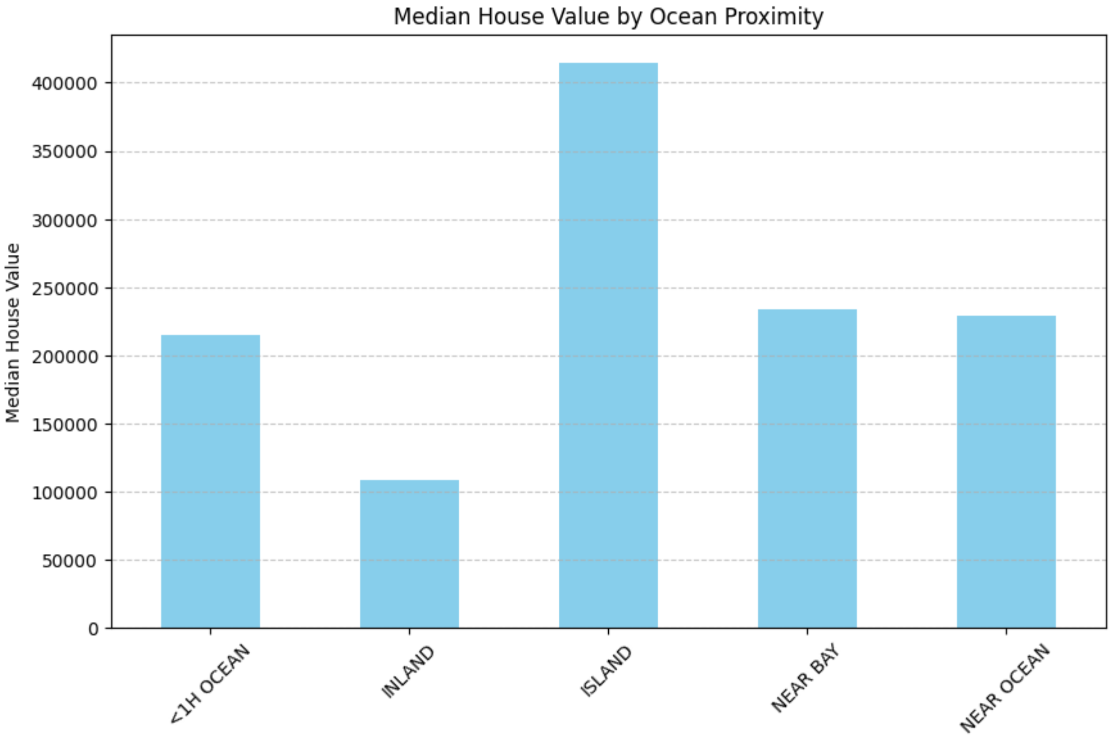
```
Median for each ocean proximity:
                 longitude  latitude  housing_median_age  total_rooms  \
ocean_proximity
<1H OCEAN          -118.28     34.03                30.0       2107.0
INLAND             -120.00     36.97                23.0       2136.0
ISLAND             -118.32     33.34                52.0       1675.0
NEAR BAY           -122.25     37.79                39.0       2082.5
NEAR OCEAN         -118.25     33.79                29.0       2197.0

                 total_bedrooms  population  households  median_income  \
ocean_proximity
<1H OCEAN                 438.0      1246.0       420.0        3.87900
INLAND                    423.0      1124.5       385.0        2.98980
ISLAND                    512.0       733.0       288.0        2.73610
NEAR BAY                  423.0      1033.5       404.5        3.81865
NEAR OCEAN                464.0      1137.5       429.0        3.64830

                 median_house_value
ocean_proximity
<1H OCEAN                  215000.0
INLAND                     108700.0
ISLAND                     414700.0
NEAR BAY                   233800.0
NEAR OCEAN                 228750.0
```

[10]:

```
median_by_proximity = housing_data.groupby('ocean_proximity').median()
```

[11]:

```
plt.figure(figsize=(10, 6))
median_by_proximity['median_house_value'].plot(kind='bar', color='skyblue')
plt.title('Median House Value by Ocean Proximity')
plt.xlabel('Ocean Proximity')
plt.ylabel('Median House Value')
plt.xticks(rotation=45)  # Rotate x-axis labels for better readability
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```


Median House Value by Ocean Proximity
Median House Value
0
50000
100000
150000
200000
250000
300000
350000
400000
<1H OCEAN
INLAND
ISLAND
NEAR BAY
NEAR OCEAN
Ocean Proximity