Updated

March 5, 2024

[27]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
```

[17]:
```
# Read the CSV file
url = r"C:\Users\MUBASHIR KHAN\Desktop\jupyter\Internship Project\California␣
 ↪Housing Prices Dataset.csv"
housing_data = pd.read_csv(url)
```

[18]:
```
# Filter data
california_data = housing_data[housing_data['ocean_proximity'] == 'NEAR OCEAN']
```

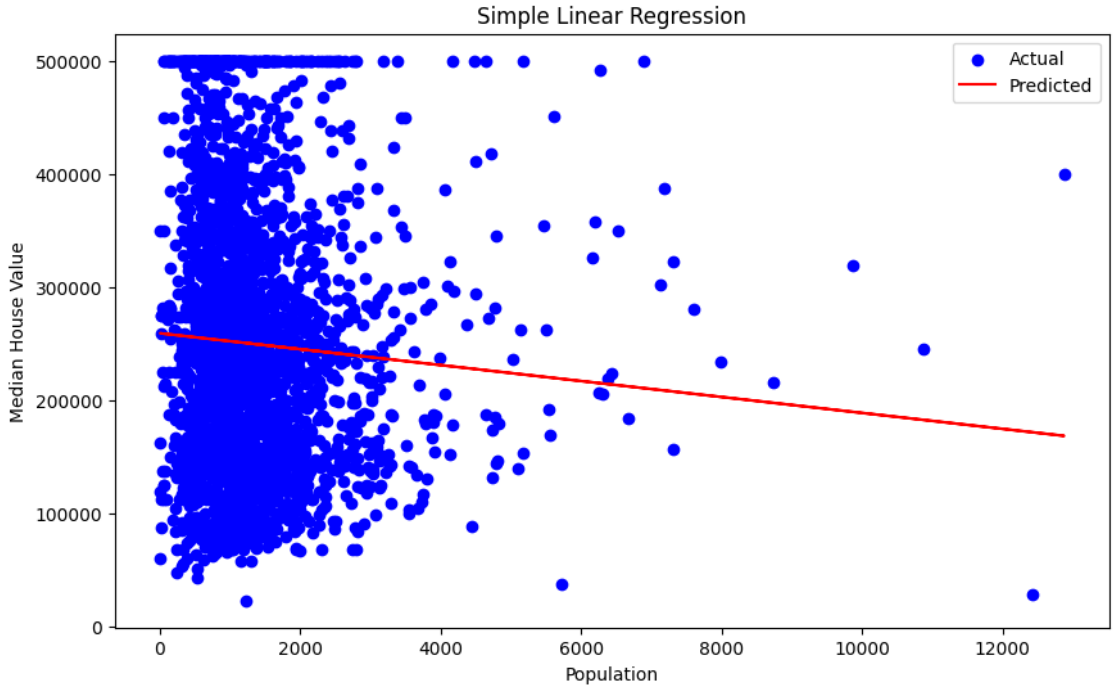
[28]:
```
X = california_data[['population']]  # independent variable
y = california_data['median_house_value']  # dependent variable
```

[30]:
```
model = LinearRegression()
model.fit(X, y)
```

[30]:
```
LinearRegression()
```

[31]:
```
y_pred = model.predict(X)
```

[34]:
```
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()
```



[35]:
```
print("Intercept:", model.intercept_)
print("Coefficient:", model.coef_)
```

```
Intercept: 258984.21877324715
Coefficient: [-7.05330895]
```

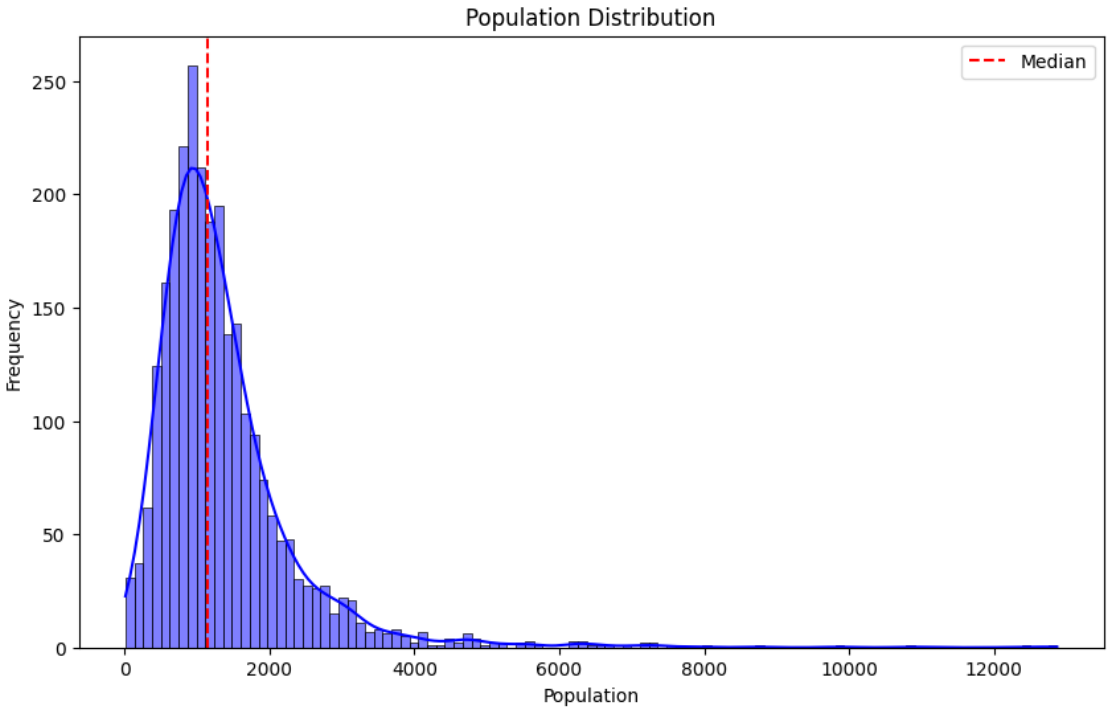
[29]:
```
# Population statistics
population_median = california_data['population'].median()
population_mean = california_data['population'].mean()
population_std = california_data['population'].std()
```

[20]:
```
print("Population Statistics:")
print("Median:", population_median)
print("Mean:", population_mean)
print("Standard Deviation:", population_std)
```

```
Population Statistics:
Median: 1136.5
Mean: 1354.0086531226486
Standard Deviation: 1005.5631663130899
```

[4]:
```
palette = {'population': 'blue', 'median_income': 'green', 'median_house_value':
  ↪ 'red'}
```

[10]:
```
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()
```

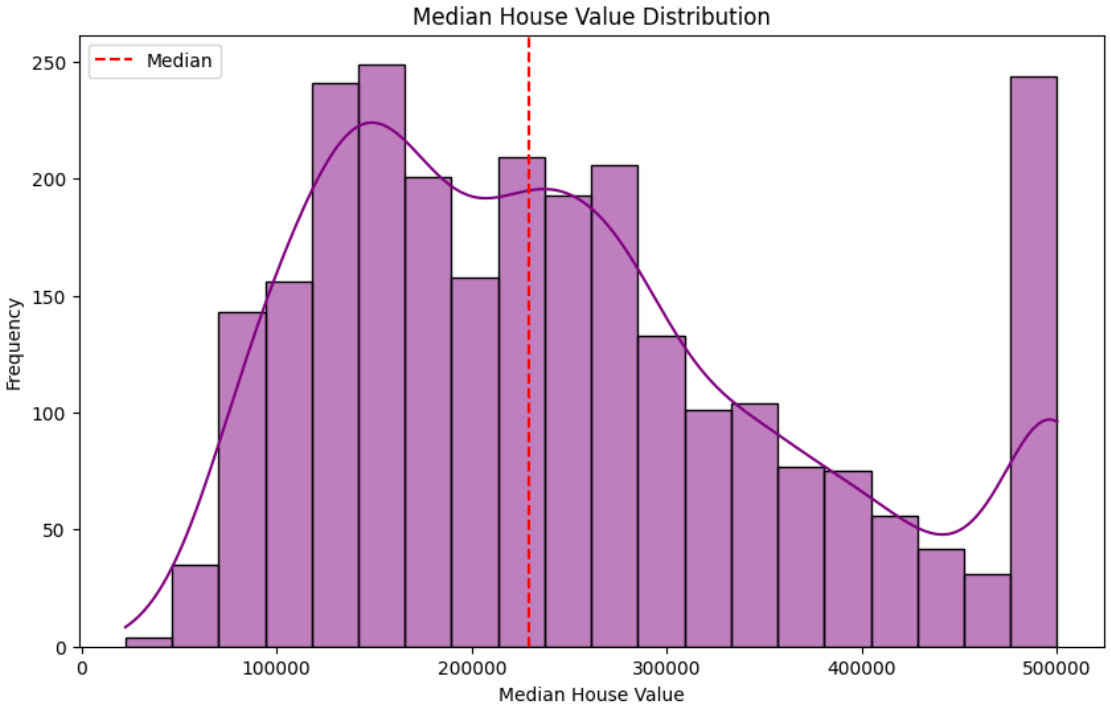


[11]:
```
# Median house value statistics
median_house_value_median = california_data['median_house_value'].median()
median_house_value_mean = california_data['median_house_value'].mean()
median_house_value_std = california_data['median_house_value'].std()
```

[12]:
```
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: 229450.0
Mean: 249433.97742663656
Standard Deviation: 122477.14592684481
```

[13]:
```
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()
```

[16]:
```
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         -118.275     34.03                30.0       2108.0
INLAND            -120.000     36.97                23.0       2131.0
ISLAND            -118.320     33.34                52.0       1675.0
NEAR BAY          -122.250     37.79                39.0       2083.0
NEAR OCEAN        -118.260     33.79                29.0       2195.0
```

ocean_proximity	total_bedrooms	population	households	median_income	\
<1H OCEAN	438.0	1247.0	421.0	3.87500	
INLAND	423.0	1124.0	385.0	2.98770	
ISLAND	512.0	733.0	288.0	2.73610	
NEAR BAY	423.0	1033.5	406.0	3.81865	
NEAR OCEAN	464.0	1136.5	429.0	3.64705	

ocean_proximity	median_house_value
<1H OCEAN	214850.0
INLAND	108500.0
ISLAND	414700.0
NEAR BAY	233800.0
NEAR OCEAN	229450.0