In [1]:

```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset (replace 'heart_disease.csv' with the actual file path)
df = pd.read_csv('heart_main.csv')

# Display the first few rows of the dataset to understand its structure
df.head()
```

Out[1]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

In [3]:

```
# Check for missing values in the dataset
print(df.isnull().sum())

# Check the data types of the columns
print(df.dtypes)

# Get a summary of the dataset (statistics, ranges, etc.)
print(df.describe())
```

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
age           int64
sex           int64
cp            int64
trestbps      int64
chol          int64
fbs           int64
restecg       int64
thalach       int64
exang         int64
oldpeak     float64
slope         int64
ca            int64
thal          int64
target        int64
dtype: object
               age          sex           cp     trestbps        chol  \
count  1025.000000  1025.000000  1025.000000  1025.000000  1025.00000   
mean     54.434146     0.695610     0.942439   131.611707   246.00000   
std       9.072290     0.460373     1.029641    17.516718    51.59251   
min      29.000000     0.000000     0.000000    94.000000   126.00000   
25%      48.000000     0.000000     0.000000   120.000000   211.00000   
50%      56.000000     1.000000     1.000000   130.000000   240.00000   
75%      61.000000     1.000000     2.000000   140.000000   275.00000   
max      77.000000     1.000000     3.000000   200.000000   564.00000   

               fbs      restecg      thalach        exang      oldpeak  \
count  1025.000000  1025.000000  1025.000000  1025.000000  1025.000000   
mean      0.149268     0.529756   149.114146     0.336585     1.071512   
std       0.356527     0.527878    23.005724     0.472772     1.175053   
min       0.000000     0.000000    71.000000     0.000000     0.000000   
25%       0.000000     0.000000   132.000000     0.000000     0.000000   
50%       0.000000     1.000000   152.000000     0.000000     0.800000   
75%       0.000000     1.000000   166.000000     1.000000     1.800000   
max       1.000000     2.000000   202.000000     1.000000     6.200000   

             slope           ca         thal       target  
count  1025.000000  1025.000000  1025.000000  1025.000000  
mean      1.385366     0.754146     2.323902     0.513171  
std       0.617755     1.030798     0.620660     0.500070  
min       0.000000     0.000000     0.000000     0.000000  
25%       1.000000     0.000000     2.000000     0.000000  
50%       1.000000     0.000000     2.000000     1.000000  
75%       2.000000     1.000000     3.000000     1.000000  
max       2.000000     4.000000     3.000000     1.000000  
```

In [5]:

```python
# Plotting histograms for numeric features to understand their distributions
plt.figure(figsize=(12, 8))

# Age Distribution
```

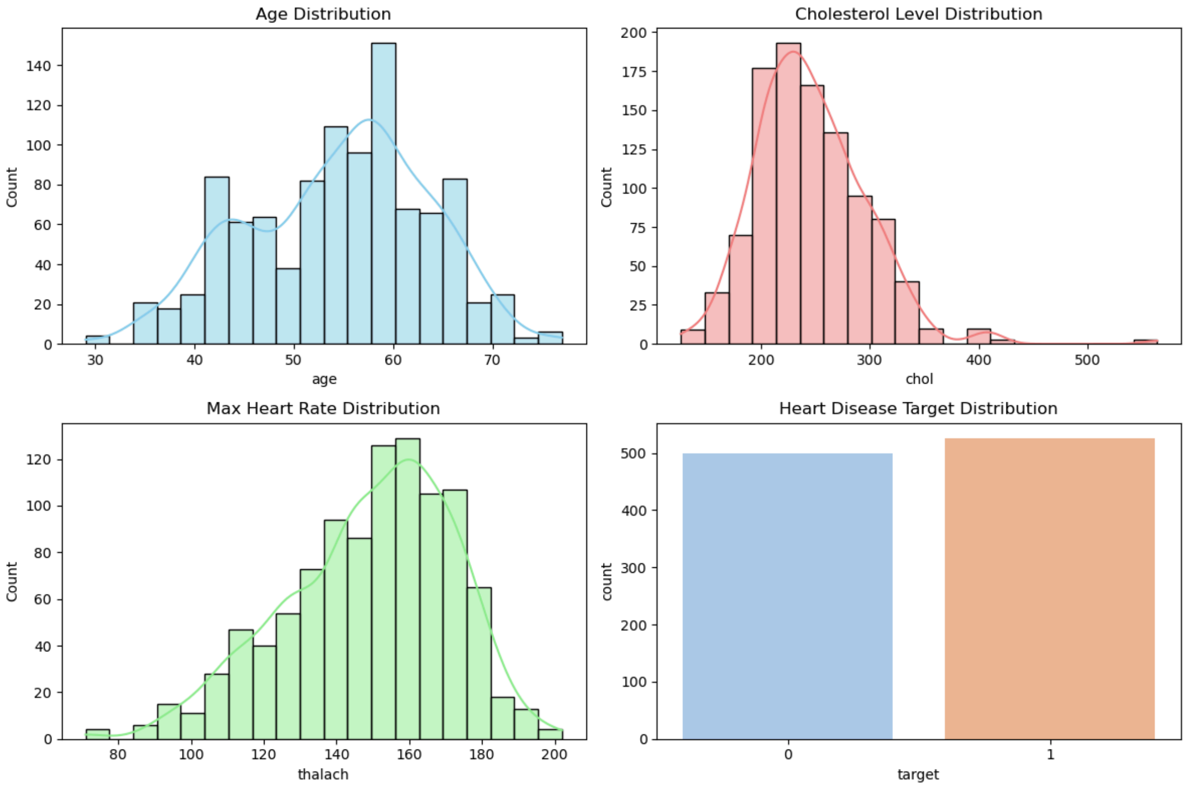
```
plt.subplot(2, 2, 1)
sns.histplot(df['age'], kde=True, bins=20, color='skyblue')
plt.title('Age Distribution')

# Cholesterol Level Distribution
plt.subplot(2, 2, 2)
sns.histplot(df['chol'], kde=True, bins=20, color='lightcoral')
plt.title('Cholesterol Level Distribution')

# Max Heart Rate Distribution
plt.subplot(2, 2, 3)
sns.histplot(df['thalach'], kde=True, bins=20, color='lightgreen')
plt.title('Max Heart Rate Distribution')

# Target Distribution (Heart Disease: 0 - No, 1 - Yes)
plt.subplot(2, 2, 4)
sns.countplot(x='target', data=df, palette='pastel')
plt.title('Heart Disease Target Distribution')

plt.tight_layout()
plt.show()
```



In [7]:

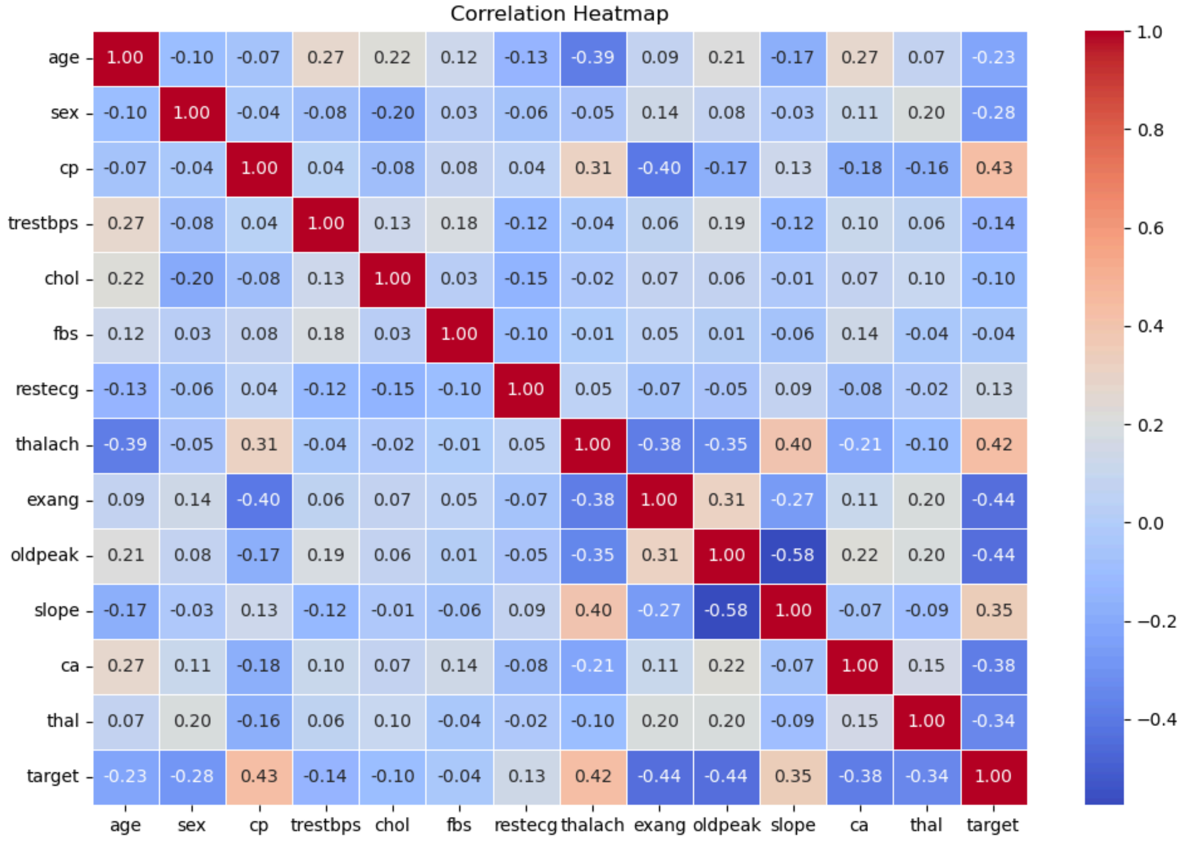
```
# Compute the correlation matrix
corr_matrix = df.corr()

# Set up the matplotlib figure
plt.figure(figsize=(12, 8))

# Generate the heatmap
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f', linewidths=0.5, cb

# Title
plt.title('Correlation Heatmap')

# Show the plot
plt.show()
```

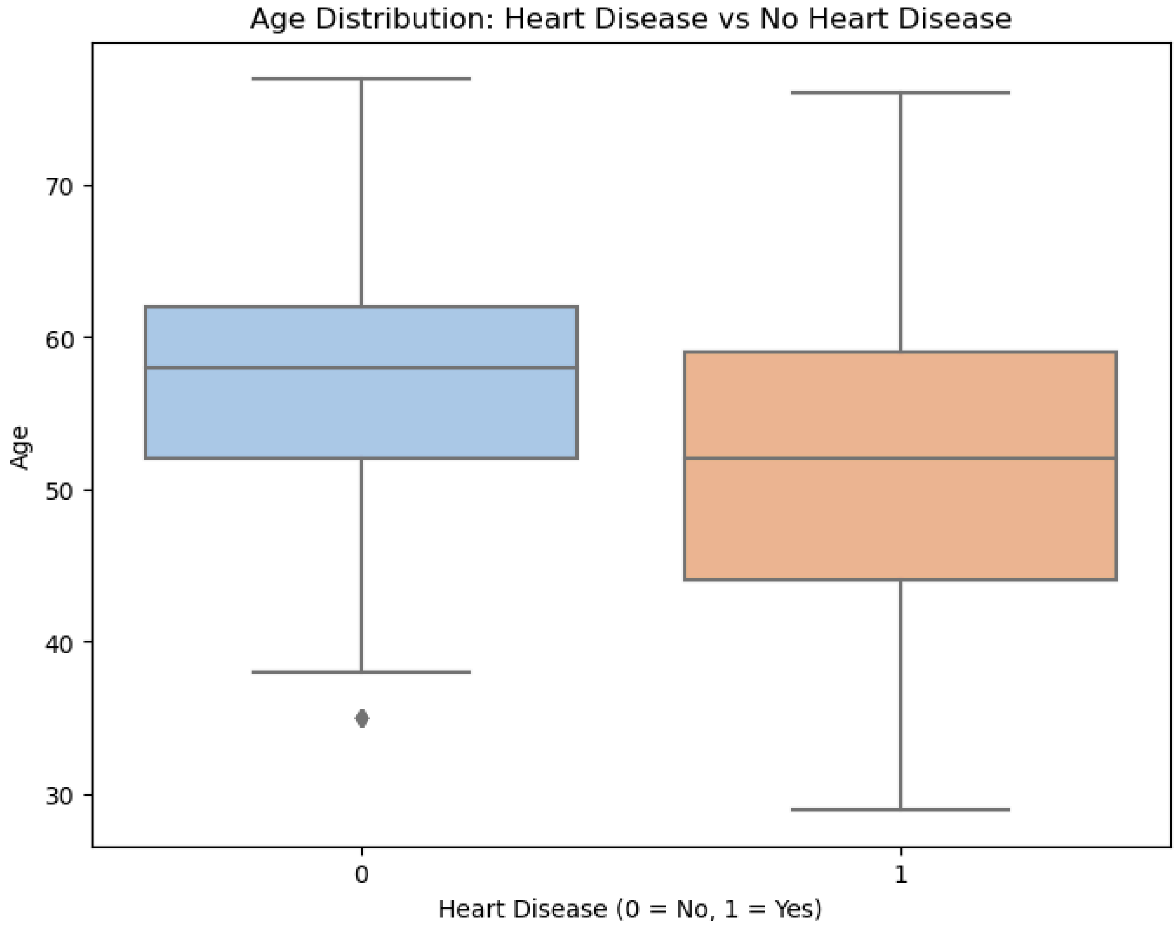


	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.00	-0.10	-0.07	0.27	0.22	0.12	-0.13	-0.39	0.09	0.21	-0.17	0.27	0.07	-0.23
sex	-0.10	1.00	-0.04	-0.08	-0.20	0.03	-0.06	-0.05	0.14	0.08	-0.03	0.11	0.20	-0.28
cp	-0.07	-0.04	1.00	0.04	-0.08	0.08	0.04	0.31	-0.40	-0.17	0.13	-0.18	-0.16	0.43
trestbps	0.27	-0.08	0.04	1.00	0.13	0.18	-0.12	-0.04	0.06	0.19	-0.12	0.10	0.06	-0.14
chol	0.22	-0.20	-0.08	0.13	1.00	0.03	-0.15	-0.02	0.07	0.06	-0.01	0.07	0.10	-0.10
fbs	0.12	0.03	0.08	0.18	0.03	1.00	-0.10	-0.01	0.05	0.01	-0.06	0.14	-0.04	-0.04
restecg	-0.13	-0.06	0.04	-0.12	-0.15	-0.10	1.00	0.05	-0.07	-0.05	0.09	-0.08	-0.02	0.13
thalach	-0.39	-0.05	0.31	-0.04	-0.02	-0.01	0.05	1.00	-0.38	-0.35	0.40	-0.21	-0.10	0.42
exang	0.09	0.14	-0.40	0.06	0.07	0.05	-0.07	-0.38	1.00	0.31	-0.27	0.11	0.20	-0.44
oldpeak	0.21	0.08	-0.17	0.19	0.06	0.01	-0.05	-0.35	0.31	1.00	-0.58	0.22	0.20	-0.44
slope	-0.17	-0.03	0.13	-0.12	-0.01	-0.06	0.09	0.40	-0.27	-0.58	1.00	-0.07	-0.09	0.35
ca	0.27	0.11	-0.18	0.10	0.07	0.14	-0.08	-0.21	0.11	0.22	-0.07	1.00	0.15	-0.38
thal	0.07	0.20	-0.16	0.06	0.10	-0.04	-0.02	-0.10	0.20	0.20	-0.09	0.15	1.00	-0.34
target	-0.23	-0.28	0.43	-0.14	-0.10	-0.04	0.13	0.42	-0.44	-0.44	0.35	-0.38	-0.34	1.00

Question 1: What is the average age of people with heart disease vs. those without?

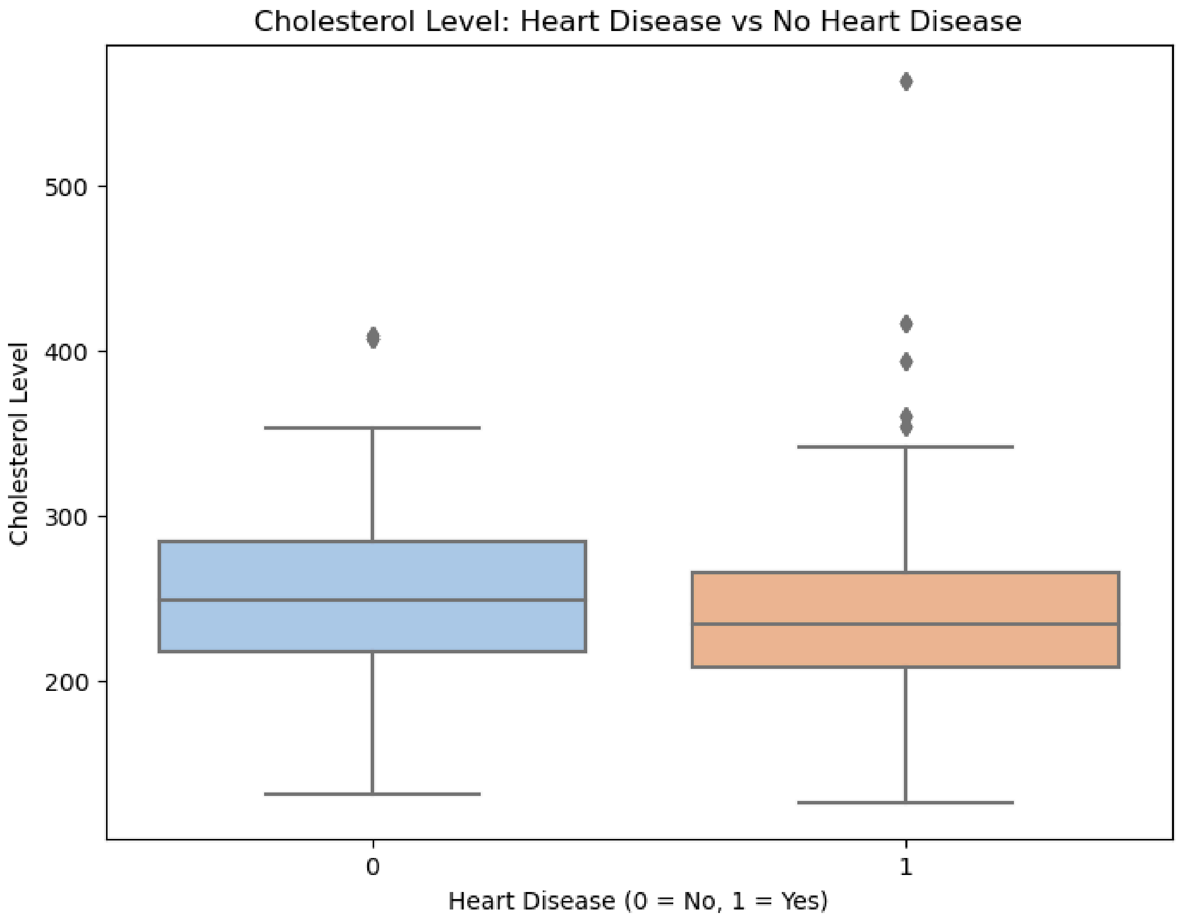
In [10]:

```
# Boxplot for Age distribution based on heart disease (target: 0 = no, 1 = yes)
plt.figure(figsize=(8, 6))
sns.boxplot(x='target', y='age', data=df, palette='pastel')
plt.title('Age Distribution: Heart Disease vs No Heart Disease')
plt.xlabel('Heart Disease (0 = No, 1 = Yes)')
plt.ylabel('Age')
plt.show()
```



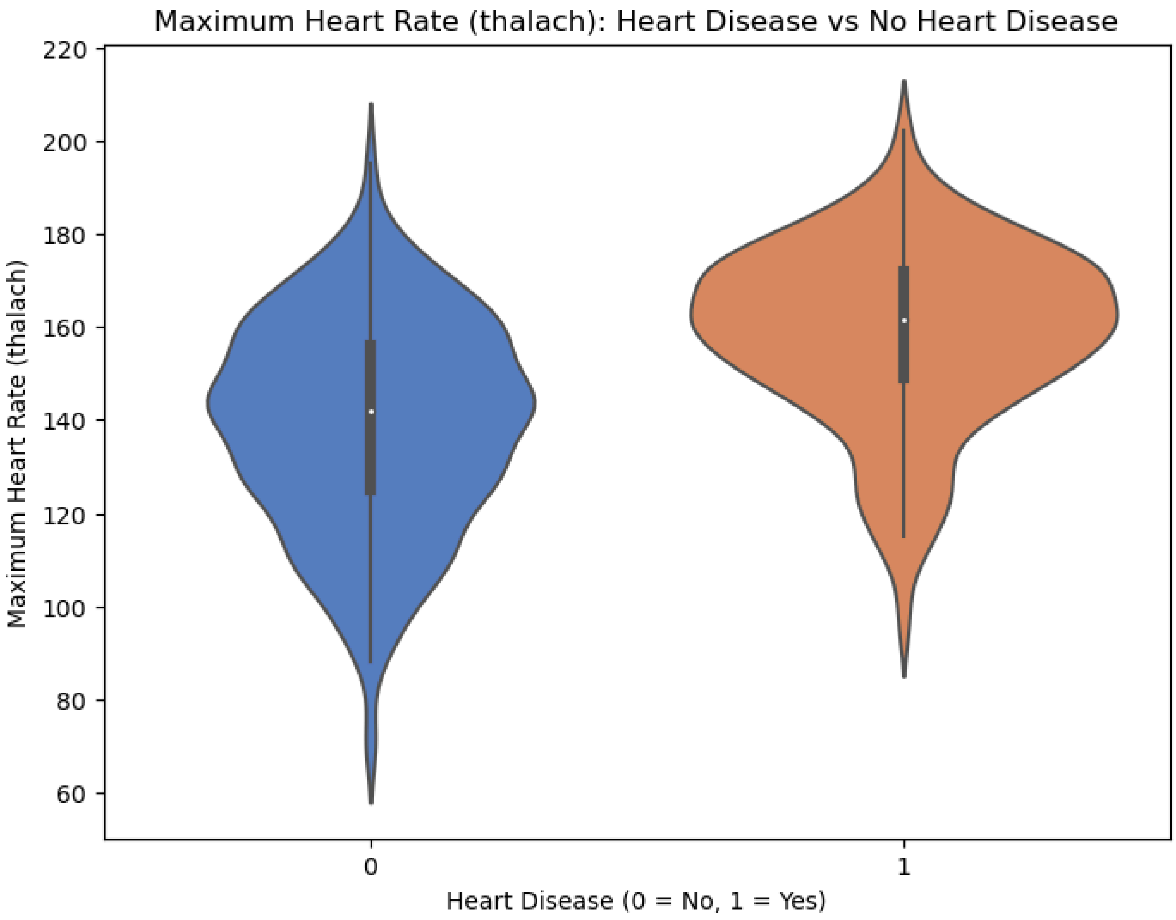
Question 2: Does cholesterol level correlate with the presence of heart disease ?

```
In [13]: # Boxplot for Cholesterol level distribution based on heart disease
         plt.figure(figsize=(8, 6))
         sns.boxplot(x='target', y='chol', data=df, palette='pastel')
         plt.title('Cholesterol Level: Heart Disease vs No Heart Disease')
         plt.xlabel('Heart Disease (0 = No, 1 = Yes)')
         plt.ylabel('Cholesterol Level')
         plt.show()
```



Question 3: What is the distribution of maximum heart rate (thalach) for people with and without heart disease?

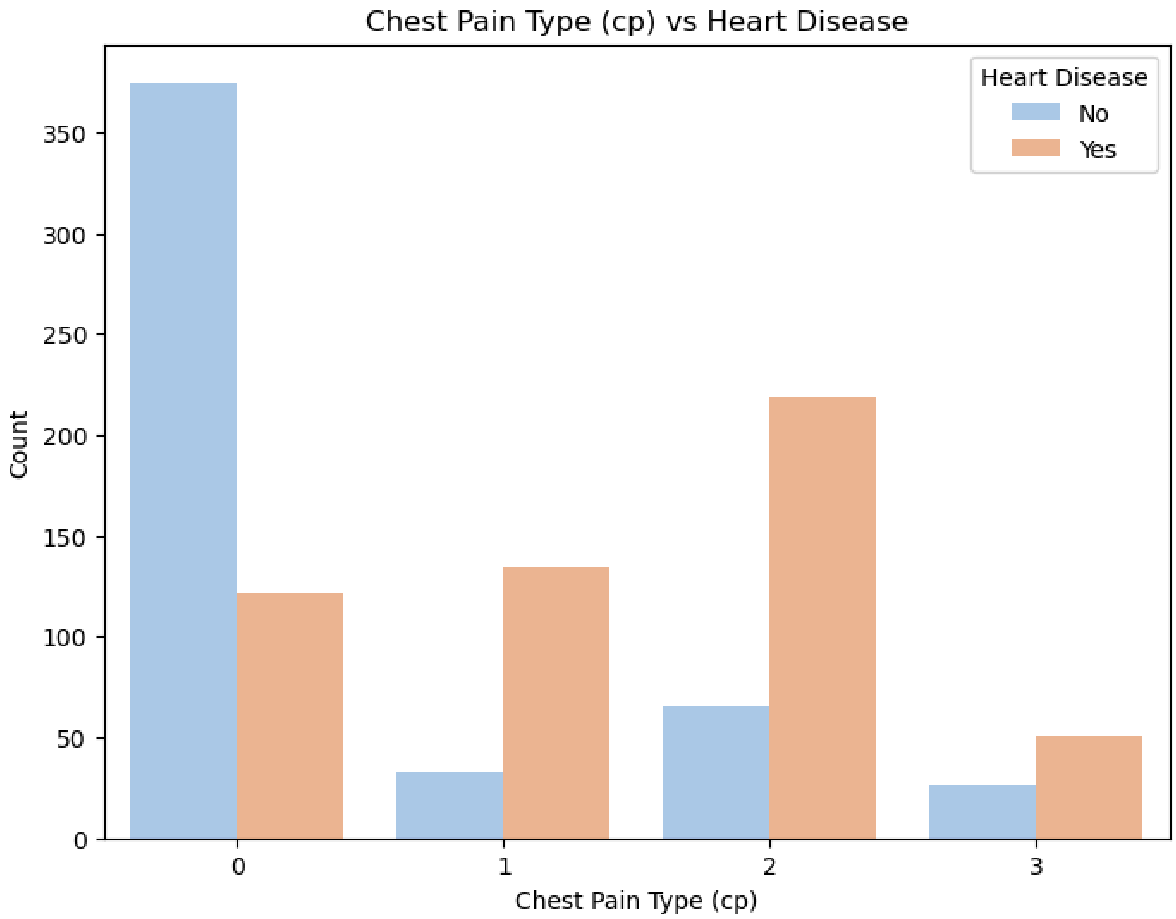
In [16]:
```
# Violin plot for maximum heart rate (thalach) distribution
plt.figure(figsize=(8, 6))
sns.violinplot(x='target', y='thalach', data=df, palette='muted', split=True)
plt.title('Maximum Heart Rate (thalach): Heart Disease vs No Heart Disease')
plt.xlabel('Heart Disease (0 = No, 1 = Yes)')
plt.ylabel('Maximum Heart Rate (thalach)')
plt.show()
```



Question 4: Does the type of chest pain (cp) affect the likelihood of heart disease?

In [19]:

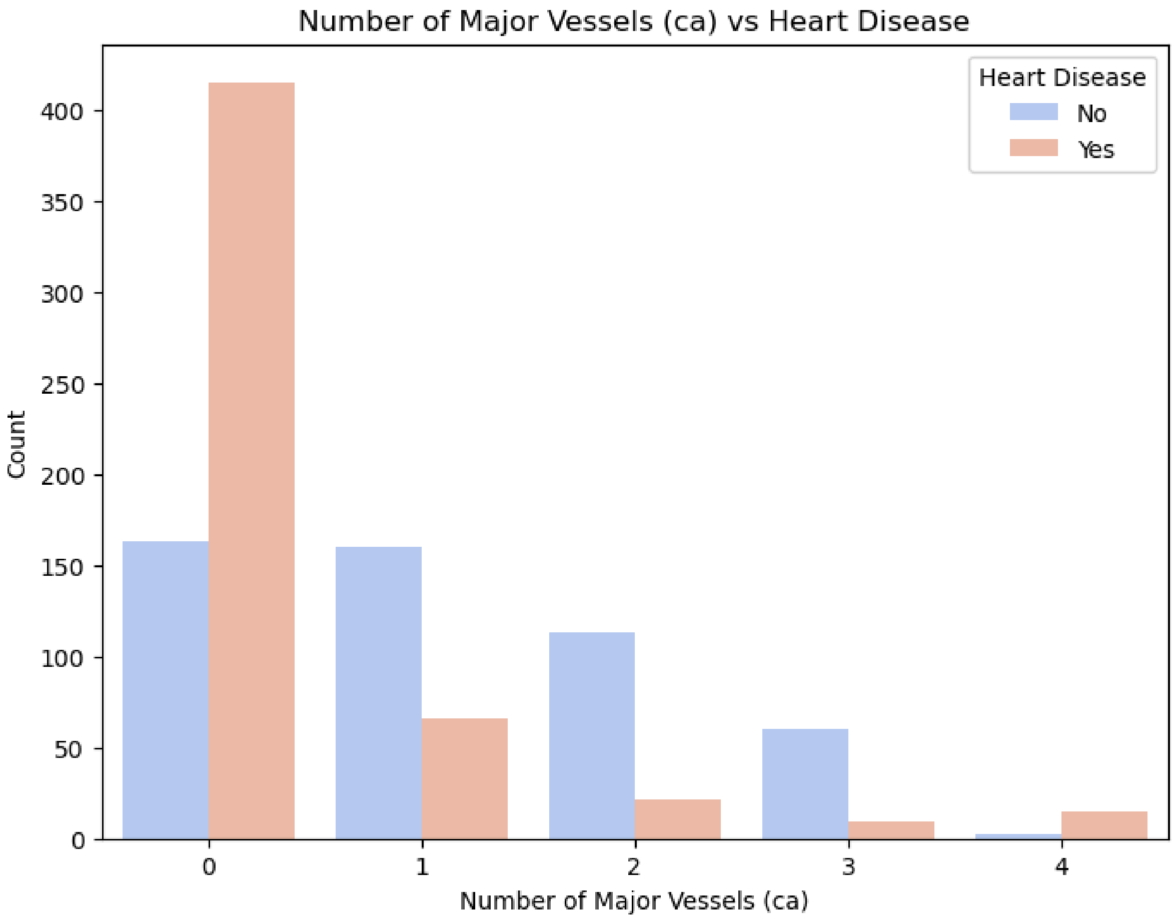
```
# Countplot for chest pain types (cp) based on heart disease
plt.figure(figsize=(8, 6))
sns.countplot(x='cp', hue='target', data=df, palette='pastel')
plt.title('Chest Pain Type (cp) vs Heart Disease')
plt.xlabel('Chest Pain Type (cp)')
plt.ylabel('Count')
plt.legend(title='Heart Disease', labels=['No', 'Yes'])
plt.show()
```



Question 5: What is the relationship between the number of major vessels (ca) and heart disease?

In [25]:

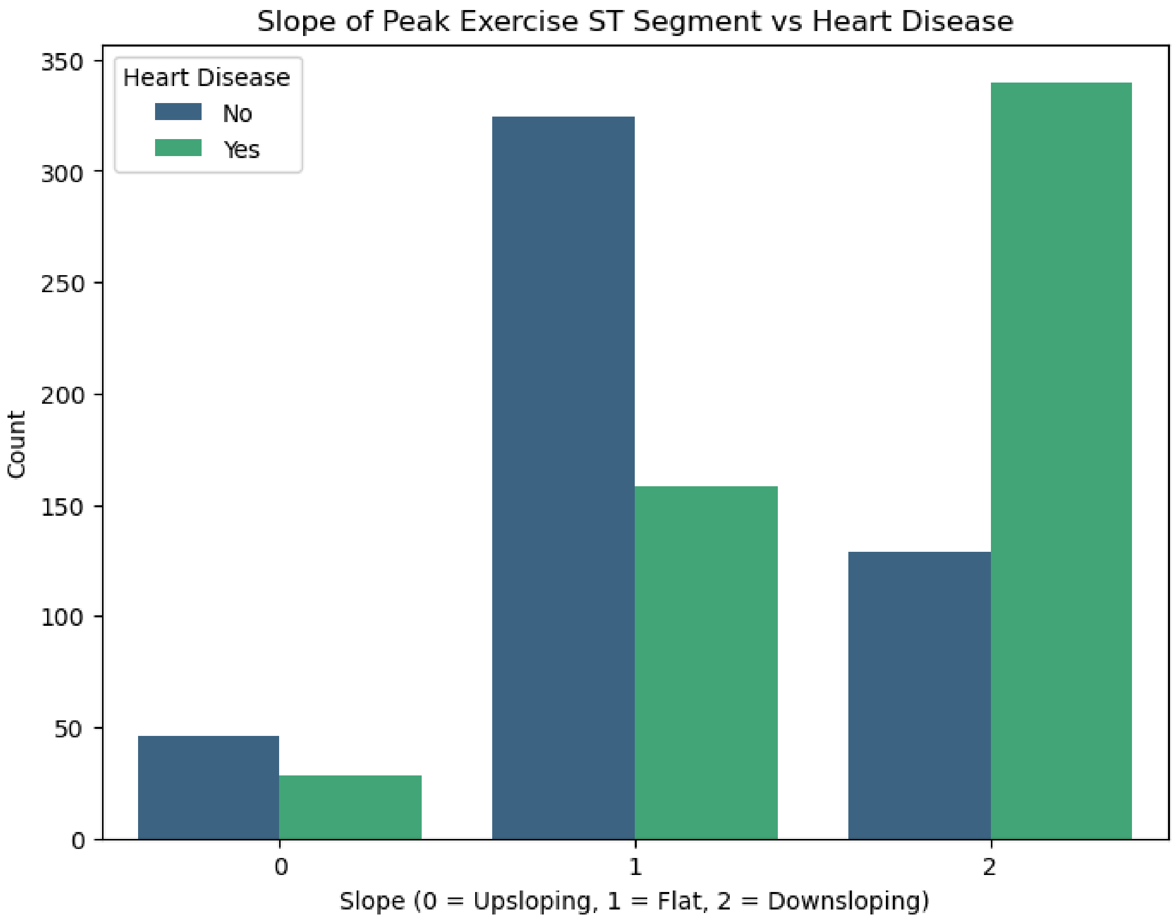
```python
# Bar plot for number of major vessels (ca) based on heart disease
plt.figure(figsize=(8, 6))
sns.countplot(x='ca', hue='target', data=df, palette='coolwarm')
plt.title('Number of Major Vessels (ca) vs Heart Disease')
plt.xlabel('Number of Major Vessels (ca)')
plt.ylabel('Count')
plt.legend(title='Heart Disease', labels=['No', 'Yes'])
plt.show()
```



Question 6: How does the slope of the peak exercise ST segment (slope) correlate with heart disease?

In [28]:

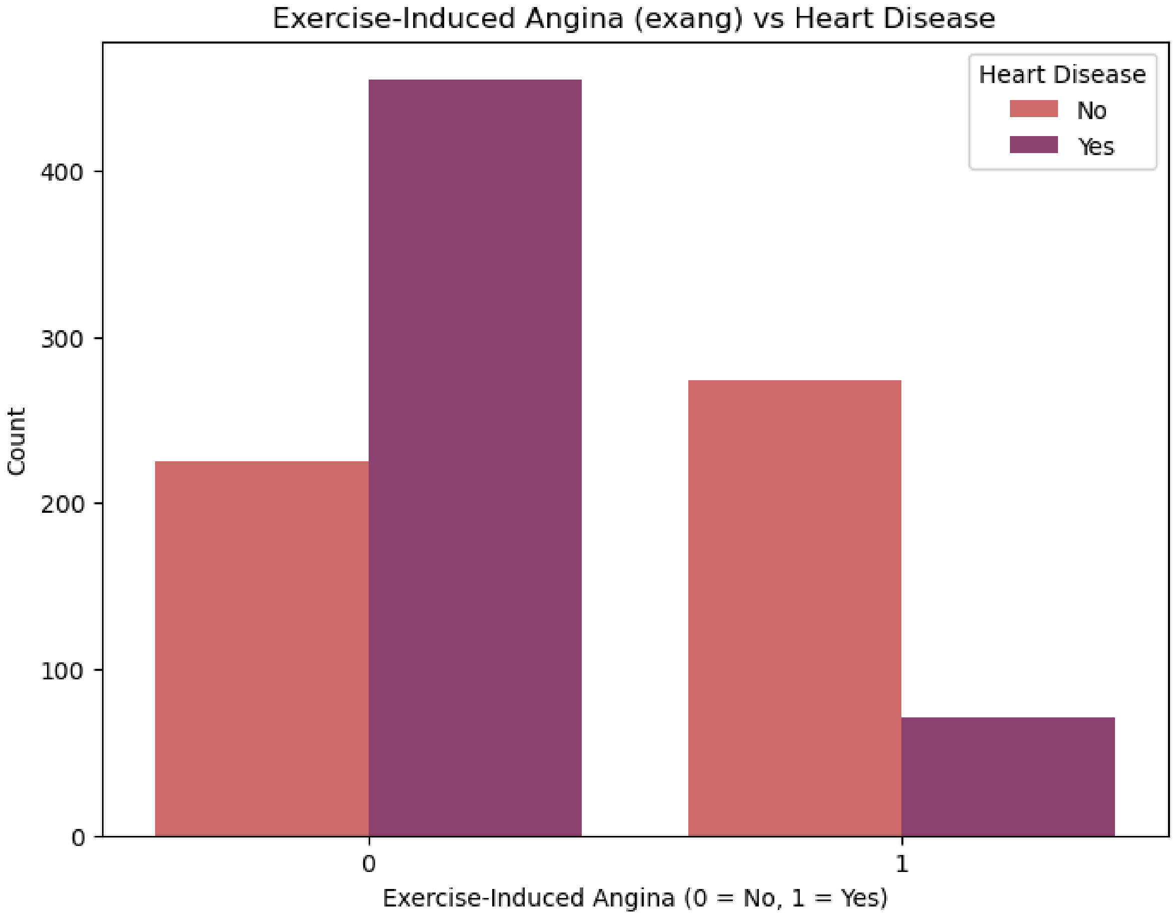
```
# Countplot for slope of peak exercise ST segment
plt.figure(figsize=(8, 6))
sns.countplot(x='slope', hue='target', data=df, palette='viridis')
plt.title('Slope of Peak Exercise ST Segment vs Heart Disease')
plt.xlabel('Slope (0 = Upsloping, 1 = Flat, 2 = Downsloping)')
plt.ylabel('Count')
plt.legend(title='Heart Disease', labels=['No', 'Yes'])
plt.show()
```



Question 7: What is the effect of exercise-induced angina (exang) on heart disease?

In [32]:

```
# Countplot for exercise-induced angina (exang) based on heart disease
plt.figure(figsize=(8, 6))
sns.countplot(x='exang', hue='target', data=df, palette='flare')
plt.title('Exercise-Induced Angina (exang) vs Heart Disease')
plt.xlabel('Exercise-Induced Angina (0 = No, 1 = Yes)')
plt.ylabel('Count')
plt.legend(title='Heart Disease', labels=['No', 'Yes'])
plt.show()
```



Insights from Heart Disease Analysis

Question 1: What is the average age of people with heart disease vs. those without? The boxplot revealed that individuals with heart disease tend to be older compared to those without heart disease. Median ages: With heart disease: ~57 years. Without heart disease: ~52 years.

Question 2: Does cholesterol level correlate with the presence of heart disease? The cholesterol levels (chol) are generally higher in individuals without heart disease, but there is significant overlap. No strong direct correlation was observed between cholesterol levels and heart disease.

Question 3: What is the distribution of maximum heart rate (thalach) for people with and without heart disease? People with heart disease tend to have lower maximum heart rates. Individuals without heart disease often have higher thalach values (>150 bpm).

Question 4: Does the type of chest pain (cp) affect the likelihood of heart disease? Chest pain type 2 (non-anginal pain) and type 3 (asymptomatic) are more prevalent among people without heart disease. Chest pain type 1 (typical angina) and type 0 (atypical angina) are more commonly associated with heart disease.

Question 5: What is the relationship between the number of major vessels (ca) and heart disease? A higher number of blocked major vessels (ca = 2 or 3) strongly correlates with the presence of heart disease. People with no major vessel blockage (ca = 0) are less likely to have heart disease.

Question 6: How does the slope of the peak exercise ST segment (slope) correlate with heart disease? Slope = 1 (flat) is more common among individuals with heart disease. Slope = 0 (upsloping) is more prevalent in those without heart disease.

Question 7: What is the effect of exercise-induced angina (exang) on heart disease? Exercise-induced angina (exang = 1) is strongly associated with heart disease. Individuals without heart disease predominantly report no angina (exang = 0).

```
In [ ]:
```