Case Study

The objective of this assignment is to analyze the Heart Disease dataset and create various visualizations using Matplotlib and Seaborn. The goal is to gain insights into the data and understand the relationships between different variables.

Tasks

Task 1: Age Distribution

Create a histogram to visualize the distribution of ages in the dataset.

• Use Matplotlib to create the histogram.

• Label the axes and provide a title.

Task 2: Gender Distribution

Create a bar plot to visualize the distribution of gender in the dataset.

• Use Seaborn to create the bar plot.

• Label the axes and provide a title.

Task 3: Chest Pain Type vs. Heart Disease

Create a count plot to visualize the relationship between chest pain type (cp) and the

presence of heart disease (target).

• Use Seaborn to create the count plot.

• Use different colors to differentiate between the presence and absence of heart

disease.

• Label the axes and provide a title.

Task 4: Cholesterol Levels

Create a box plot to visualize the distribution of cholesterol levels (chol) for patients with

and without heart disease.

• Use Seaborn to create the box plot.

• Label the axes and provide a title.

Task 5: Pair Plot

Create a pair plot to visualize relationships between multiple variables.

• Use Seaborn to create the pair plot.

• Include the following variables: age, trestbps, chol, thalach, and target.

• Differentiate the points based on the target variable.

Task 6: Correlation Heatmap

Create a heatmap to visualize the correlation between different attributes in the dataset.

• Use Seaborn to create the heatmap.

• Display the correlation values on the heatmap.

• Provide a title.

Task 7: Exercise Induced Angina vs. Maximum Heart Rate

Create a scatter plot to visualize the relationship between exercise-induced angina (exang) and maximum heart rate (thalach).

• Use Matplotlib to create the scatter plot.

• Color the points based on the presence of heart disease (target).

• Label the axes and provide a title.

CODE

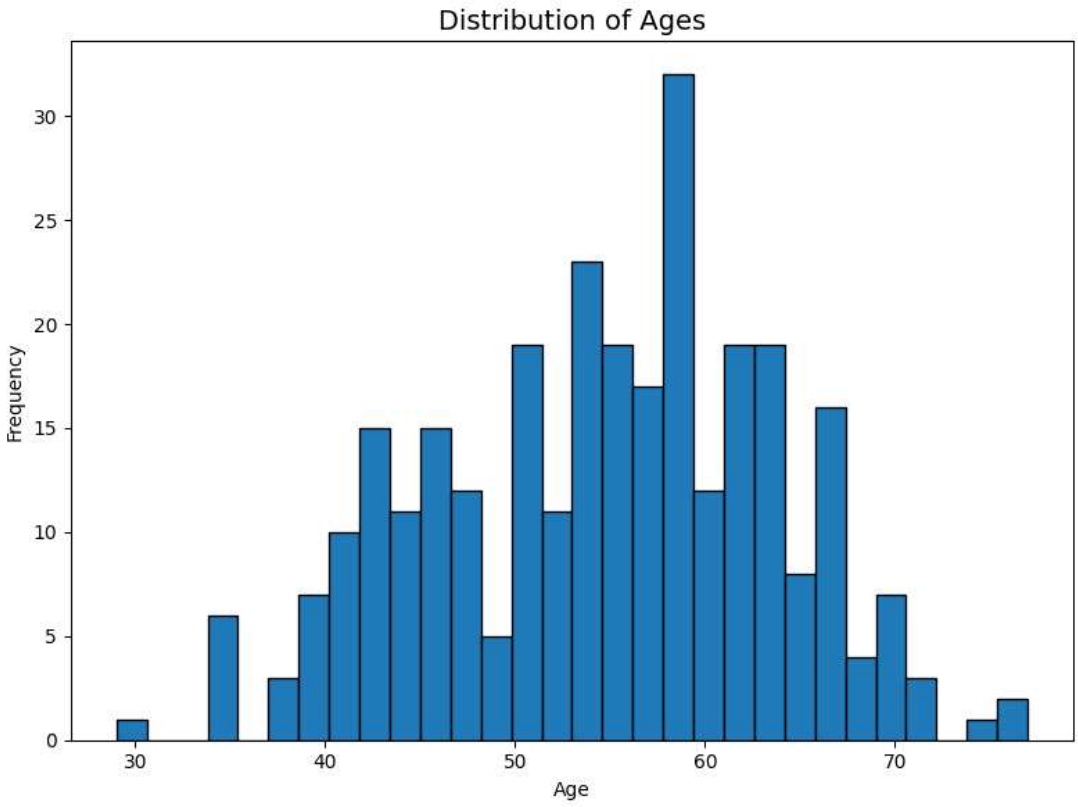
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load Dataset
data_url = 'https://archive.ics.uci.edu/ml/machine-learning-
databases/heart-disease/processed.cleveland.data'
```

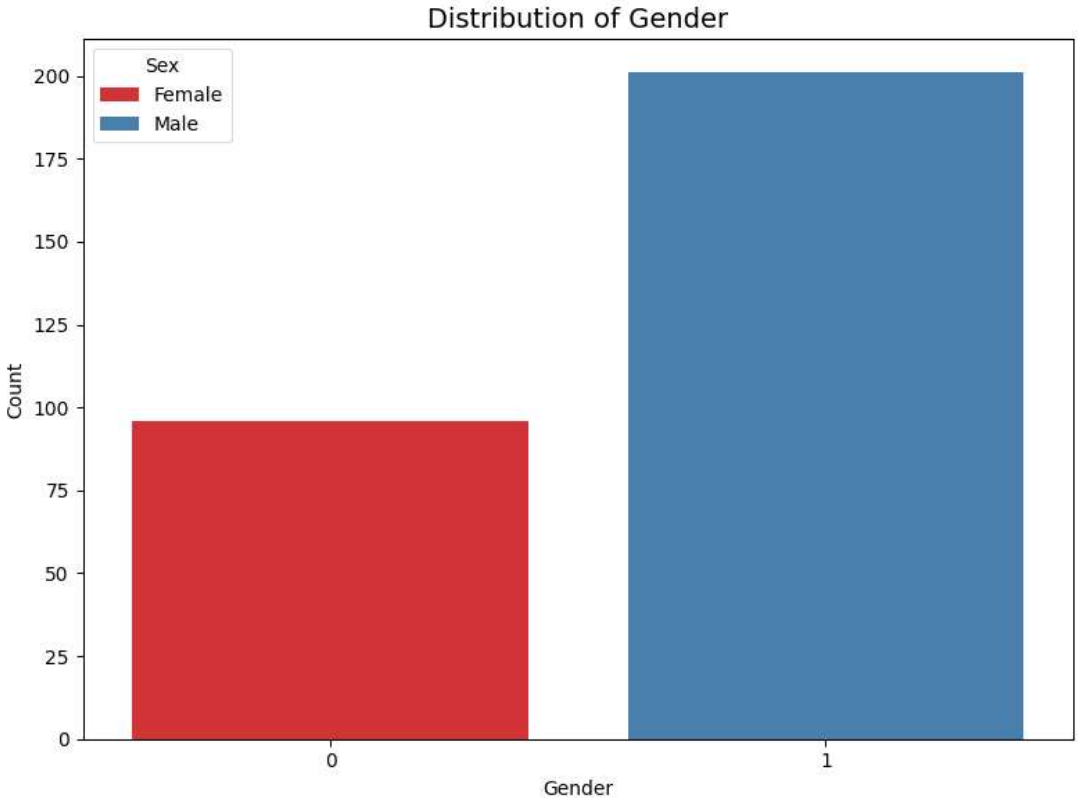
```
columns = ['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg',
'thalach', 'exang',
           'oldpeak', 'slope', 'ca', 'thal', 'target']

data = pd.read_csv(data_url, names=columns)

# Data cleaning
data['sex'] = data['sex'].apply(lambda x: 1 if x > 0 else 0)
data['cp'] = data['cp'].apply(lambda x: min(x, 3))
data['target'] = data['target'].apply(lambda x: 1 if x > 0 else 0)

# Convert columns to numeric, coerce errors to NaN
for column in data.columns:
    data[column] = pd.to_numeric(data[column], errors='coerce')

# Drop rows with NaN values
data_cleaned = data.dropna()

# Plot 1: Age Distribution Histogram
plt.figure(figsize=(8, 6))
plt.hist(data_cleaned['age'], bins=30, edgecolor='black')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.title('Distribution of Ages', fontsize=14)
plt.tight_layout()
plt.savefig('age_distribution_histogram.png')
plt.show()

"""
This histogram shows the distribution of ages in the cleaned dataset.
Most of the patients are between 40 and 70 years old, with fewer
patients in the extremes of the age range.
"""

# Plot 2: Gender Distribution
plt.figure(figsize=(8, 6))
sns.countplot(x='sex', data=data_cleaned, hue='sex', palette='Set1')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.title('Distribution of Gender', fontsize=14)
plt.legend(title='Sex', labels=['Female', 'Male'])
plt.tight_layout()
plt.savefig('gender_distribution.png')
plt.show()

"""
This count plot shows the distribution of genders in the cleaned
dataset.
It highlights that there are more male patients than female patients
in the dataset.
"""

# Plot 3: Chest Pain Type vs Heart Disease
plt.figure(figsize=(8, 6))
sns.countplot(x='cp', hue='target', data=data_cleaned,
palette='Set2')
```

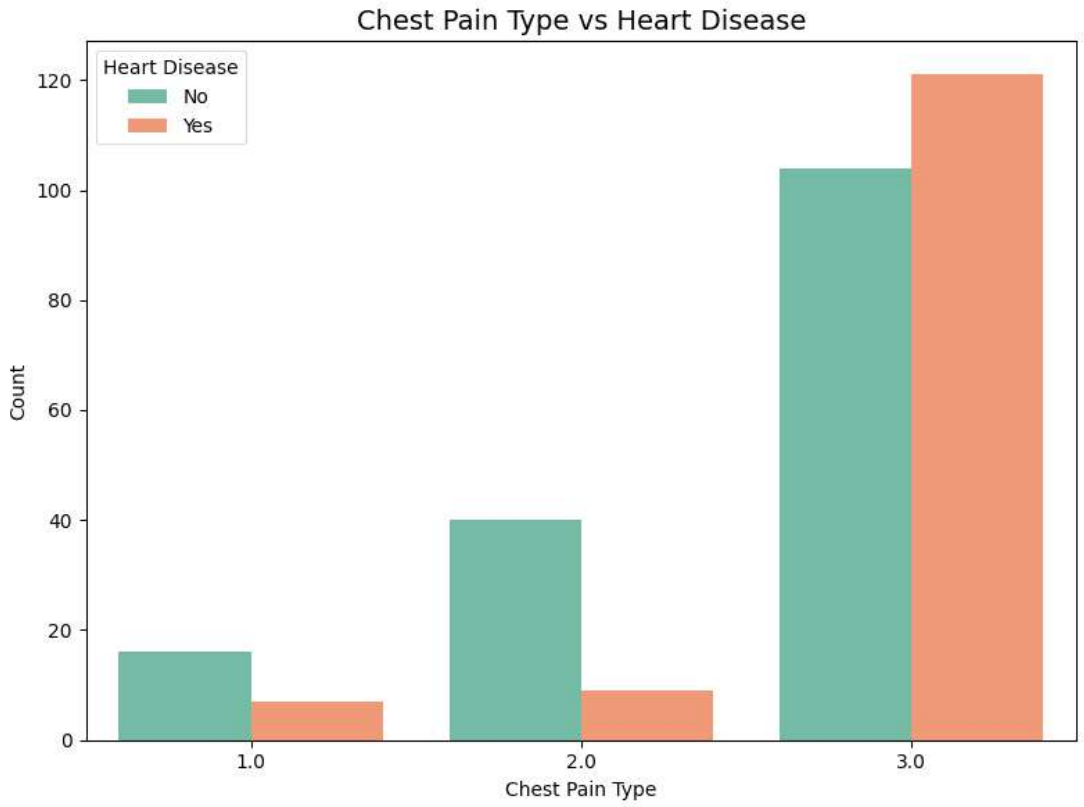
```
plt.xlabel('Chest Pain Type')
plt.ylabel('Count')
plt.title('Chest Pain Type vs Heart Disease', fontsize=14)
plt.legend(title='Heart Disease', labels=['No', 'Yes'])
plt.tight_layout()
plt.savefig('chest_pain_type_vs_heart_disease.png')
plt.show()

"""
This count plot shows the relationship between chest pain type and
the presence of heart disease.
Different types of chest pain are categorized and their occurrence is
compared between patients with and without heart disease.
It appears that certain types of chest pain are more common in
patients with heart disease.
"""

# Plot 4: Cholesterol Levels Box Plot
plt.figure(figsize=(8, 6))
sns.boxplot(x='target', y='chol', data=data_cleaned, palette='Set2',
hue='target')
plt.xlabel('Heart Disease')
plt.ylabel('Cholesterol Levels')
plt.title('Cholesterol Levels for Patients with/without Heart
Disease', fontsize=14)
plt.xticks(ticks=[0,1],labels=['No', 'Yes'])
plt.legend(title='Heart Disease', labels=['No', 'Yes'])
plt.tight_layout()
plt.savefig('cholesterol_levels_box_plot.png')
plt.show()

"""
This box plot compares cholesterol levels between patients with and
without heart disease.
The distribution, median, and outliers of cholesterol levels are
visualized for both groups.
Generally, it appears that patients with heart disease tend to have
higher cholesterol levels.
"""

# Plot 5: Pair plot to visualize relationships between multiple
variables.
pair_plot = sns.pairplot(data_cleaned[['age', 'trestbps', 'chol',
'thalach', 'target']], hue='target', palette='Set2', diag_kind='kde')
pair_plot.savefig('pair_plot.png')
plt.tight_layout()
plt.show()

"""
The pair plot shows the relationships between several variables: age,
resting blood pressure, cholesterol, maximum heart rate, and the
presence of heart disease.
Each scatter plot within the pair plot is colored by the presence of
heart disease, allowing for the visualization of patterns and
relationships between the variables.
This plot helps identify potential correlations and distributions
```

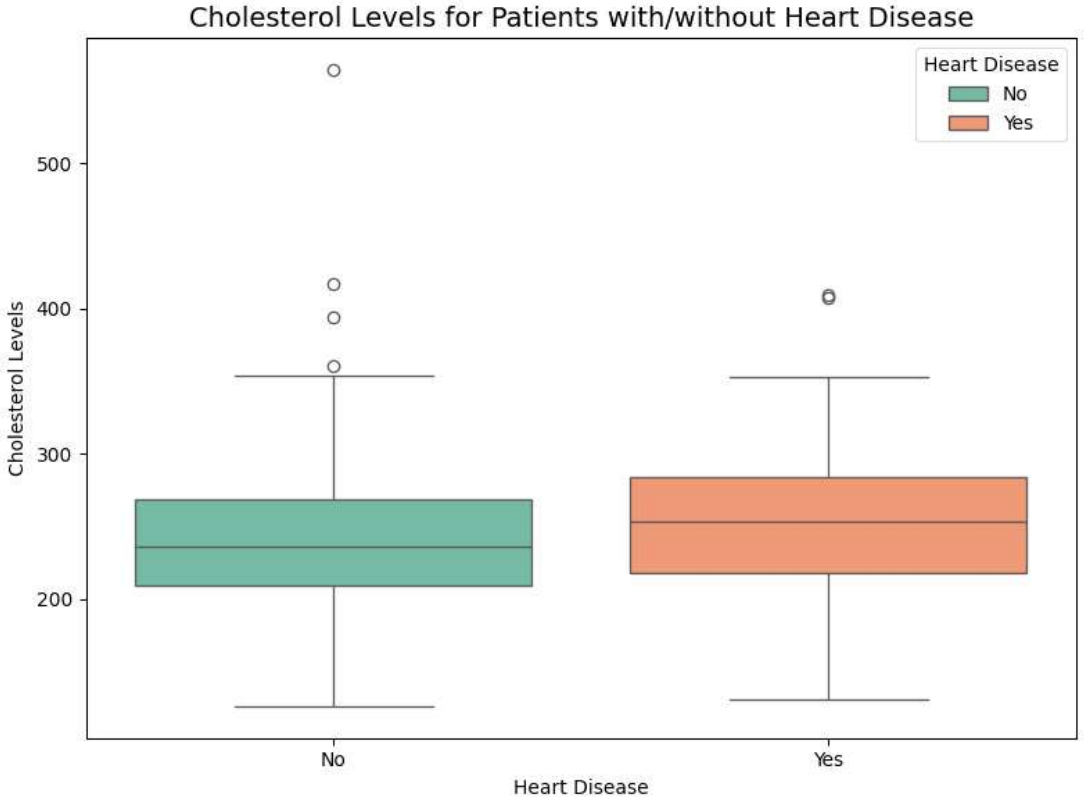
```
among the selected variables.
"""

# Plot 6: Heatmap to visualize the correlation between different
attributes

# Compute the correlation matrix
correlation_matrix = data_cleaned.corr()

plt.figure(figsize=(12, 8))
heatmap = sns.heatmap(correlation_matrix, annot=True,
cmap='coolwarm', fmt='.2f', linewidths=0.5)
plt.title('Correlation Heatmap of Heart Disease Dataset', fontsize=
16)
plt.tight_layout()
plt.savefig('correlation_heatmap.png')
plt.show()

"""
The heatmap displays the correlation between different attributes in
the dataset.
Strong positive and negative correlations are highlighted, which can
help in understanding the relationships between variables.
For example, there might be a strong negative correlation between age
and maximum heart rate.
"""

# Plot 7: Scatter Plot to visualize the relationship between
exercise-induced angina and maximum heart rate
plt.figure(figsize=(12, 8))
plt.scatter(data_cleaned['exang'], data_cleaned['thalach'],
c=data_cleaned['target'], cmap='coolwarm', edgecolor='black', alpha=
0.7)
plt.xlabel('Exercise-Induced Angina (exang)', fontsize=12)
plt.ylabel('Maximum Heart Rate (thalach)', fontsize=12)
plt.title('Relationship between Exercise-Induced Angina and Maximum
Heart Rate', fontsize=16)
plt.tight_layout()
plt.savefig('exercise_induced_angina_vs_max_heart_rate.png')
plt.show()

"""
This scatter plot visualizes the relationship between exercise-
induced angina (exang) and maximum heart rate (thalach).
The points are colored based on the presence of heart disease.
Generally, patients with exercise-induced angina tend to have lower
maximum heart rates, and this relationship is further distinguished
by the presence of heart disease.
"""
```

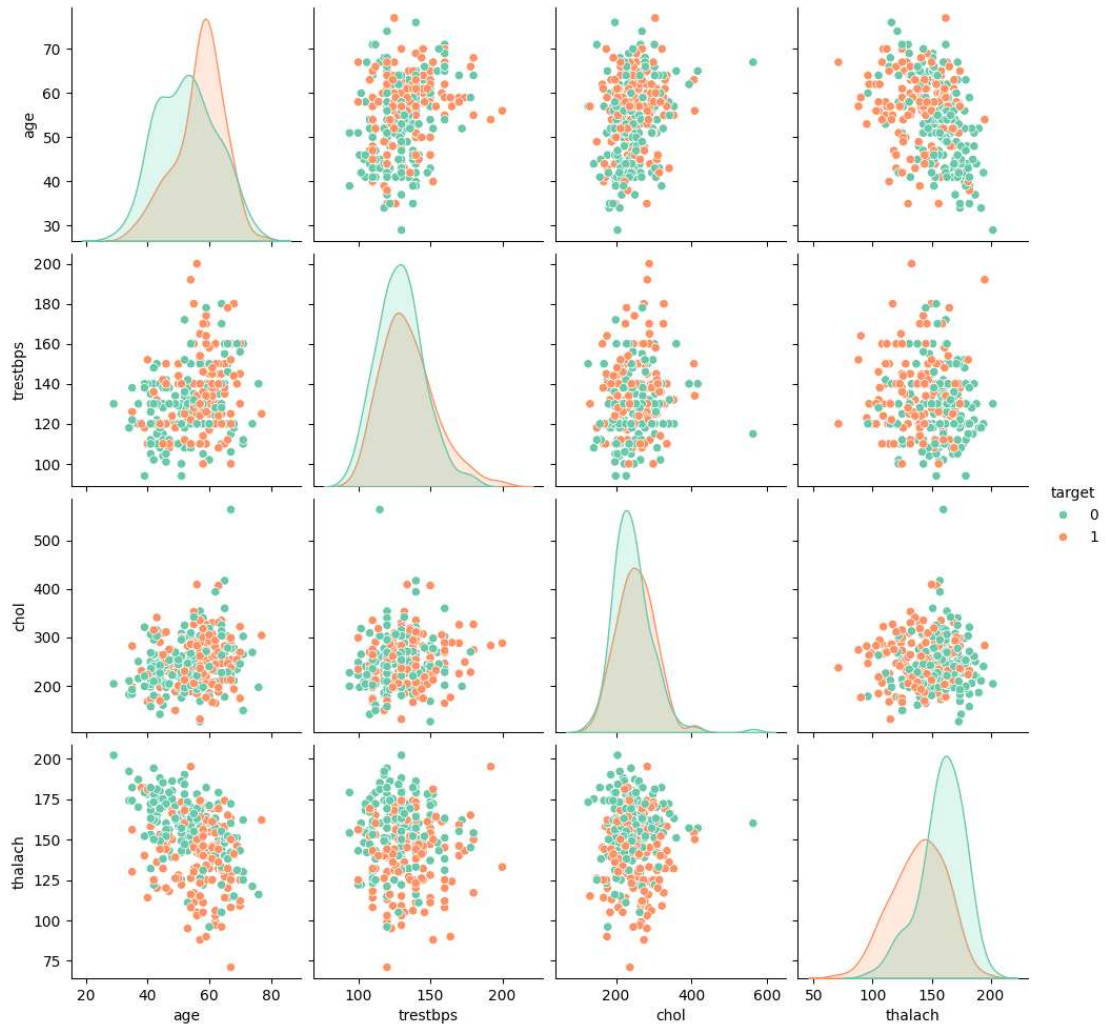
OUTPUT

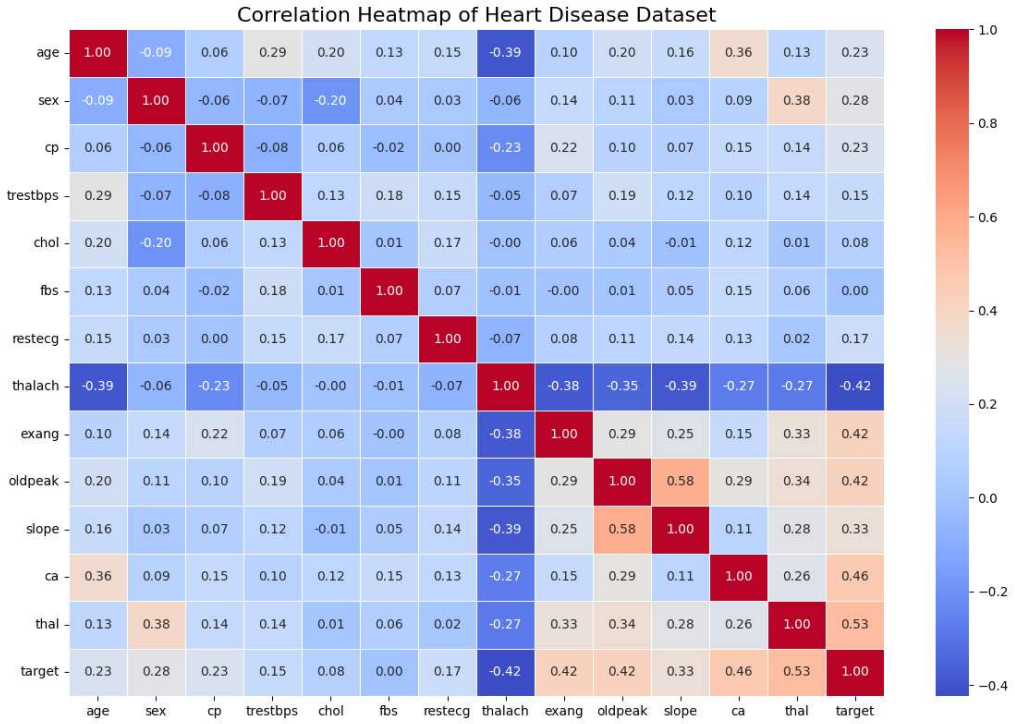

Distribution of Ages
Frequency
Age
0
5
10
15
20
25
30
30
40
50
60
70

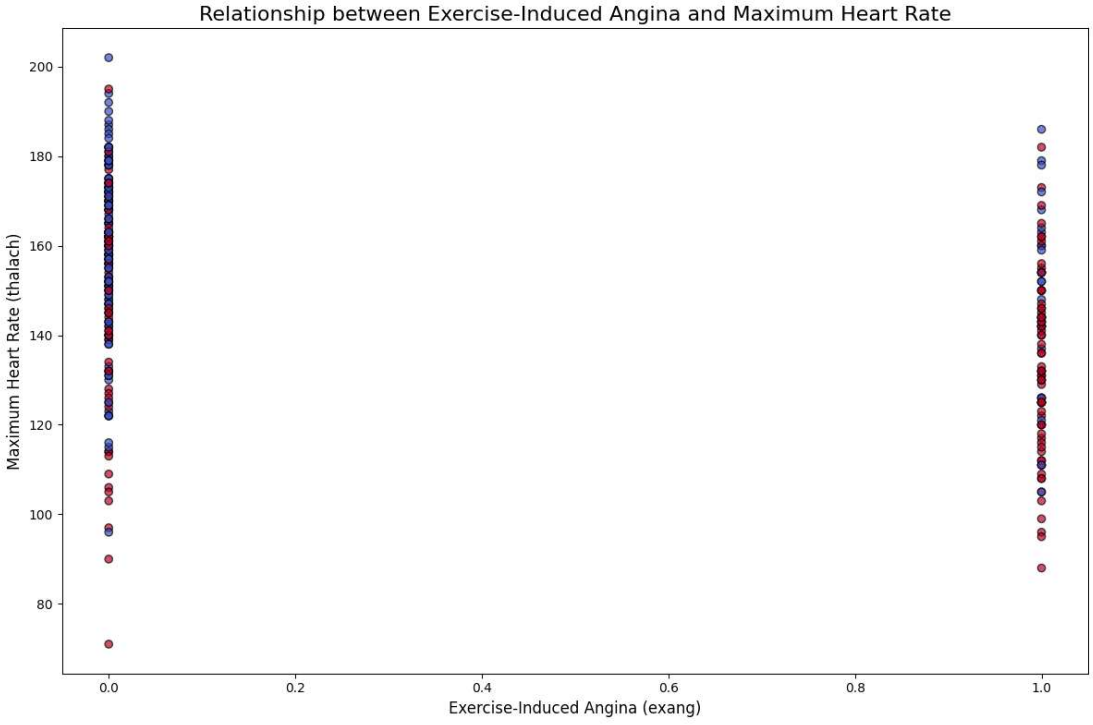

Distribution of Gender
Sex
Female
Male
Count
Gender
0
25
50
75
100
125
150
175
200
0
1


Chest Pain Type vs Heart Disease
Heart Disease
No
Yes
Count
0
20
40
60
80
100
120
1.0
2.0
3.0
Chest Pain Type


Cholesterol Levels for Patients with/without Heart Disease
Heart Disease
No
Yes
Cholesterol Levels
200
300
400
500
No
Yes
Heart Disease


age
trestbps
chol
thalach
target
0
1


Correlation Heatmap of Heart Disease Dataset
age 1.00 -0.09 0.06 0.29 0.20 0.13 0.15 -0.39 0.10 0.20 0.16 0.36 0.13 0.23
sex -0.09 1.00 -0.06 -0.07 -0.20 0.04 0.03 -0.06 0.14 0.11 0.03 0.09 0.38 0.28
cp 0.06 -0.06 1.00 -0.08 0.06 -0.02 0.00 -0.23 0.22 0.10 0.07 0.15 0.14 0.23
trestbps 0.29 -0.07 -0.08 1.00 0.13 0.18 0.15 -0.05 0.07 0.19 0.12 0.10 0.14 0.15
chol 0.20 -0.20 0.06 0.13 1.00 0.01 0.17 -0.00 0.06 0.04 -0.01 0.12 0.01 0.08
fbs 0.13 0.04 -0.02 0.18 0.01 1.00 0.07 -0.01 -0.00 0.01 0.05 0.15 0.06 0.00
restecg 0.15 0.03 0.00 0.15 0.17 0.07 1.00 -0.07 0.08 0.11 0.14 0.13 0.02 0.17
thalach -0.39 -0.06 -0.23 -0.05 -0.00 -0.01 -0.07 1.00 -0.38 -0.35 -0.39 -0.27 -0.27 -0.42
exang 0.10 0.14 0.22 0.07 0.06 -0.00 0.08 -0.38 1.00 0.29 0.25 0.15 0.33 0.42
oldpeak 0.20 0.11 0.10 0.19 0.04 0.01 0.11 -0.35 0.29 1.00 0.58 0.29 0.34 0.42
slope 0.16 0.03 0.07 0.12 -0.01 0.05 0.14 -0.39 0.25 0.58 1.00 0.11 0.28 0.33
ca 0.36 0.09 0.15 0.10 0.12 0.15 0.13 -0.27 0.15 0.29 0.11 1.00 0.26 0.46
thal 0.13 0.38 0.14 0.14 0.01 0.06 0.02 -0.27 0.33 0.34 0.28 0.26 1.00 0.53
target 0.23 0.28 0.23 0.15 0.08 0.00 0.17 -0.42 0.42 0.42 0.33 0.46 0.53 1.00
age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal target
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4


Relationship between Exercise-Induced Angina and Maximum Heart Rate
Maximum Heart Rate (thalach)
200
180
160
140
120
100
80
0.0
0.2
0.4
0.6
0.8
1.0
Exercise-Induced Angina (exang)