heart_disease_eda

June 26, 2025

1 Heart Disease Dataset - Exploratory Data Analysis (EDA)

Author: *Harsh Mishra*
Date: *29 June 2025*
Dataset: Heart Disease Dataset

1.1 Introduction

This notebook presents an in-depth Exploratory Data Analysis (EDA) of a heart disease dataset. The goal is to explore data distribution, detect missing values and outliers, and uncover relationships between features and the target variable.

1.2 1. Load and Inspect the Dataset

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Set style
sns.set(style="whitegrid")
plt.rcParams["figure.figsize"] = (10, 6)

# Load dataset
df = pd.read_csv("heart.csv")
df.head()
```

[1]:

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	\
0	40	M	ATA	140	289	0	Normal	172	
1	49	F	NAP	160	180	0	Normal	156	
2	37	M	ATA	130	283	0	ST	98	
3	48	F	ASY	138	214	0	Normal	108	
4	54	M	NAP	150	195	0	Normal	122	

	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	N	0.0	Up	0
1	N	1.0	Flat	1
2	N	0.0	Up	0

```
3              Y      1.5     Flat             1
4              N      0.0       Up             0
```

1.3 2. Dataset Summary

[2]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Age             918 non-null    int64  
 1   Sex             918 non-null    object 
 2   ChestPainType   918 non-null    object 
 3   RestingBP       918 non-null    int64  
 4   Cholesterol     918 non-null    int64  
 5   FastingBS       918 non-null    int64  
 6   RestingECG      918 non-null    object 
 7   MaxHR           918 non-null    int64  
 8   ExerciseAngina  918 non-null    object 
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    object 
 11  HeartDisease    918 non-null    int64  
dtypes: float64(1), int64(6), object(5)
memory usage: 86.2+ KB
```

[3]:
```
df.describe()
```

[3]:

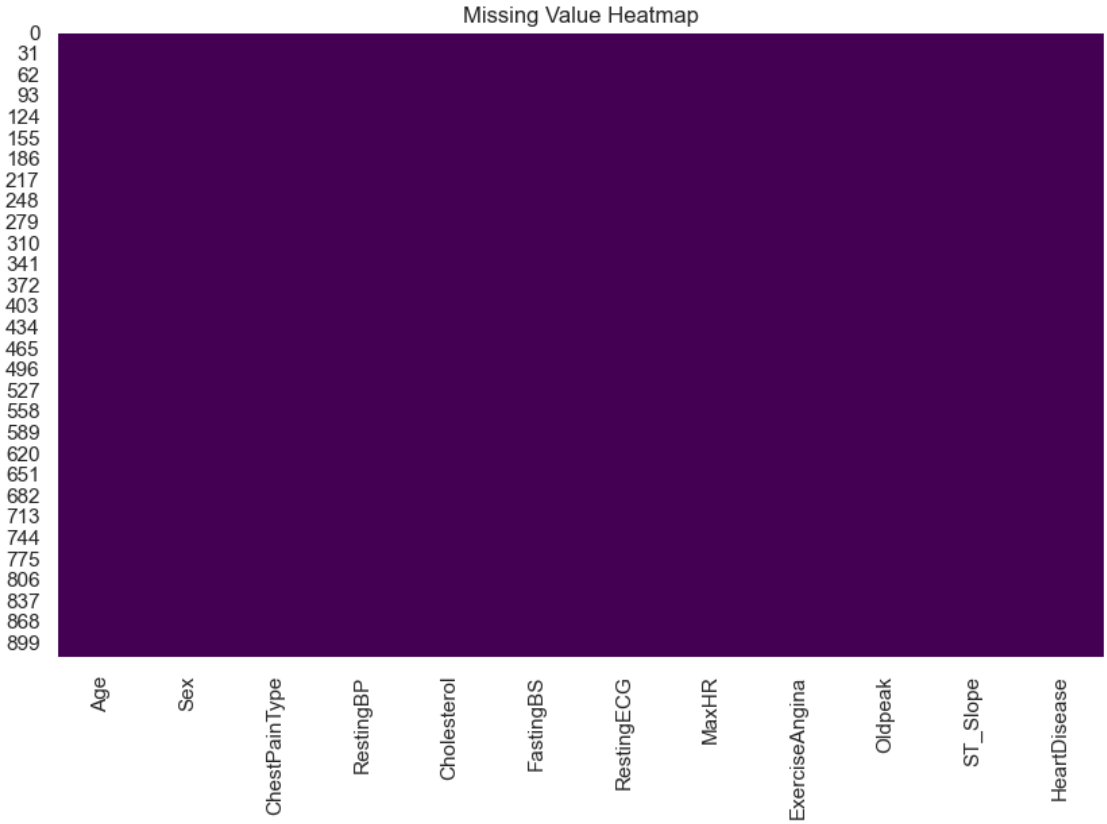
	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
count	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000
mean	53.510893	132.396514	198.799564	0.233115	136.809368	0.887364	0.553377
std	9.432617	18.514154	109.384145	0.423046	25.460334	1.066570	0.497414
min	28.000000	0.000000	0.000000	0.000000	60.000000	-2.600000	0.000000
25%	47.000000	120.000000	173.250000	0.000000	120.000000	0.000000	0.000000
50%	54.000000	130.000000	223.000000	0.000000	138.000000	0.600000	1.000000
75%	60.000000	140.000000	267.000000	0.000000	156.000000	1.500000	1.000000
max	77.000000	200.000000	603.000000	1.000000	202.000000	6.200000	1.000000

1.4 3. Missing Value Analysis

[4]:
```
df.isnull().sum()
```

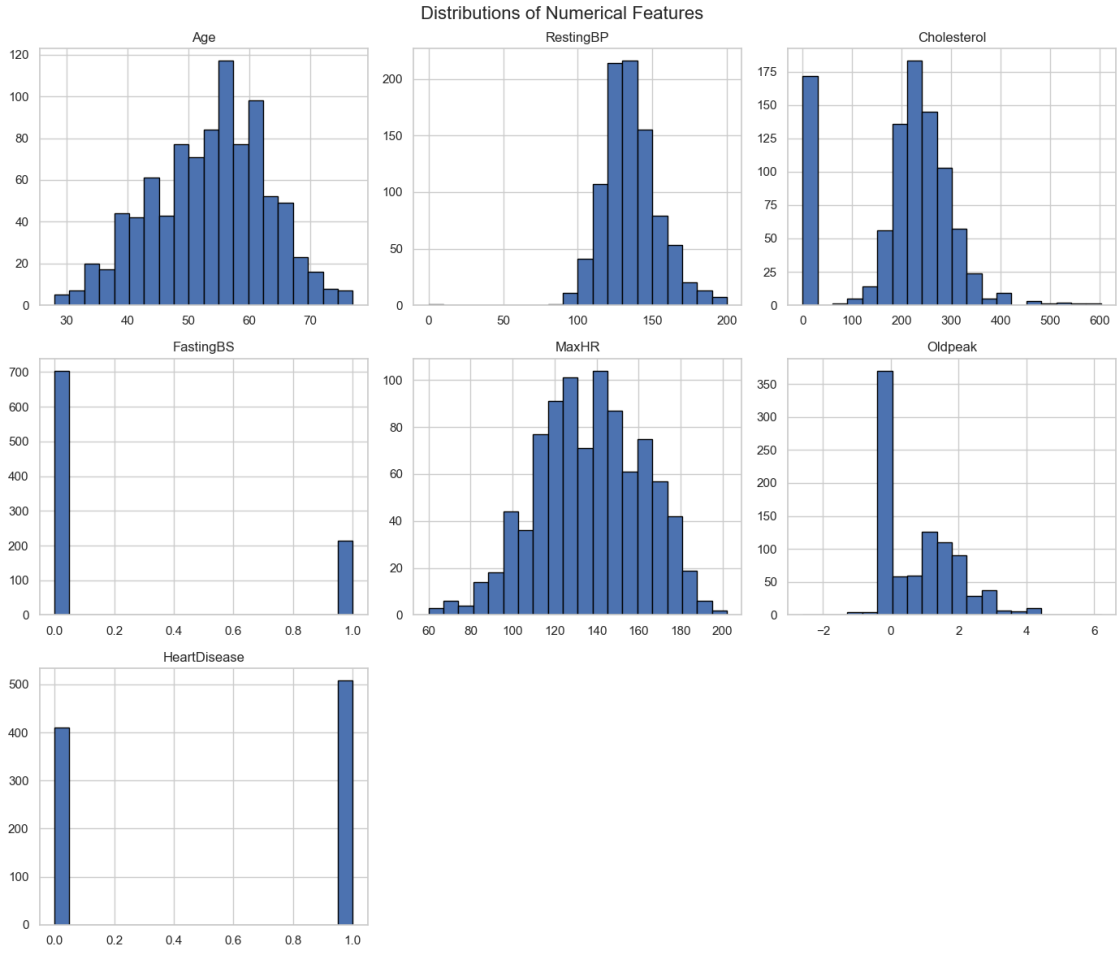
[4]:
```
Age               0
Sex               0
ChestPainType     0
RestingBP         0
Cholesterol       0
FastingBS         0
RestingECG        0
MaxHR             0
ExerciseAngina    0
Oldpeak           0
ST_Slope          0
HeartDisease      0
dtype: int64
```

[5]:
```
sns.heatmap(df.isnull(), cbar=False, cmap='viridis')
plt.title("Missing Value Heatmap")
plt.show()
```

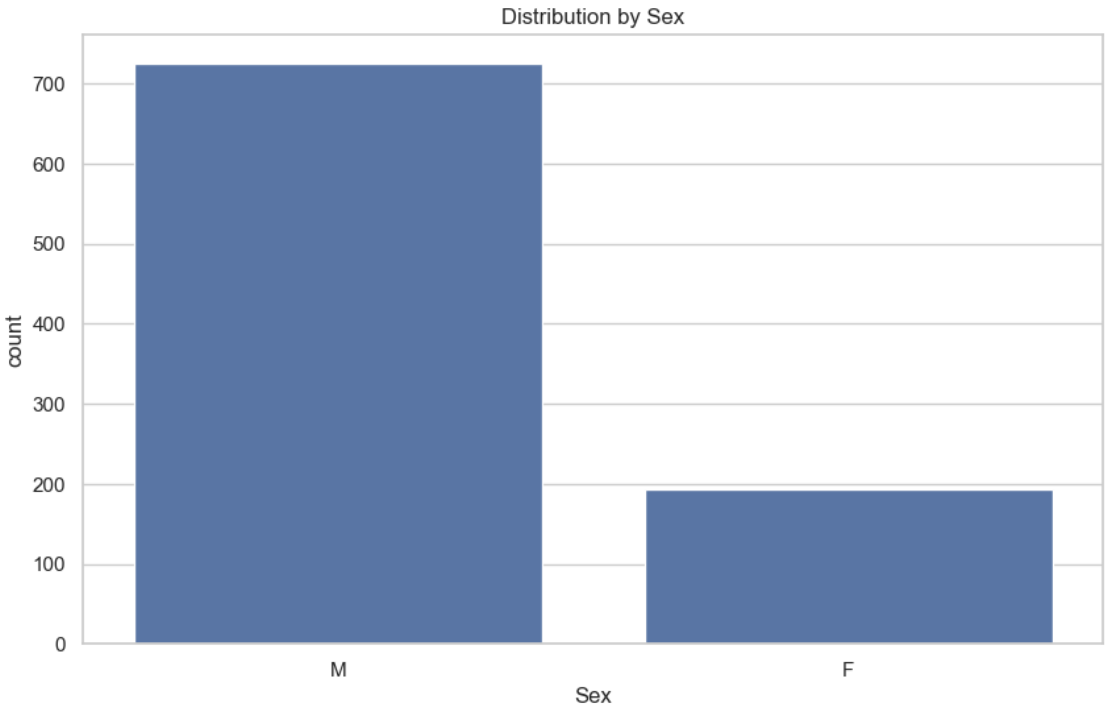


1.5 4. Univariate Analysis

[6]:
```
df.hist(figsize=(14, 12), bins=20, edgecolor='black')
plt.suptitle("Distributions of Numerical Features", fontsize=16)
plt.tight_layout()
plt.show()
```

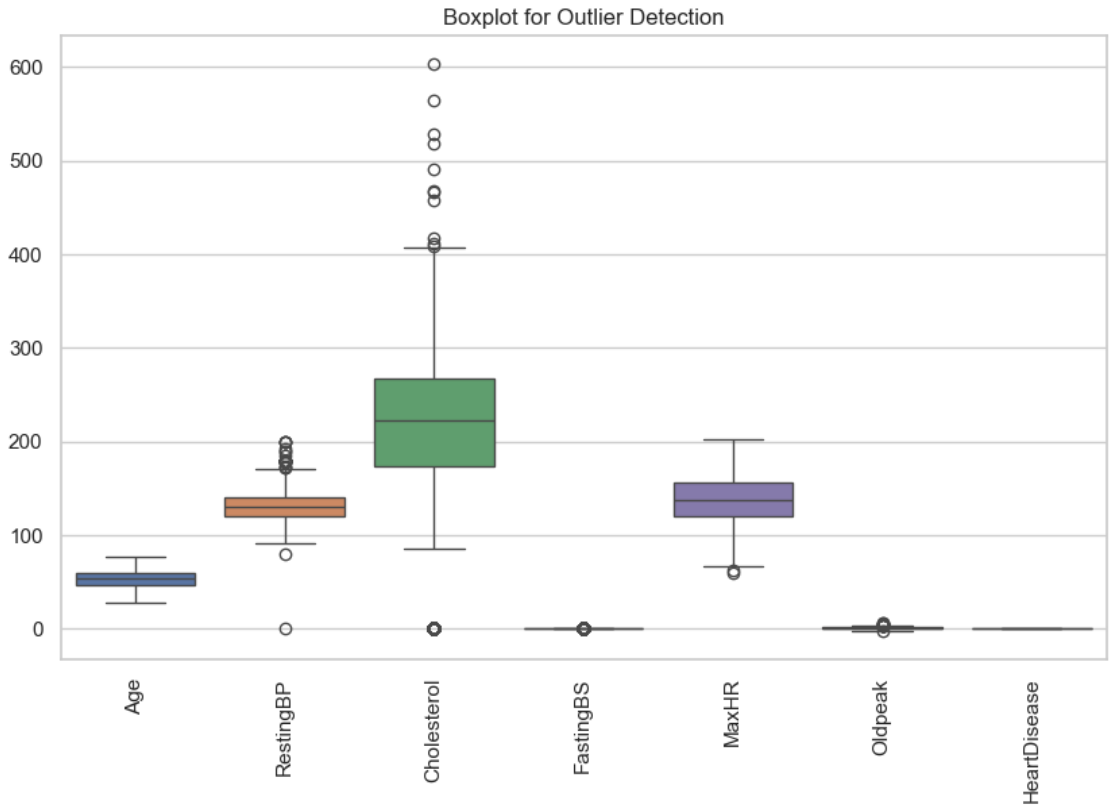


[8]:
```
sns.countplot(x='Sex', data=df)
plt.title("Distribution by Sex")
plt.show()
```

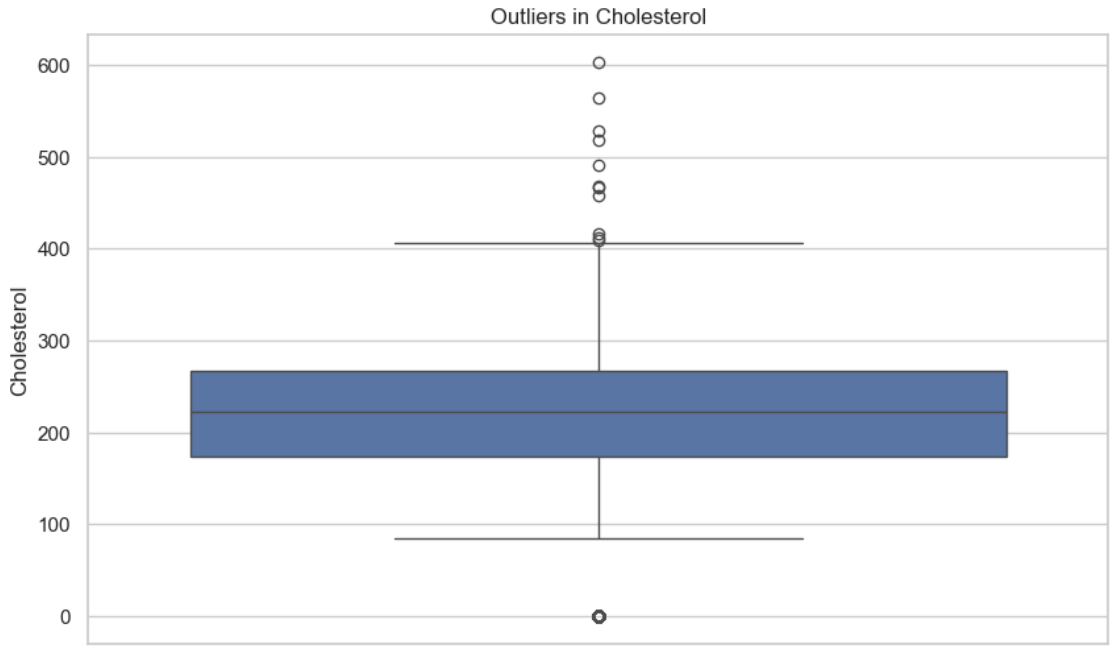


1.6 5. Outlier Detection

[9]:
```
sns.boxplot(data=df)
plt.xticks(rotation=90)
plt.title("Boxplot for Outlier Detection")
plt.show()
```

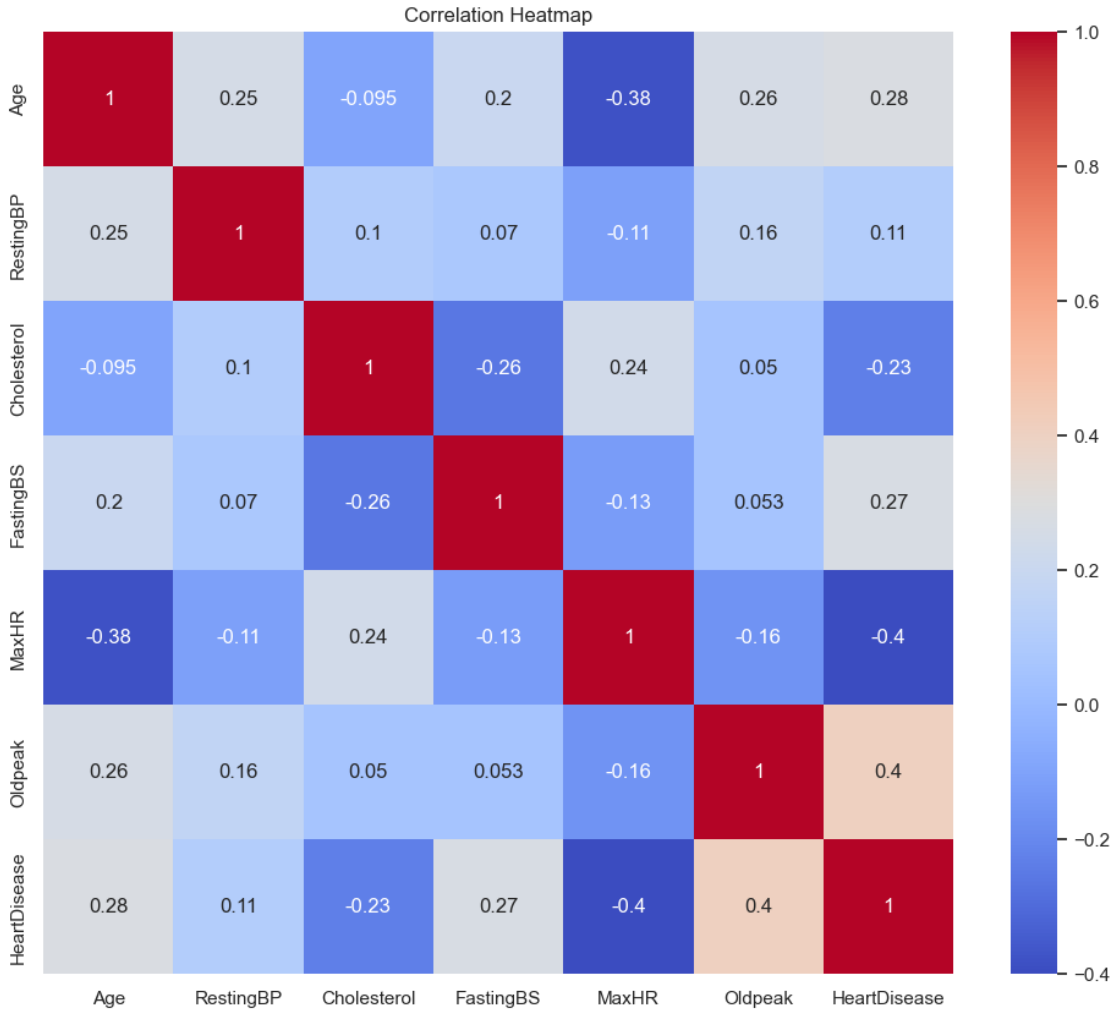


[11]:
```
sns.boxplot(y='Cholesterol', data=df)
plt.title("Outliers in Cholesterol")
plt.show()
```



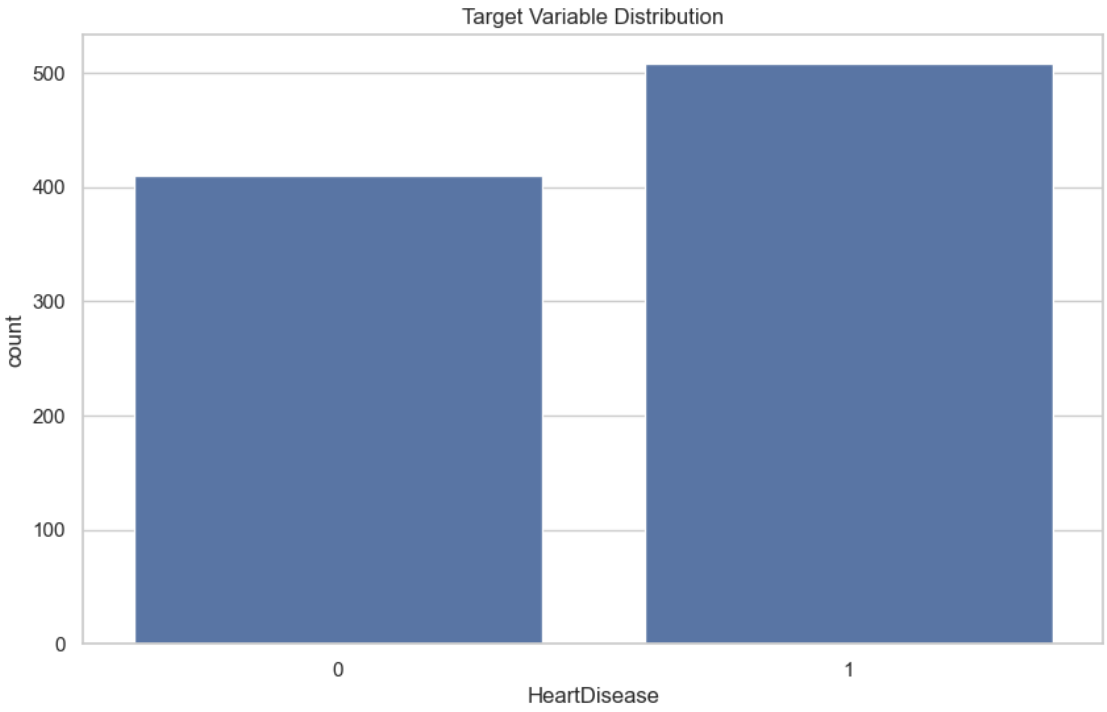
1.7 6. Correlation Analysis

[13]:
```
plt.figure(figsize=(12, 10))
sns.heatmap(df.select_dtypes(include='number').corr(), annot=True,␣
 ↪cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```

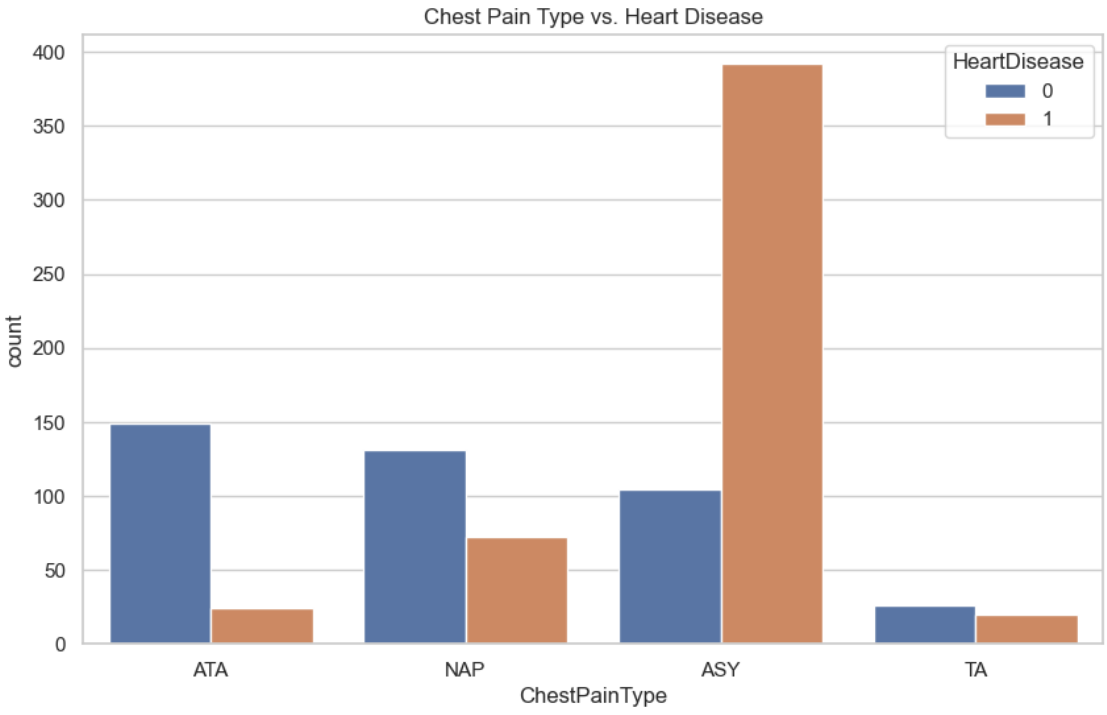


1.8 7. Relationships with Target Variable

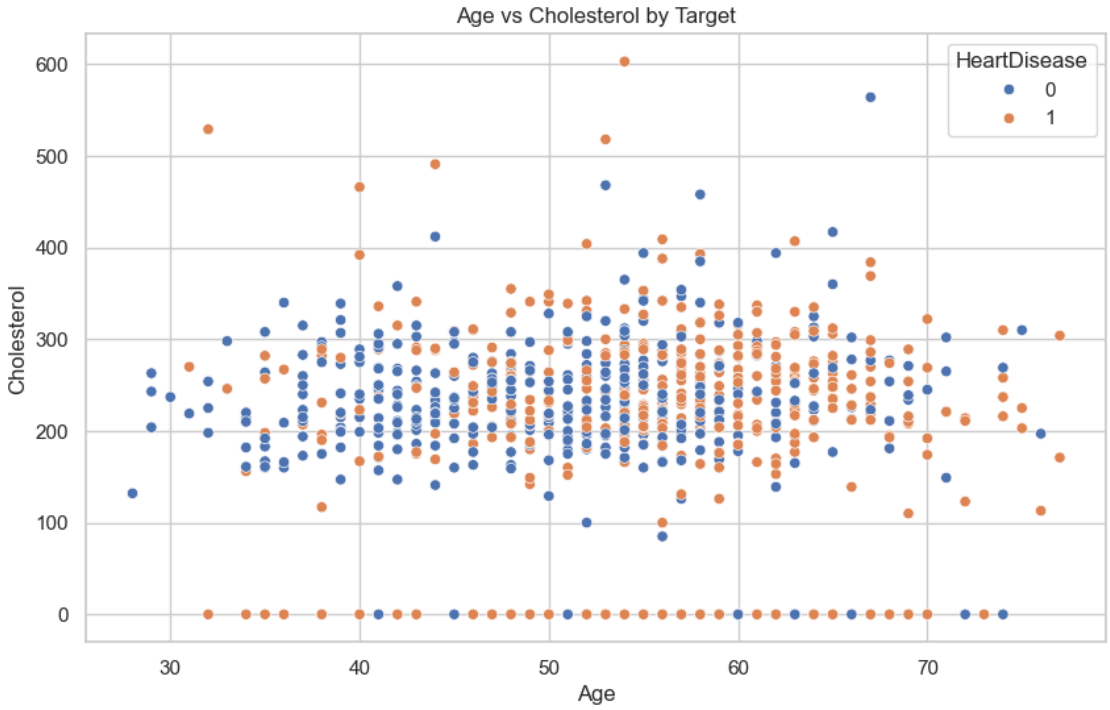
[15]:
```
sns.countplot(x='HeartDisease', data=df)
plt.title('Target Variable Distribution')
plt.show()
```



[16]:
```
sns.countplot(x='ChestPainType', hue='HeartDisease', data=df)
plt.title('Chest Pain Type vs. Heart Disease')
plt.show()
```



[17]:
```
sns.scatterplot(x='Age', y='Cholesterol', hue='HeartDisease', data=df)
plt.title('Age vs Cholesterol by Target')
plt.show()
```



1.9 8. Summary of Findings

- No missing values in the dataset.
- Some outliers exist in features like cholesterol (`chol`) and max heart rate (`thalach`).
- Features like `cp` (chest pain type) and `exang` (exercise-induced angina) show strong correlation with heart disease.
- Correlation heatmap reveals interesting inter-feature relationships worth exploring in modeling.
