Heart Disease Data Explorer: Learning the Basics with heart.csv

Dataset loaded successfully!

Select an analysis view:

1. Intial Data Overview

1. Initial Data Overview

1.1. First 5 Rows (df.head())

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca
0	63	1	3	145	233	1	0	150	0	2.3	0	0
1	37	1	2	130	250	0	1	187	0	3.5	0	0
2	41	0	1	130	204	0	0	172	0	1.4	2	0
3	56	1	1	120	236	0	1	178	0	0.8	2	0
4	57	0	0	120	354	0	1	163	1	0.6	2	0

👉 **What this tells you about** `heart.csv` **:** You see the column names (like 'age', 'sex', 'cp', 'trestbps', 'chol', 'target'), and the first few values. This gives you a quick feel for the type of data in each column.

1.2. Data Dimensions (df.shape)

The dataset has **303 rows** and **14 columns**.

👉 **What this tells you about** `heart.csv` **:** This tells you the total number of patient records (rows) and the number of features/variables (columns) for each patient. For `heart.csv` , it's usually around (303, 14), meaning 303 patients and 14 pieces of information about each.

1.3. Column Information (df.info())

	Data Type	Non-Null Count	Null Count
age	int64	303	0
sex	int64	303	0
cp	int64	303	0
trestbps	int64	303	0
chol	int64	303	0
fbs	int64	303	0
restecg	int64	303	0
thalach	int64	303	0
exang	int64	303	0
oldpeak	float64	303	0

👉 **What this tells you about** `heart.csv` **:**

- **Data Type (** `dtype` **):** Tells you if a column contains numbers (`int64` , `float64`) or text (`object`). For `heart.csv` , most are numbers. `age` , `trestbps` , `chol` , `thalach` , `oldpeak` are usually integers or floats.
- **Non-Null Count:** Shows how many actual data points are in each column. If this number is **less than the total number of rows** (from `df.shape[0]`), it means there are **missing values** in that column. `heart.csv` is usually very clean with no missing values, which is great!

1.4. Descriptive Statistics for Numerical Columns

(heart_disease_df.describe())

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope
count	303	303	303	303	303	303	303	303	303	303	303
mean	54.3663	0.6832	0.967	131.6238	246.264	0.1485	0.5281	149.6469	0.3267	1.0396	1.3993
std	9.0821	0.466	1.0321	17.5381	51.8308	0.3562	0.5259	22.9052	0.4698	1.1611	0.6162
min	29	0	0	94	126	0	0	71	0	0	0
25%	47.5	0	0	120	211	0	0	133.5	0	0	1
50%	55	1	1	130	240	0	1	153	0	0.8	1
75%	61	1	2	140	274.5	0	1	166	1	1.6	2
max	77	1	3	200	564	1	2	202	1	6.2	2

👉 **What this tells you about** `heart.csv` **:** This table summarizes all your *numerical* columns:

- `count` : Same as non-null count.
- `mean` : The average value of the column (e.g., average age, average cholesterol).
- `std` : Standard deviation, how spread out the data is around the mean.
- `min` / `max` : The smallest and largest values. Useful for checking for impossible values (e.g., negative age).
- `25%` , `50%` **(median),** `75%` : Quartiles. They divide the data into quarters. `50%` is the median (middle value). These help understand the distribution without being affected by extreme values.
- **Example for** `age` **:** You can see the minimum age, maximum age, and the average age of patients in this dataset.

Heart Disease Data Explorer: Learning the Basics with heart.csv

Dataset loaded successfully!

Select an analysis view:

2. Understanding Age

2. Understanding the 'age' Column

	age	cp	target
0	63	3	1
1	37	2	1
2	41	1	1
3	56	1	1
4	57	0	1
5	57	0	1
6	56	1	1
7	44	1	1
8	52	2	1
9	57	2	1

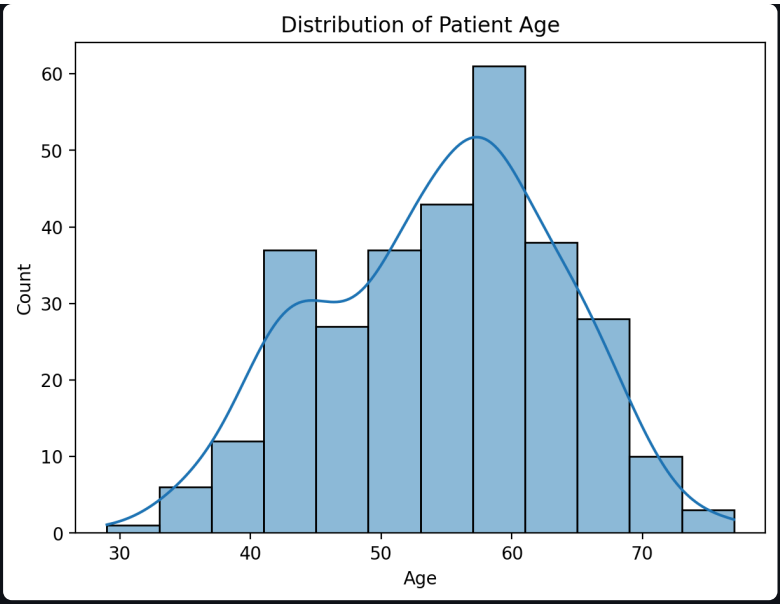
👉 **What this tells you about** `heart.csv` **:** The 'age' column contains the age of each patient. It helps us understand the age distribution of patients with heart disease.

2.1. Basic Statistics for 'age'

	age
count	303
mean	54.3663
std	9.0821
min	29
25%	47.5
50%	55
75%	61
max	77

👉 **What this tells you:** You can quickly see the min, max, average, and spread of ages in your dataset.

2.2. Distribution of 'age' (Histogram & KDE)

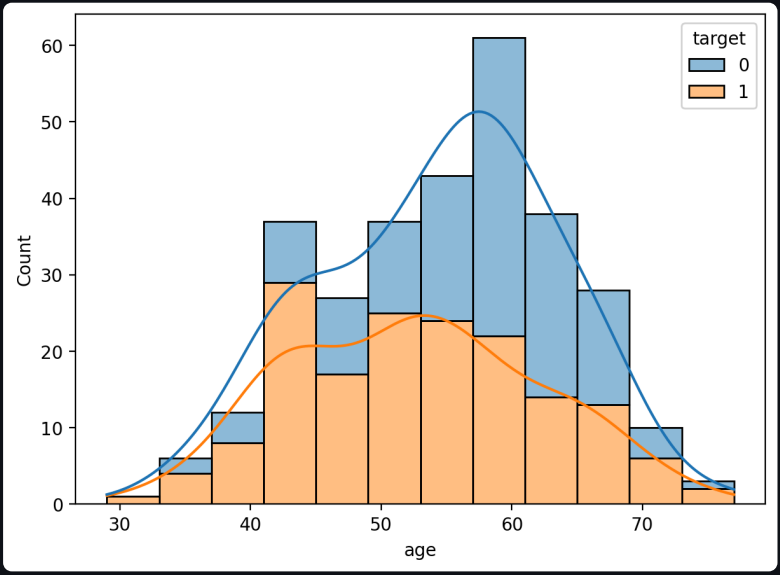


👉 **What this tells you:**

- The **histogram (bars)** shows how many patients fall into different age groups.
- The **KDE line (smooth curve)** provides a smoothed estimate of the distribution.
- For `heart.csv`, you'll likely see that most patients are in their **50s and 60s**, with fewer younger or much older patients.
- This plot helps confirm that 'age' looks like a typical age distribution and doesn't have weird spikes or gaps.

2.3. Age Distribution by Target Variable

2.3.1. Age Distribution by Target Variable (0=No Disease, 1=Disease)



Heart Disease Data Explorer: Learning the Basics with heart.csv

Dataset loaded successfully!

Select an analysis view:

3. Understanding Gender (Sex)

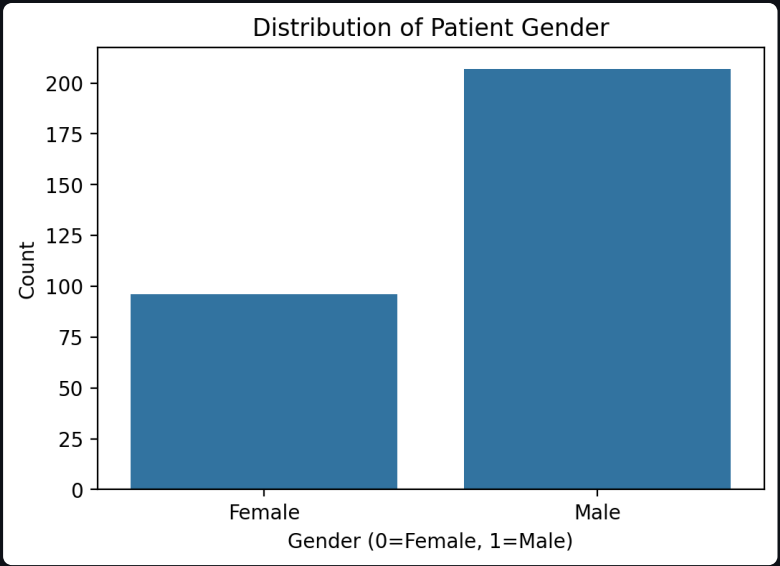
3. Understanding the 'sex' Column

3.1. Unique Values and Counts for 'sex'

sex	count
Male	207
Female	96

👉 **What this tells you:** You can see how many female (0) and male (1) patients are in your dataset. In `heart.csv`, you'll typically find more male patients than female patients.

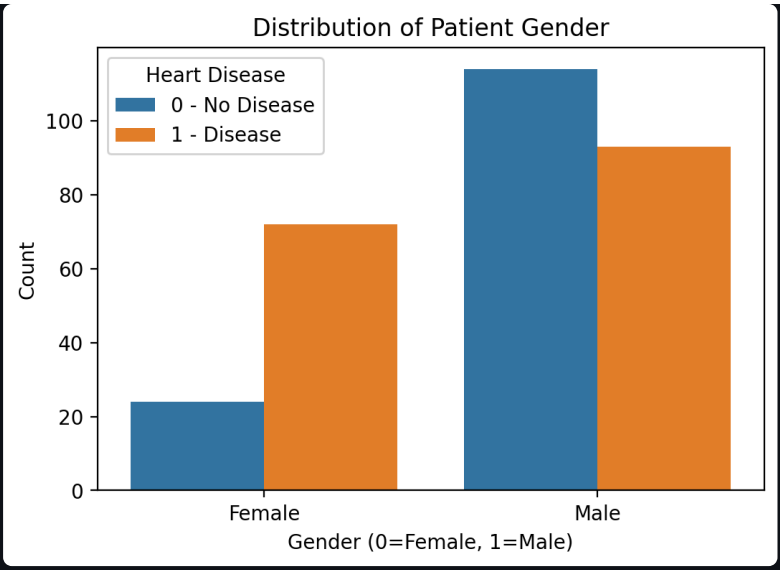
3.2. Distribution of 'sex' (Count Plot)



👉 **What this tells you:**

- The bars visually represent the counts you saw in the table. This confirms the balance (or imbalance) of genders.
- Visuals are often easier to quickly grasp than just numbers.

3.3 Distubtion of sex with respect to target variable



Gender-wise Distribution of Heart Disease (Target)

target	Female	Male
No Disease	24	114
Disease	72	93

👉 **What this tells you:**

- The bars visually represent the counts you saw in the table. This confirms the balance (or imbalance) of genders.
- Visuals are often easier to quickly grasp than just numbers.
- **Insight:** If you see a much higher bar for one gender, it means your dataset is imbalanced for that gender. For example, if there are more males than females, your analysis and conclusions may be more representative of male patients. This is important to keep in mind when interpreting results or building predictive models.

Heart Disease Data Explorer: Learning the Basics with heart.csv

Dataset loaded successfully!

Select an analysis view:

4. Understanding Target (Heart Disease Status)

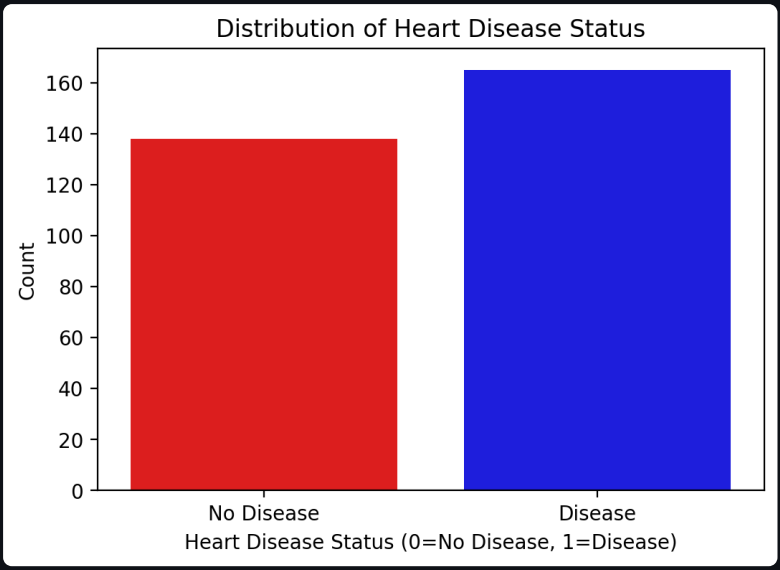
4. Understanding the 'target' Column

4.1. Unique Values and Counts for 'target'

target	count
Disease	165
No Disea	138

👉 **What this tells you:** This is the most crucial column! `0` typically means the patient **does NOT have heart disease**, and `1` means they **DO have heart disease**. This table shows how many patients fall into each category. In `heart.csv`, the counts are usually quite balanced, which is good for machine learning models.

4.2. Distribution of 'target' (Count Plot)



👉 **What this tells you:**

- This plot visually confirms the balance of your main prediction outcome.
- If one category was much, much smaller, it would be an 'imbalanced dataset', which can be tricky for models.

Heart Disease Data Explorer: Learning the Basics with heart.csv

Dataset loaded successfully!

Select an analysis view:

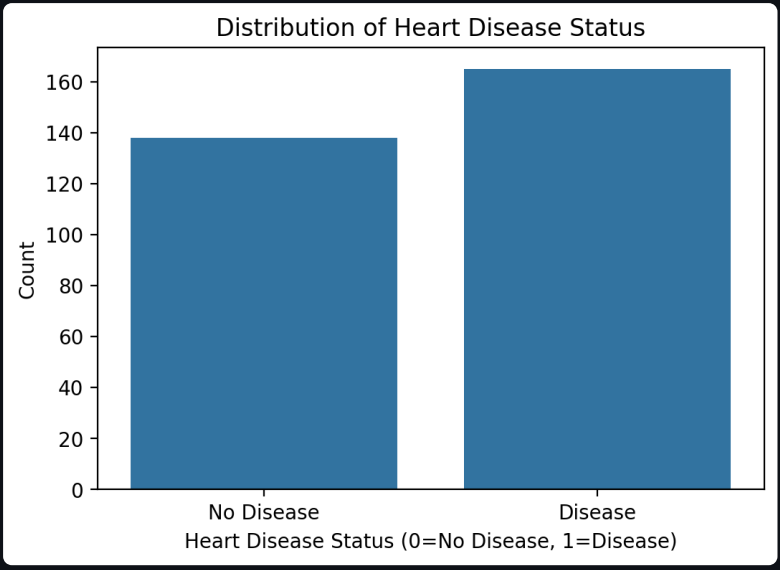
5. Understanding Cholesterol Levels

5.1. Understanding the 'chol' Column

	chol	target
0	233	1
1	250	1
2	204	1
3	236	1
4	354	1
5	192	1
6	294	1
7	263	1
8	199	1
9	168	1

👉 **What this tells you about** `heart.csv` **:** The 'chol' column contains the cholesterol levels of each patient. It helps us understand the cholesterol distribution among patients with heart disease.

5.2. Distribution of 'target' (Count Plot)



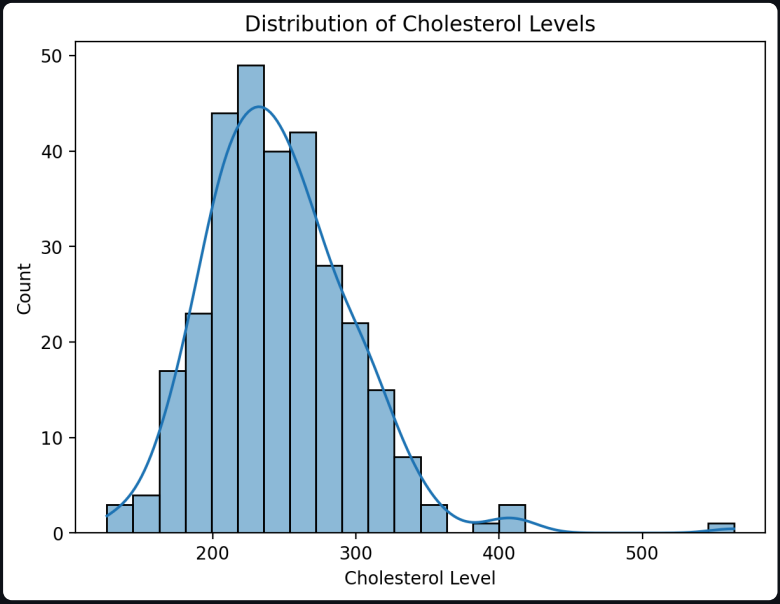
👉 **What this tells you:** This plot visually confirms the balance of your main prediction outcome.

5.3. Basic Statistics for 'chol'

	chol
count	303
mean	246.264
std	51.8308
min	126
25%	211
50%	240
75%	274.5
max	564

👉 **What this tells you:** You can quickly see the min, max, average, and spread of cholesterol levels in your dataset.

5.4. Distribution of 'chol' (Histogram & KDE)

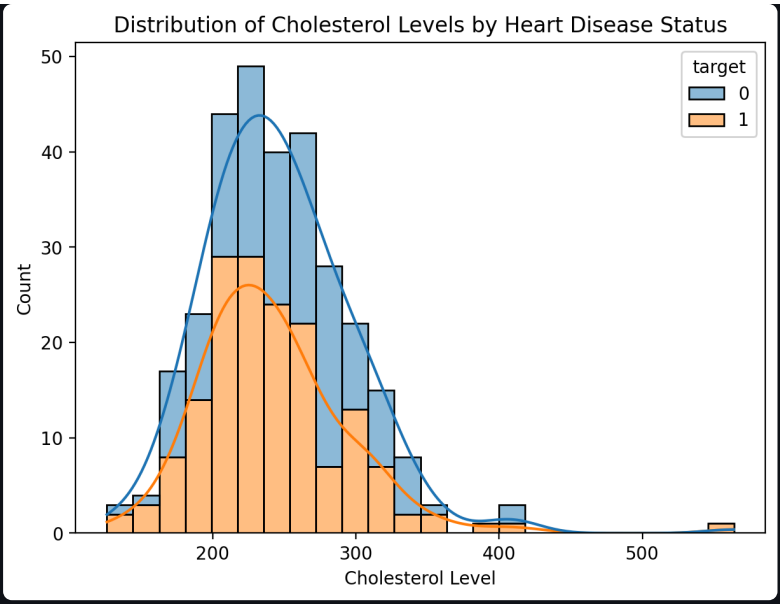


👉 **What this tells you:**

- The **histogram (bars)** shows how many patients fall into different cholesterol level ranges.
- The **KDE line (smooth curve)** provides a smoothed estimate of the distribution.
- For `heart.csv`, you'll likely see that most patients have cholesterol levels between **150 and 300 mg/dl**, with fewer patients having very high or very low levels.

5.5. Cholesterol Levels by Target Variable

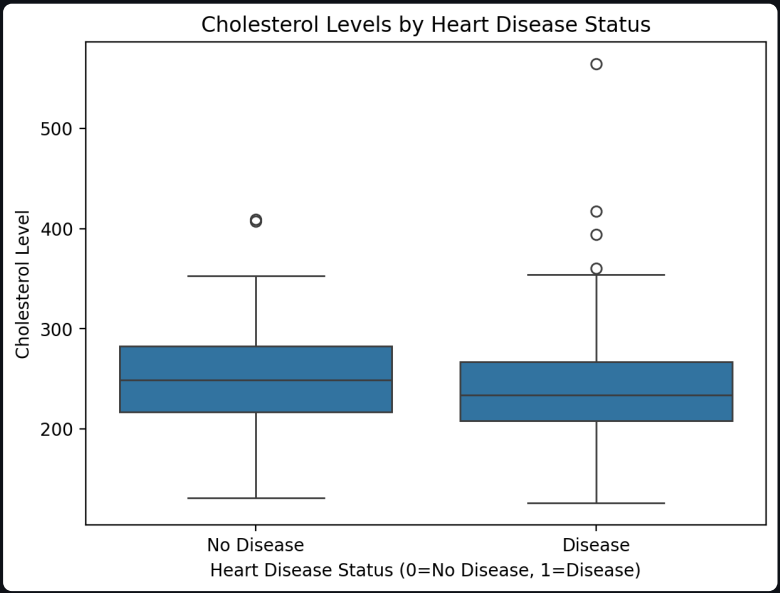
5.5.1. Cholesterol Levels by Target Variable (0=No Disease, 1=Disease)



👉 **What this tells you:**

- This plot shows how cholesterol levels differ between patients with and without heart disease.
- You can see if patients with heart disease tend to have higher or lower cholesterol levels compared to those without.

5.6. Cholesterol Levels by Target Variable (Box Plot)



👉 **What this tells you:**

- The box plot shows the distribution of cholesterol levels for patients with and without heart disease.
- You can see the median cholesterol level, the interquartile range (IQR), and any potential outliers.

Heart Disease Data Explorer: Learning the Basics with heart.csv

Dataset loaded successfully!

Select an analysis view:

7. Understanding Chest Pain Type

7. Understanding the 'cp' Column

	cp	target
0	3	1
1	2	1
2	1	1
3	1	1
4	0	1
5	0	1
6	1	1
7	1	1
8	2	1
9	2	1

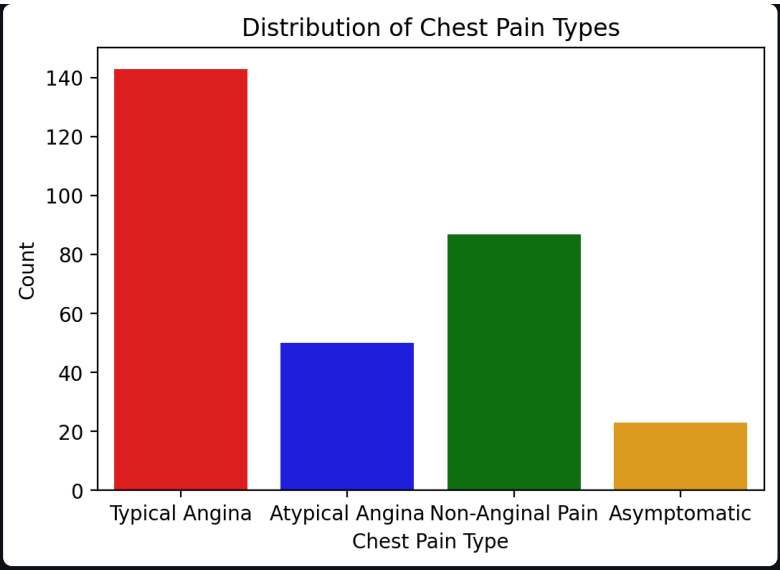
👉 **What this tells you about** `heart.csv` **:** The 'cp' column contains the chest pain type experienced by each patient. It helps us understand the distribution of chest pain types among patients with heart disease.

7.1. Unique Values and Counts for 'cp'

cp	count
Typical Angina	143
Non-Anginal Pain	87
Atypical Angina	50
Asymptomatic	23

👉 **What this tells you:** This table shows how many patients fall into each chest pain type category. In `heart.csv`, you'll typically find a mix of chest pain types, which is important for understanding heart disease symptoms.

7.2. Distribution of 'cp' (Count Plot)

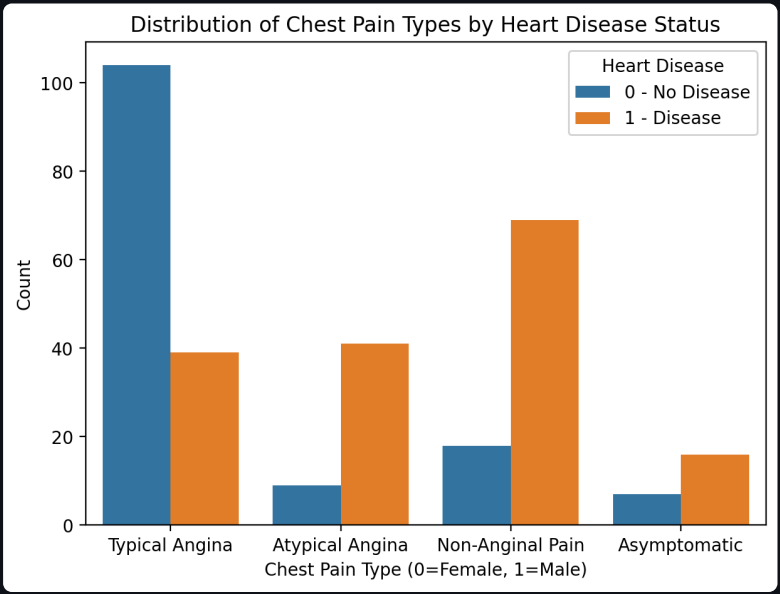


👉 **What this tells you:**

- This plot visually confirms the distribution of chest pain types.
- You can see which chest pain types are most common among patients with heart disease.

7.3. Chest Pain Types by Target Variable

7.3.1. Chest Pain Types by Target Variable (0=No Disease, 1=Disease)



Heart Disease Data Explorer: Learning the Basics with heart.csv

Dataset loaded successfully!

Select an analysis view:

8. Understanding talach

8. Understanding the 'thalach' Column

	thalach	target
0	150	1
1	187	1
2	172	1
3	178	1
4	163	1
5	148	1
6	153	1
7	173	1
8	162	1
9	174	1

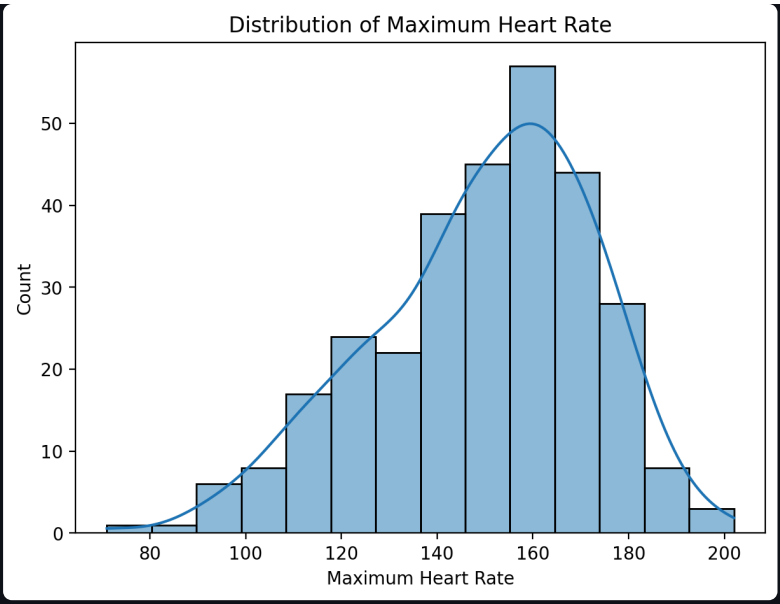
👉 **What this tells you about** `heart.csv` **:** The 'thalach' column contains the maximum heart rate achieved by each patient. It helps us understand the distribution of heart rates among patients with heart disease.

8.1. Basic Statistics for 'thalach'

	thalach
count	303
mean	149.6469
std	22.9052
min	71
25%	133.5
50%	153
75%	166
max	202

👉 **What this tells you:** You can quickly see the min, max, average, and spread of heart rates in your dataset.

8.2. Distribution of 'thalach' (Histogram & KDE)

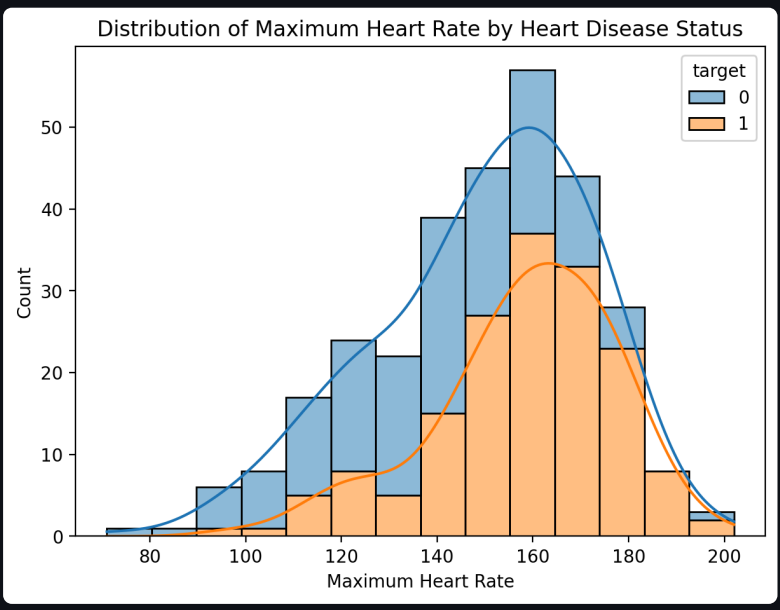


👉 **What this tells you:**

- The **histogram (bars)** shows how many patients fall into different heart rate ranges.
- The **KDE line (smooth curve)** provides a smoothed estimate of the distribution.
- For `heart.csv`, you'll likely see that most patients have maximum heart rates between **100 and 200 bpm**, with fewer patients having very high or very low heart rates.

8.3. Maximum Heart Rate by Target Variable

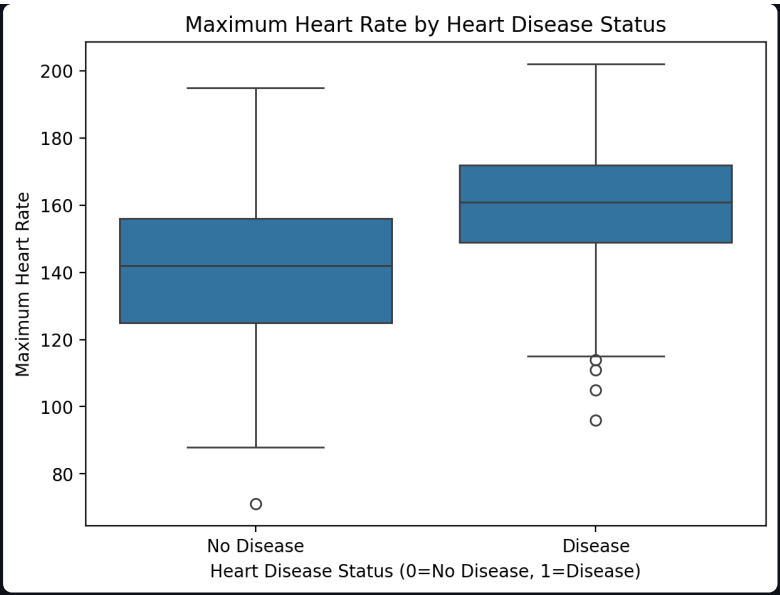
8.3.1. Maximum Heart Rate by Target Variable (0=No Disease, 1=Disease)



👉 **What this tells you:**

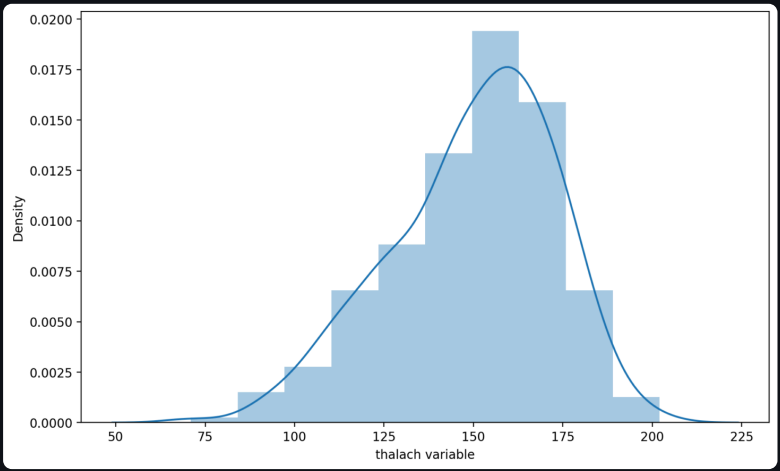
- This plot shows how maximum heart rates differ between patients with and without heart disease.
- You can see if patients with heart disease tend to have higher or lower maximum heart rates compared to those without.

8.4. Maximum Heart Rate by Target Variable (Box Plot)



👉 **What this tells you:**

- The box plot shows the distribution of maximum heart rates for patients with and without heart disease.
- You can see the median maximum heart rate, the interquartile range (IQR), and any potential outliers.



Heart Disease Data Explorer: Learning the Basics with heart.csv

Dataset loaded successfully!

Select an analysis view:

10. Pair Plot

10. Pair Plot of Numerical Variables

10.1. Pair Plot

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slop
age	1	-0.0984	-0.0687	0.2794	0.2137	0.1213	-0.1162	-0.3985	0.0968	0.21	-0.1
sex	-0.0984	1	-0.0494	-0.0568	-0.1979	0.045	-0.0582	-0.044	0.1417	0.0961	-0.0
cp	-0.0687	-0.0494	1	0.0476	-0.0769	0.0944	0.0444	0.2958	-0.3943	-0.1492	0.1
trestbps	0.2794	-0.0568	0.0476	1	0.1232	0.1775	-0.1141	-0.0467	0.0676	0.1932	-0.1
chol	0.2137	-0.1979	-0.0769	0.1232	1	0.0133	-0.151	-0.0099	0.067	0.054	-0.0
fbs	0.1213	0.045	0.0944	0.1775	0.0133	1	-0.0842	-0.0086	0.0257	0.0057	-0.0
restecg	-0.1162	-0.0582	0.0444	-0.1141	-0.151	-0.0842	1	0.0441	-0.0707	-0.0588	0.0
thalach	-0.3985	-0.044	0.2958	-0.0467	-0.0099	-0.0086	0.0441	1	-0.3788	-0.3442	0.3
exang	0.0968	0.1417	-0.3943	0.0676	0.067	0.0257	-0.0707	-0.3788	1	0.2882	-0.2
oldpeak	0.21	0.0961	-0.1492	0.1932	0.054	0.0057	-0.0588	-0.3442	0.2882	1	-0.5

👉 **What this tells you:** The correlation matrix shows how strongly each pair of variables is related. Values close to 1 or -1 indicate strong relationships, while values near 0 indicate weak relationships.

9.2. Heat Map of Correlations


age 1.00 -0.10 -0.07 0.28 0.21 0.12 -0.12 -0.40 0.10 0.21 -0.17 0.28 0.07 -0.23
sex -0.10 1.00 -0.05 -0.06 -0.20 0.05 -0.06 -0.04 0.14 0.10 -0.03 0.12 0.21 -0.28
cp -0.07 -0.05 1.00 0.05 -0.08 0.09 0.04 0.30 -0.39 -0.15 0.12 -0.18 -0.16 0.43
trestbps 0.28 -0.06 0.05 1.00 0.12 0.18 -0.11 -0.05 0.07 0.19 -0.12 0.10 0.06 -0.14
chol 0.21 -0.20 -0.08 0.12 1.00 0.01 -0.15 -0.01 0.07 0.05 -0.00 0.07 0.10 -0.09
fbs 0.12 0.05 0.09 0.18 0.01 1.00 -0.08 -0.01 0.03 0.01 -0.06 0.14 -0.03 -0.03
restecg -0.12 -0.06 0.04 -0.11 -0.15 -0.08 1.00 0.04 -0.07 -0.06 0.09 -0.07 -0.01 0.14
thalach -0.40 -0.04 0.30 -0.05 -0.01 -0.01 0.04 1.00 -0.38 -0.34 0.39 -0.21 -0.10 0.42
exang 0.10 0.14 -0.39 0.07 0.07 0.03 -0.07 -0.38 1.00 0.29 -0.26 0.12 0.21 -0.44
oldpeak 0.21 0.10 -0.15 0.19 0.05 0.01 -0.06 -0.34 0.29 1.00 -0.58 0.22 0.21 -0.43
slope -0.17 -0.03 0.12 -0.12 -0.00 -0.06 0.09 0.39 -0.26 -0.58 1.00 -0.08 -0.10 0.35
ca 0.28 0.12 -0.18 0.10 0.07 0.14 -0.07 -0.21 0.12 0.22 -0.08 1.00 0.15 -0.39
thal 0.07 0.21 -0.16 0.06 0.10 -0.03 -0.01 -0.10 0.21 0.21 -0.10 0.15 1.00 -0.34
target -0.23 -0.28 0.43 -0.14 -0.09 -0.03 0.14 0.42 -0.44 -0.43 0.35 -0.39 -0.34 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

Heart Disease Data Explorer: Learning the Basics with heart.csv

Dataset loaded successfully!

Select an analysis view:

9. Heat Map

9. Heat Map of Correlations

9.1. Correlation Matrix

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slop
age	1	-0.0984	-0.0687	0.2794	0.2137	0.1213	-0.1162	-0.3985	0.0968	0.21	-0.16
sex	-0.0984	1	-0.0494	-0.0568	-0.1979	0.045	-0.0582	-0.044	0.1417	0.0961	-0.03
cp	-0.0687	-0.0494	1	0.0476	-0.0769	0.0944	0.0444	0.2958	-0.3943	-0.1492	0.11
trestbps	0.2794	-0.0568	0.0476	1	0.1232	0.1775	-0.1141	-0.0467	0.0676	0.1932	-0.12
chol	0.2137	-0.1979	-0.0769	0.1232	1	0.0133	-0.151	-0.0099	0.067	0.054	-0.0
fbs	0.1213	0.045	0.0944	0.1775	0.0133	1	-0.0842	-0.0086	0.0257	0.0057	-0.05
restecg	-0.1162	-0.0582	0.0444	-0.1141	-0.151	-0.0842	1	0.0441	-0.0707	-0.0588	0.0
thalach	-0.3985	-0.044	0.2958	-0.0467	-0.0099	-0.0086	0.0441	1	-0.3788	-0.3442	0.38
exang	0.0968	0.1417	-0.3943	0.0676	0.067	0.0257	-0.0707	-0.3788	1	0.2882	-0.25
oldpeak	0.21	0.0961	-0.1492	0.1932	0.054	0.0057	-0.0588	-0.3442	0.2882	1	-0.57

👉 **What this tells you:** The correlation matrix shows how strongly each pair of variables is related. Values close to 1 or -1 indicate strong relationships, while values near 0 indicate weak relationships.

9.2. Heat Map of Correlations


age 1.00 -0.10 -0.07 0.28 0.21 0.12 -0.12 -0.40 0.10 0.21 -0.17 0.28 0.07 -0.23
sex -0.10 1.00 -0.05 -0.06 -0.20 0.05 -0.06 -0.04 0.14 0.10 -0.03 0.12 0.21 -0.28
cp -0.07 -0.05 1.00 0.05 -0.08 0.09 0.04 0.30 -0.39 -0.15 0.12 -0.18 -0.16 0.43
trestbps 0.28 -0.06 0.05 1.00 0.12 0.18 -0.11 -0.05 0.07 0.19 -0.12 0.10 0.06 -0.14
chol 0.21 -0.20 -0.08 0.12 1.00 0.01 -0.15 -0.01 0.07 0.05 -0.00 0.07 0.10 -0.09
fbs 0.12 0.05 0.09 0.18 0.01 1.00 -0.08 -0.01 0.03 0.01 -0.06 0.14 -0.03 -0.03
restecg -0.12 -0.06 0.04 -0.11 -0.15 -0.08 1.00 0.04 -0.07 -0.06 0.09 -0.07 -0.01 0.14
thalach -0.40 -0.04 0.30 -0.05 -0.01 -0.01 0.04 1.00 -0.38 -0.34 0.39 -0.21 -0.10 0.42
exang 0.10 0.14 -0.39 0.07 0.07 0.03 -0.07 -0.38 1.00 0.29 -0.26 0.12 0.21 -0.44
oldpeak 0.21 0.10 -0.15 0.19 0.05 0.01 -0.06 -0.34 0.29 1.00 -0.58 0.22 0.21 -0.43
slope -0.17 -0.03 0.12 -0.12 -0.00 -0.06 0.09 0.39 -0.26 -0.58 1.00 -0.08 -0.10 0.35
ca 0.28 0.12 -0.18 0.10 0.07 0.14 -0.07 -0.21 0.12 0.22 -0.08 1.00 0.15 -0.39
thal 0.07 0.21 -0.16 0.06 0.10 -0.03 -0.01 -0.10 0.21 0.21 -0.10 0.15 1.00 -0.34
target -0.23 -0.28 0.43 -0.14 -0.09 -0.03 0.14 0.42 -0.44 -0.43 0.35 -0.39 -0.34 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

Heart Disease Data Explorer: Learning the Basics with heart.csv

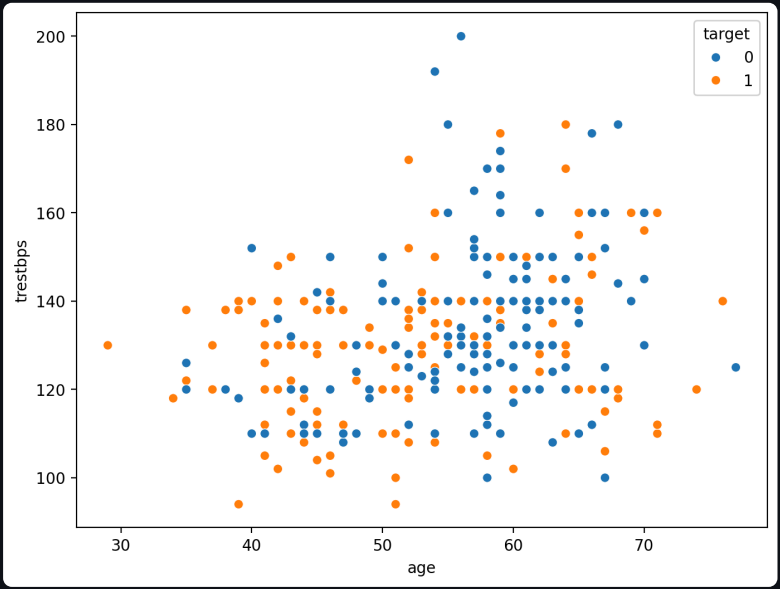
Dataset loaded successfully!

Select an analysis view:

11. Miscellaneous Analysis

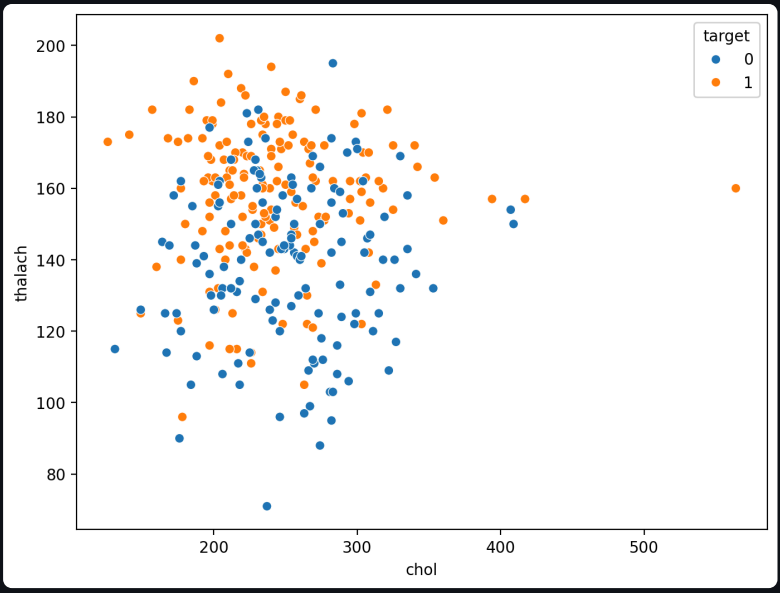
11.1. Scatter Plot of Age vs. Resting Blood Pressure

👉 **What this tells you about** `heart.csv` **:** This scatter plot shows the relationship between age and resting blood pressure (trestbps) for patients with and without heart disease. It helps us visualize how these two variables are related.



11.2. Scatter Plot of Cholesterol vs. Maximum Heart Rate

👉 **What this tells you about** `heart.csv` **:** This scatter plot shows the relationship between cholesterol levels and maximum heart rate (thalach) for patients with and without heart disease. It helps us visualize how these two variables are related.

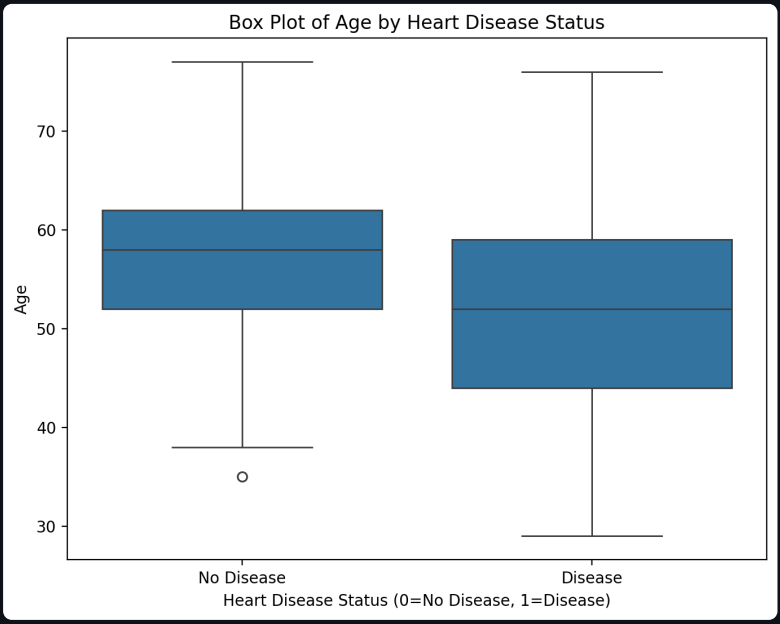


👉 **What this tells you:**

- The scatter plot shows individual data points for each patient, colored by heart disease status (target).
- You can see how cholesterol levels and maximum heart rates vary among patients with and without heart disease.
- This helps identify any potential relationships or patterns between these two variables.

11.3. Box Plot of Age by Target Variable

👉 **What this tells you about** `heart.csv` **:** This box plot shows the distribution of ages for patients with and without heart disease. It helps us visualize how age varies between these two groups.

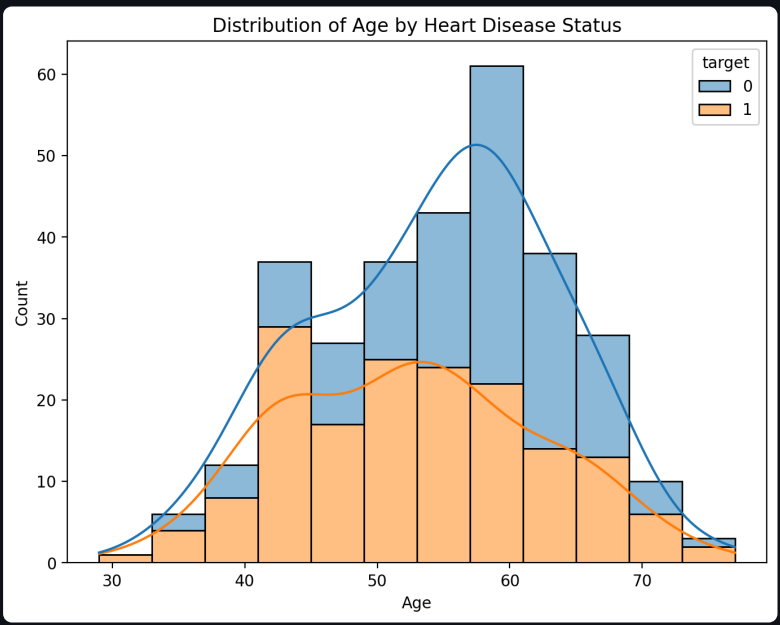


👉 **What this tells you:**

- The box plot shows the distribution of ages for patients with and without heart disease.
- You can see the median age, the interquartile range (IQR), and any potential outliers.

11.4. Distribution of Age by Target Variable

👉 **What this tells you about** `heart.csv` **:** This histogram shows the distribution of ages for patients with and without heart disease. It helps us visualize how age varies between these two groups.



👉 **What this tells you:**

- The histogram shows how many patients fall into different age groups, colored by heart disease status (target).
- You can see how age is distributed among patients with and without heart disease.
- This helps identify any potential differences in age distribution between these two groups.