HEART DISEASES DATASET EXPLORATORY DATA ANALYSIS

Your Name

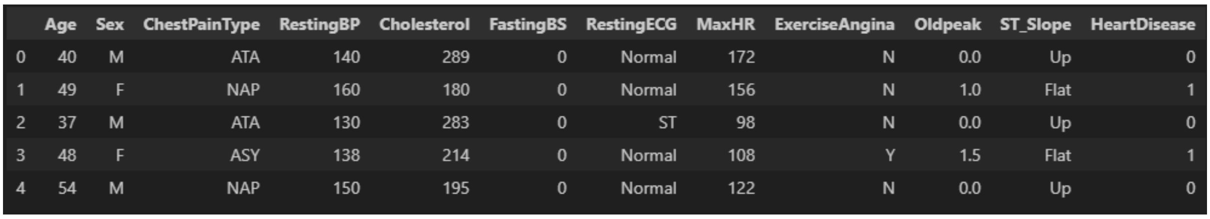
November 25, 2025

1 Introduction

The Kaggle Heart Diseases dataset contains clinical information about patients, including age, sex, resting blood pressure, cholesterol, maximum heart rate, chest pain type, ECG results, exercise-induced angina, ST slope and a binary target variable `HeartDisease`. This dataset is widely used for classification studies and enables an exploration of factors associated with heart disease, sex-related differences, and the dataset's separability.

2 Data Overview

Basic inspection (`head`, `describe`, `shape`, `info`) confirms a tabular dataset with mixed numerical and categorical features. The target distribution, visualized with a pie chart, shows the proportion of patients with and without heart disease.



	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

Figure 1: Head preview

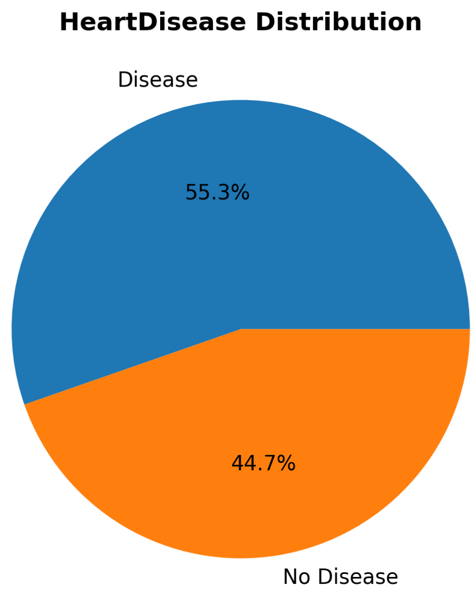
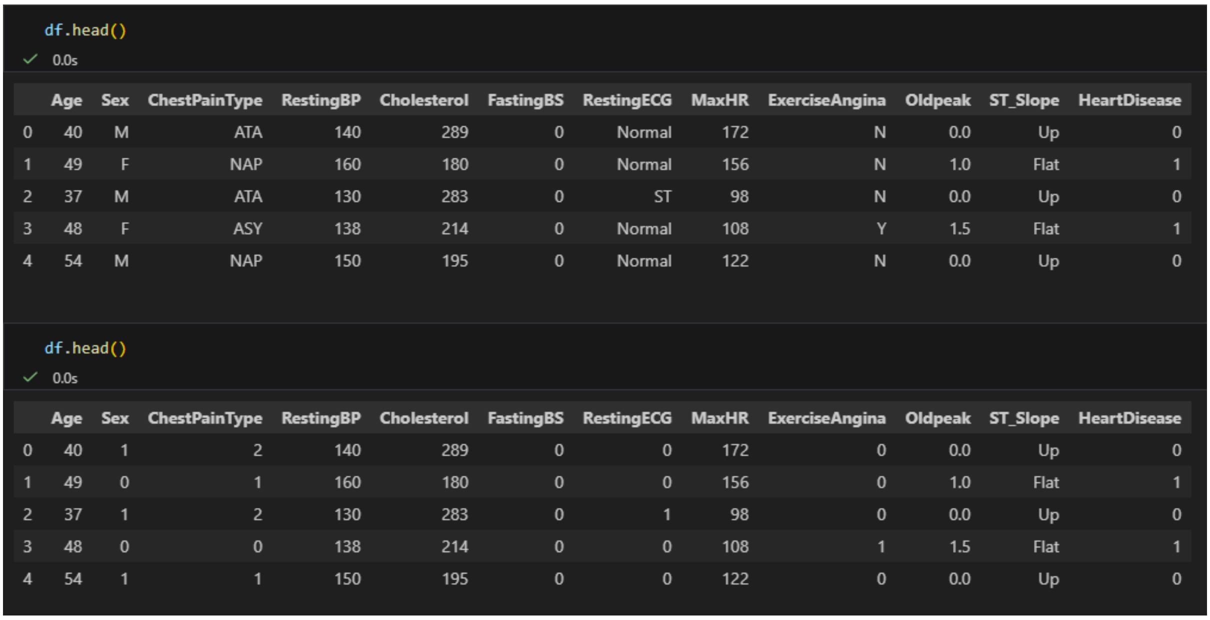


Figure 2: Heart Disease distribution

3 DATA CLEANING

The dataset was checked for missing values and inconsistent entries. Categorical variables (`Sex`, `ChestPainType`, `RestingECG`, `ExerciseAngina`) were encoded numerically using mapping and one-hot encoding for `ST_Slope`.



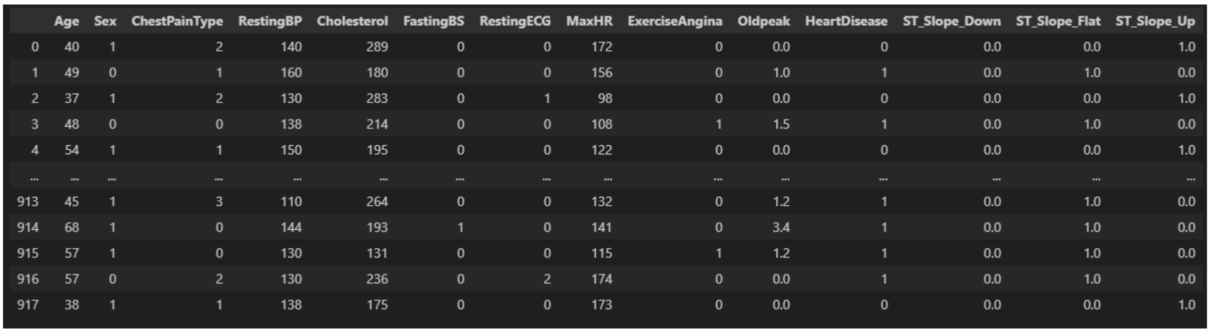
```
df.head()
✓ 0.0s
```

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

```
df.head()
✓ 0.0s
```

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	1	2	140	289	0	0	172	0	0.0	Up	0
1	49	0	1	160	180	0	0	156	0	1.0	Flat	1
2	37	1	2	130	283	0	1	98	0	0.0	Up	0
3	48	0	0	138	214	0	0	108	1	1.5	Flat	1
4	54	1	1	150	195	0	0	122	0	0.0	Up	0

Figure 3: Before and after ordinal encoding



	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	HeartDisease	ST_Slope_Down	ST_Slope_Flat	ST_Slope_Up
0	40	1	2	140	289	0	0	172	0	0.0	0	0.0	0.0	1.0
1	49	0	1	160	180	0	0	156	0	1.0	1	0.0	1.0	0.0
2	37	1	2	130	283	0	1	98	0	0.0	0	0.0	0.0	1.0
3	48	0	0	138	214	0	0	108	1	1.5	1	0.0	1.0	0.0
4	54	1	1	150	195	0	0	122	0	0.0	0	0.0	0.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
913	45	1	3	110	264	0	0	132	0	1.2	1	0.0	1.0	0.0
914	68	1	0	144	193	1	0	141	0	3.4	1	0.0	1.0	0.0
915	57	1	0	130	131	0	0	115	1	1.2	1	0.0	1.0	0.0
916	57	0	2	130	236	0	2	174	0	0.0	1	0.0	1.0	0.0
917	38	1	1	138	175	0	0	173	0	0.0	0	0.0	0.0	1.0

Figure 4: After one hot encoding ST$_Slope$

Unrealistic values were identified:

- **RestingBP**: a value of 0 mmHg is physiologically impossible and was removed.
- **Cholesterol**: many entries with value 0, also unrealistic, were treated as missing and imputed using the median cholesterol value.

4 EXPLORATORY DATA ANALYSIS

Univariate and bivariate analyses were performed to understand distributions and relationships among variables.

4.1 Univariate analysis

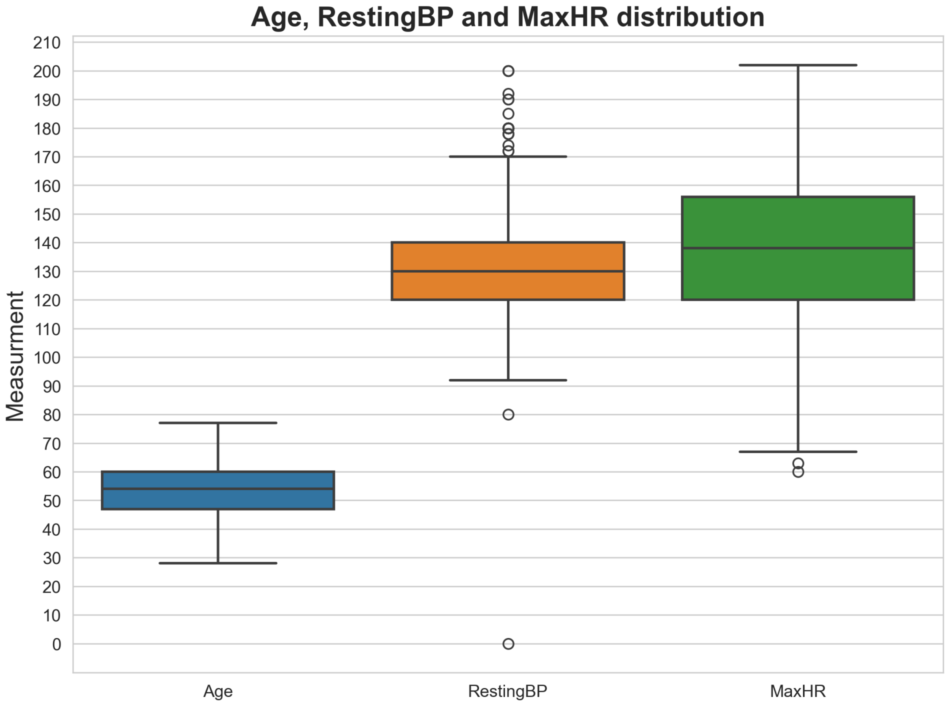


Figure 5: A boxplot of Age, RestingBP and MaxHR

A boxplot of `Age`, `RestingBP` and `MaxHR` shows:

- **Age**: Median around 55 years; most patients fall between roughly 48 and 60, with a wider range of about 27 to 76 and no major outliers.
- **RestingBP**: Median around 130 mmHg; interquartile range 120–140 mmHg; values above 170 appear as outliers, and 0 was flagged as invalid and deleted.
- **MaxHR**: Median around 138 bpm; interquartile range 120–156 bpm; a few low outliers below 68 bpm.

Cholesterol Distribution

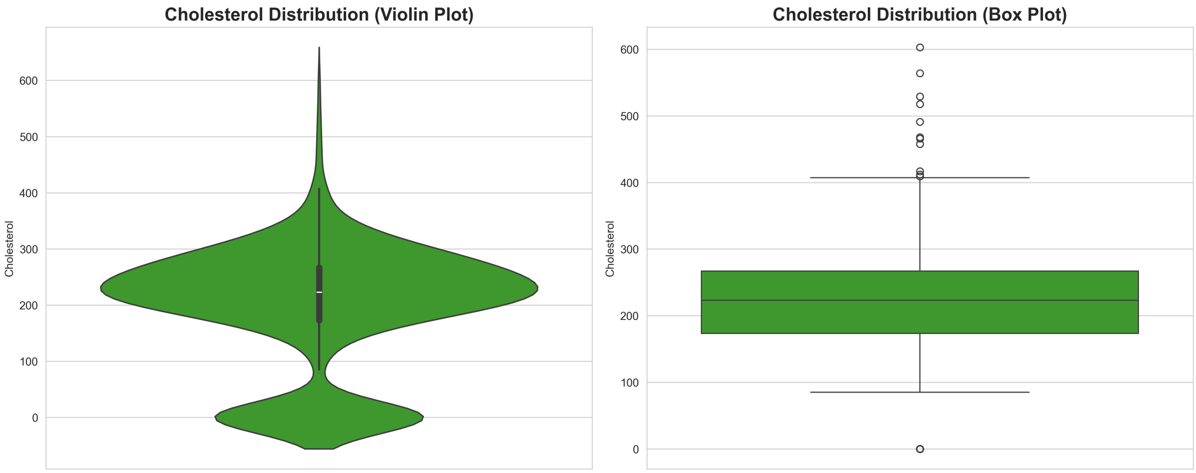


Figure 6: A boxplot of Age, RestingBP and MaxHR

A violin plot and boxplot of `Cholesterol` reveal:

- A bimodal distribution with a clear cluster around 0 (invalid) and a main cluster between 150 and 300.
- Several high outliers extending up to about 630.
- Zero values were treated as missing and imputed with the median to retain sufficient data.

ChestPainType and RestingECG

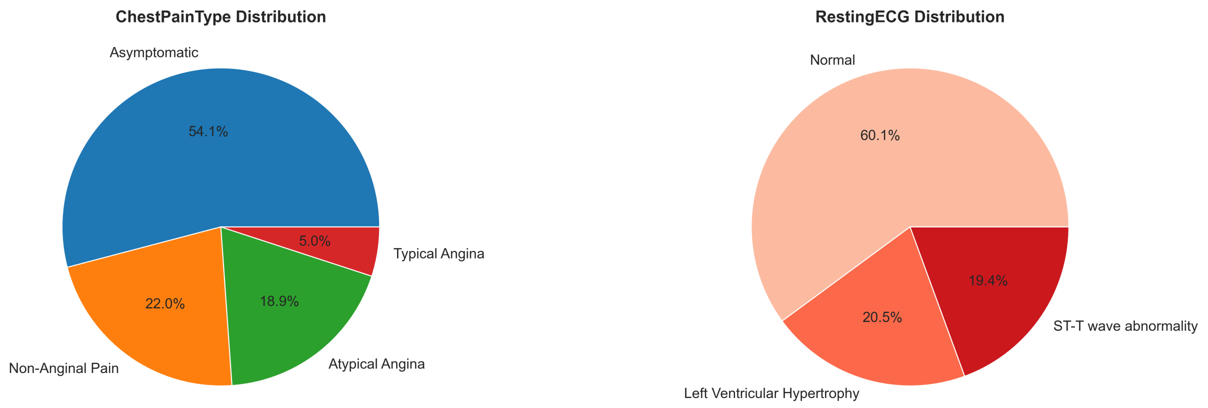


Figure 7: Cholesterol distribution

Pie charts show the distribution of `ChestPainType` and `RestingECG`:

- Most patients are classified as asymptomatic in terms of chest pain.
- Most ECG results are normal.
- About 5% exhibit typical angina (highest chest pain level), and around 20.5% show left ventricular hypertrophy, indicating structural heart changes.

4.2 Bivariate analysis

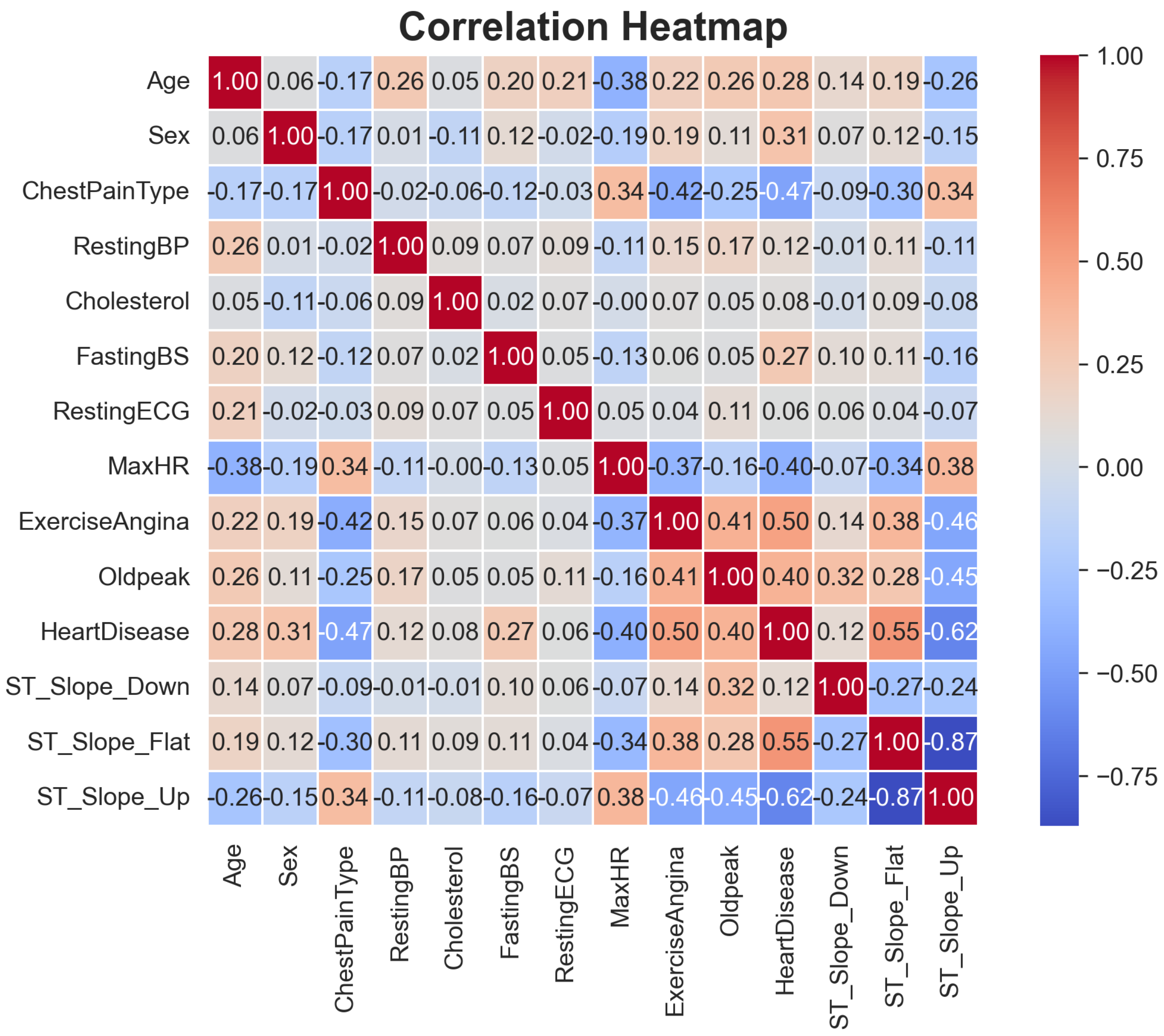


Figure 8: Correlation heatmap

A correlation heatmap of all processed features highlights several important relationships:

- **ExerciseAngina** vs **HeartDisease**: correlation of approximately +0.50, indicating that exercise-induced angina is strongly associated with heart disease.
- **ST_Slope_Flat** vs **HeartDisease**: correlation around +0.55, making flat ST slope one of the strongest predictors.
- **MaxHR** vs **HeartDisease**: moderate negative correlation (~ -0.40), suggesting that lower maximum heart rate during exercise is linked to disease.
- **Age** vs **MaxHR**: negative correlation (~ -0.38), older patients tend to achieve lower maximum heart rates.

Sex and Heart Disease

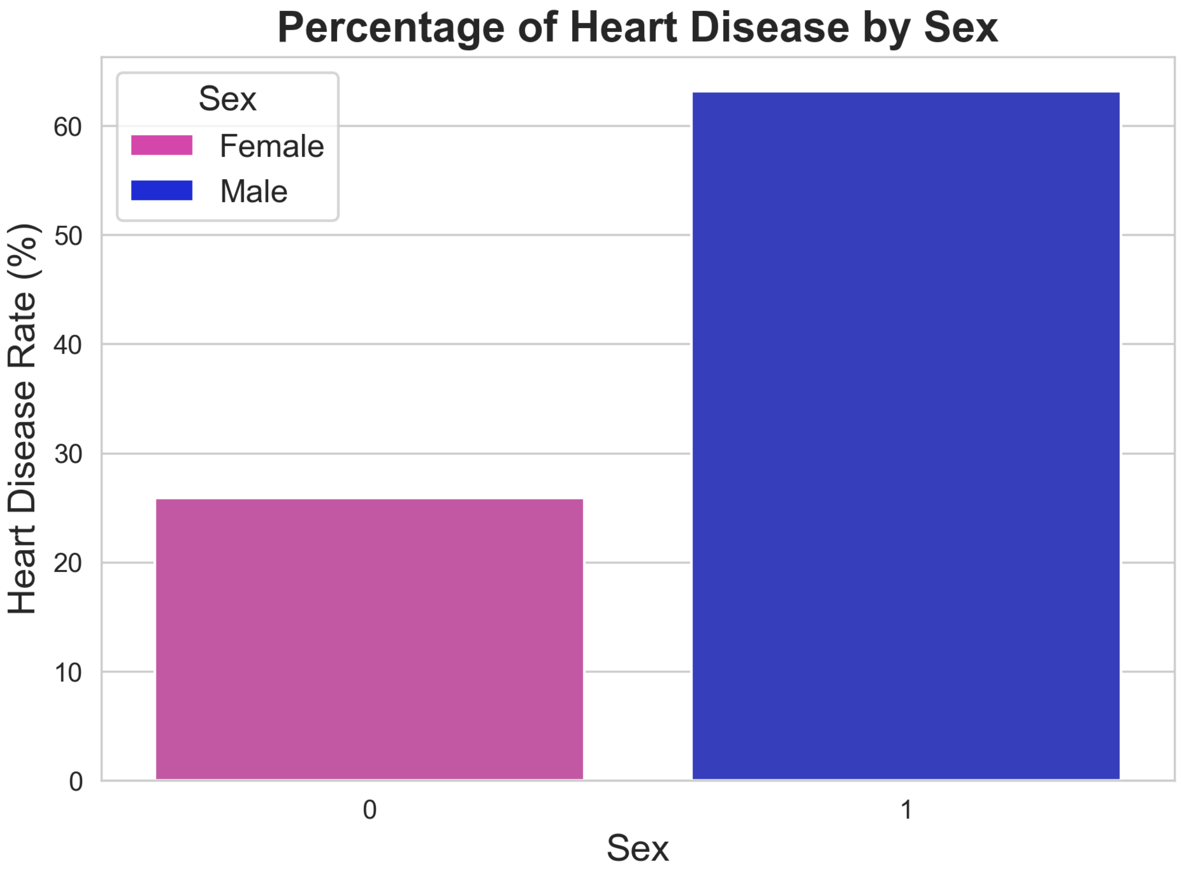


Figure 9: Heart diseases by sex

Heart disease rates by sex were computed and visualized with a bar plot:

- Males show a heart disease rate of roughly 63%.
- Females show a significantly lower rate of roughly 26%.

4.6 Pairwise Relationships and Linear Separability

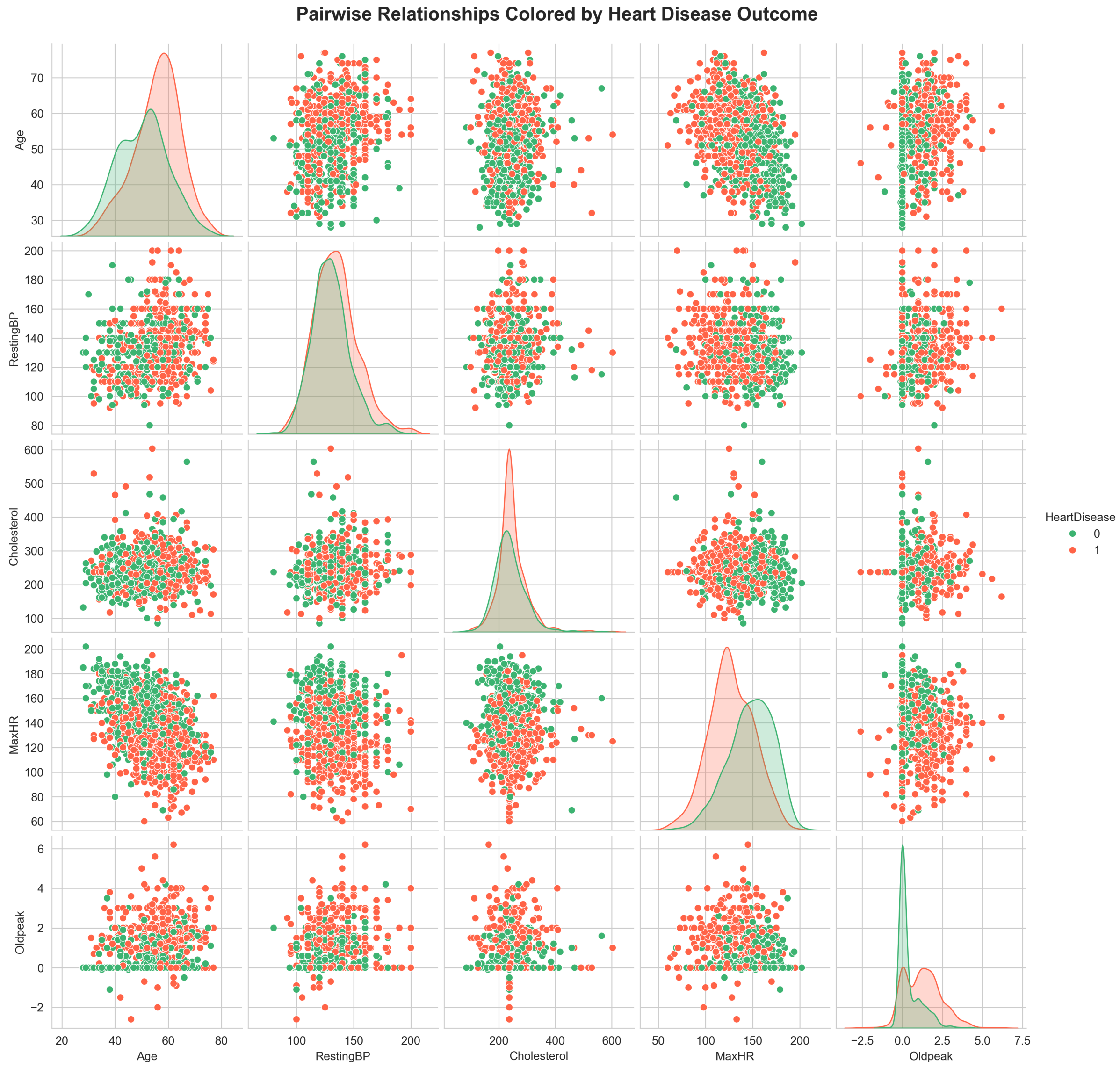


Figure 10: Pairplot

A pairplot of selected features (`Age`, `RestingBP`, `Cholesterol`, `MaxHR`, `Oldpeak`, `HeartDisease`) colored by the target reveals:

- Red (disease) and green (no disease) points are heavily mixed in almost all two-dimensional projections.
- There is no simple straight line in any 2D feature plane that perfectly separates disease from non-disease cases.

This means a simple linear model (like a basic perceptron or a linear SVM) relying on only two of these features will not be able to perfectly classify the patients.

5 CONCLUSION

This exploratory analysis shows that several clinical features are strongly associated with heart disease, especially exercise-induced angina and flat ST slope, with maximum heart rate and age also playing important roles. Data cleaning (removal of invalid blood pressure values and median imputation for cholesterol) was necessary to ensure data quality. Males exhibit a substantially higher heart disease rate than females. Finally, the lack of linear separability in pairwise plots indicates that simple linear classifiers on individual or paired features are insufficient and more expressive models are required.