Intelligent Data Analysis - Lab4 Report

Exploratory Data Analysis (EDA) of Heart Dataset

Student: Manasova Gulum 230121028 MATDAIS23

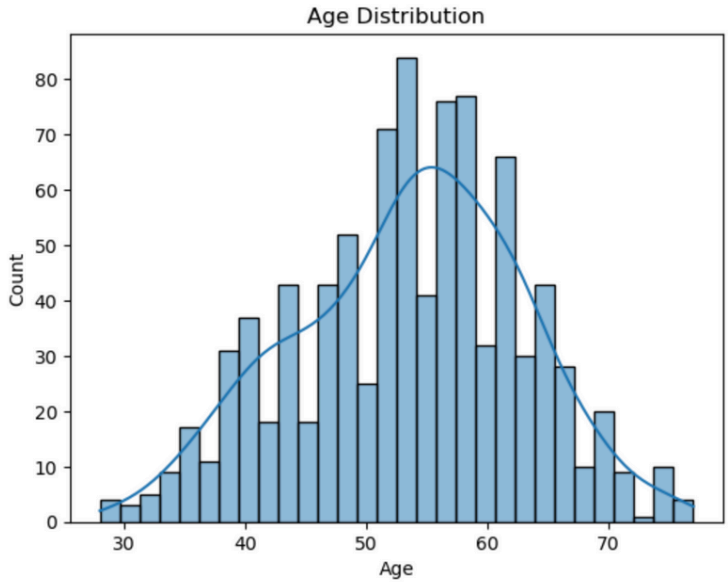
Professor: Mr.Remudin Mekuria

Lab Objective

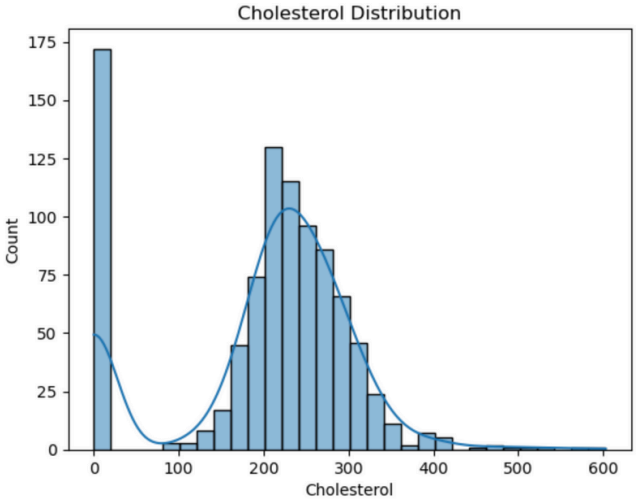
The main objective is to perform an Exploratory Data Analysis (EDA) that uncovers patterns, relationships, and trends in the dataset related to heart disease. The goal is to tell a clear story about which features are most associated with the presence or absence of heart disease, helping guide further modeling.

1. Univariate Analysis

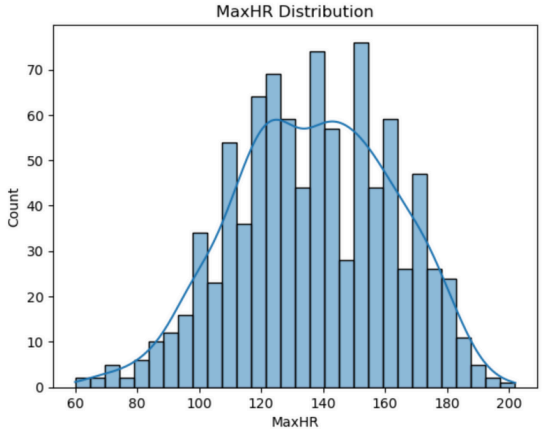
We first explored the distribution of individual features to understand their ranges and frequencies:



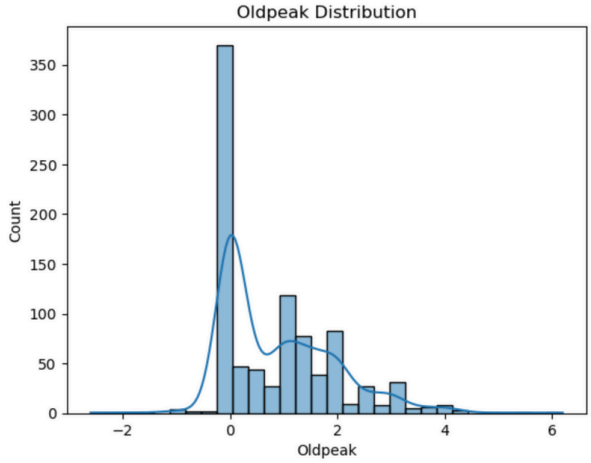
The majority of patients are between 40 and 60 years old, indicating that middle-aged adults form the largest group at risk of heart disease.



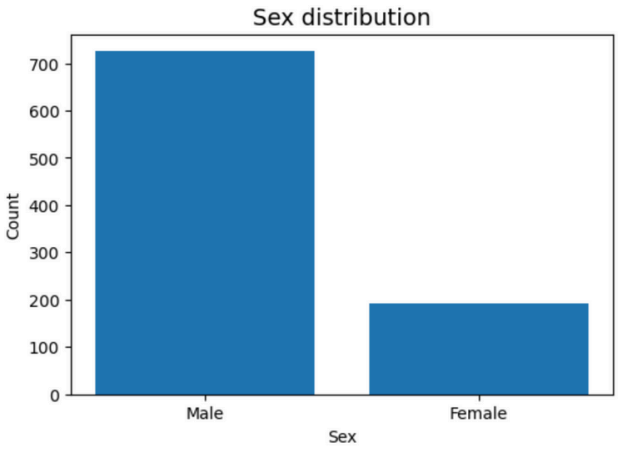
Cholesterol levels are mostly concentrated between 150 and 300 mg/dl, but a few patients have very high values above 350, showing possible outliers or hypercholesterolemia cases.



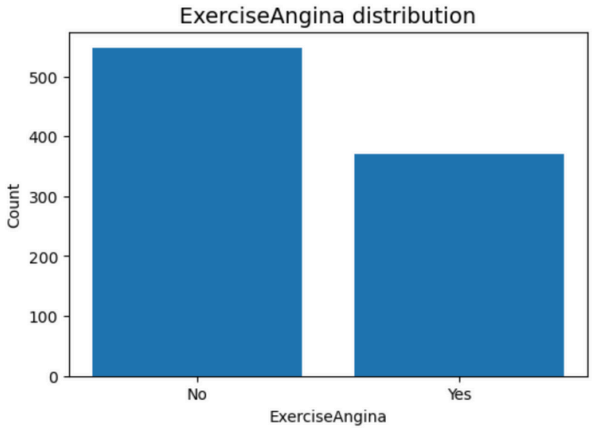
Most patients achieve a maximum heart rate between 120 and 170 bpm, while patients with lower MaxHR values tend to show higher rates of heart disease.



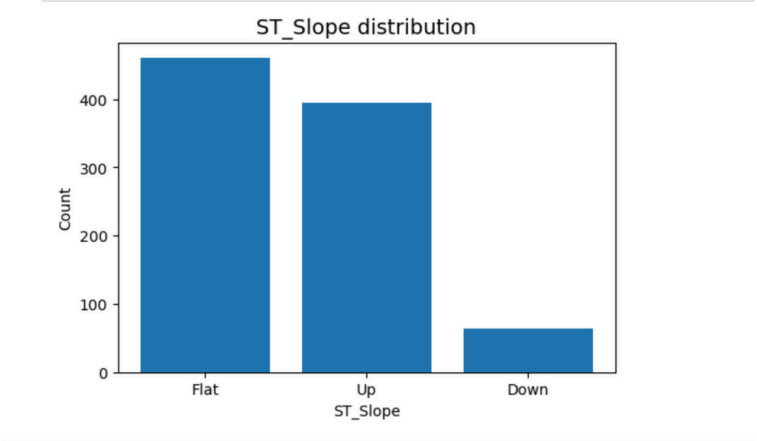
Oldpeak values are mostly below 2.0, and higher Oldpeak levels are associated with an increased likelihood of heart disease, showing a positive relationship.



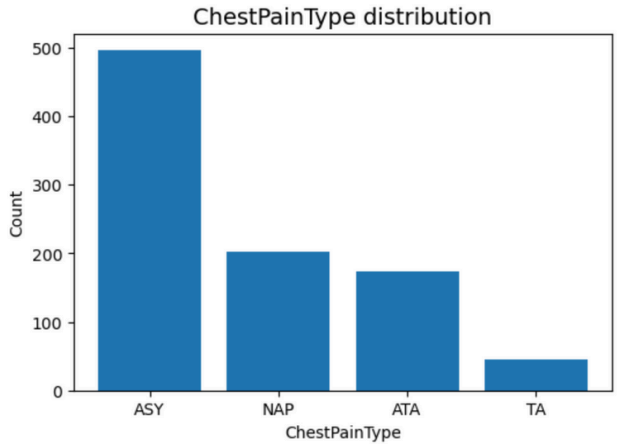
Males represent a higher proportion of patients compared to females, and the rate of heart disease is notably higher among males.



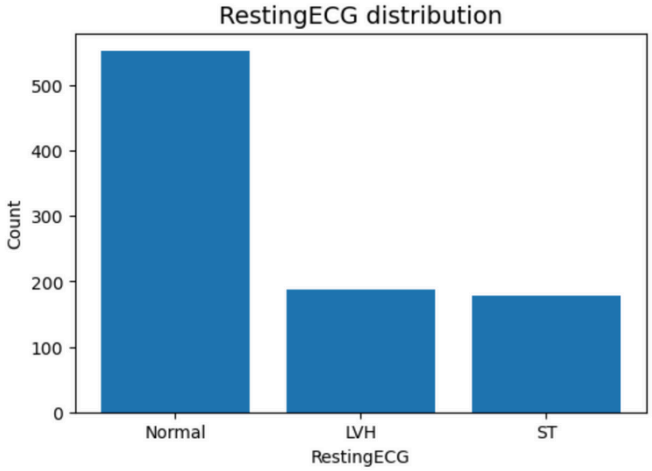
Patients reporting exercise-induced angina (Yes) show a much higher frequency of heart disease, highlighting its strong clinical importance.



The Flat slope category strongly correlates with heart disease presence, while the Up slope is mostly seen among healthy patients.



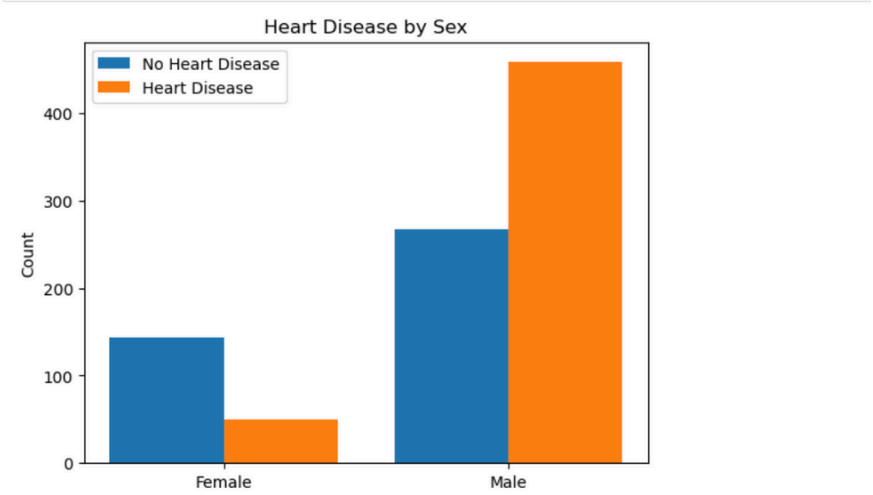
ASY (Asymptomatic) chest pain type dominates among heart disease cases, whereas **ATA** and **NAP** types are more common in healthy individuals.

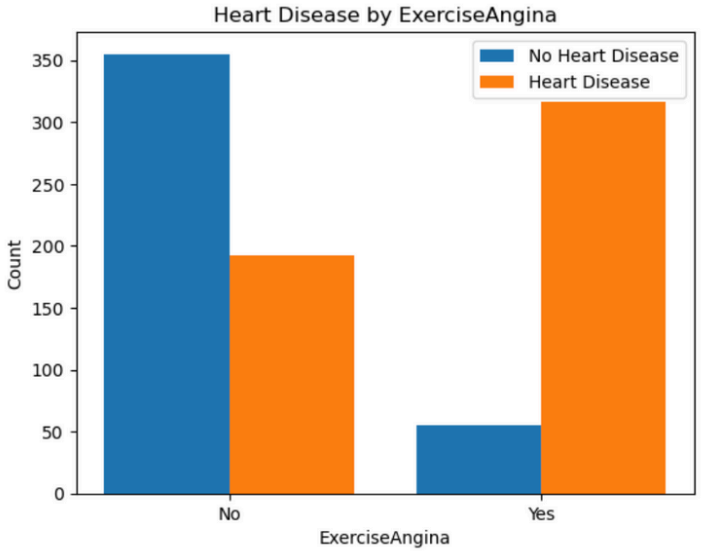


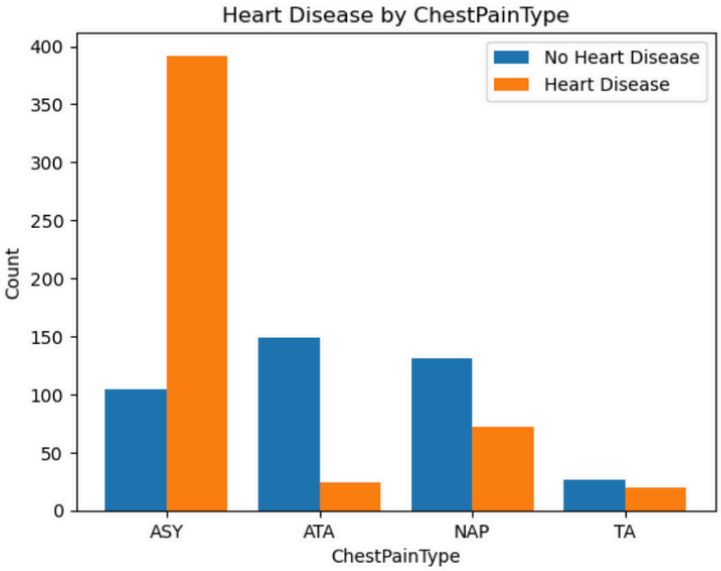
Most patients have normal ECG results, but those with ST or LVH findings show slightly higher chances of having heart disease.

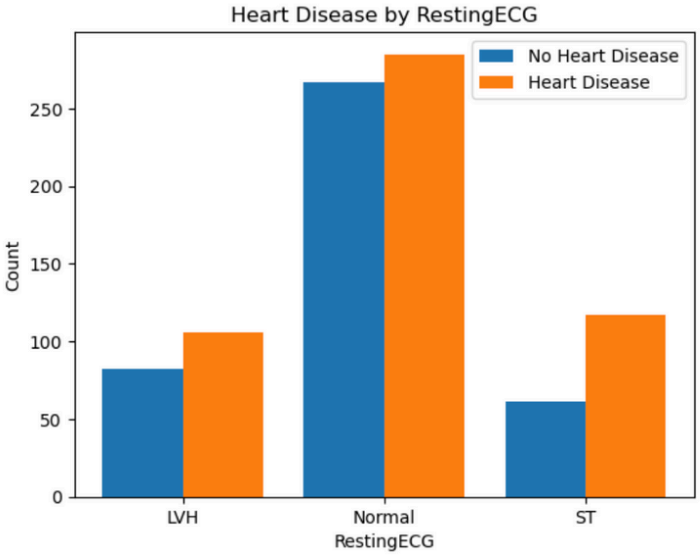
2. Bivariate Analysis

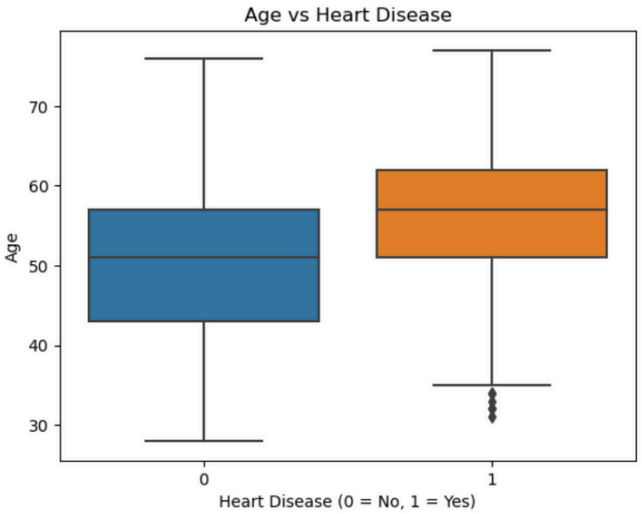
We examined relationships between features and the survival outcome:

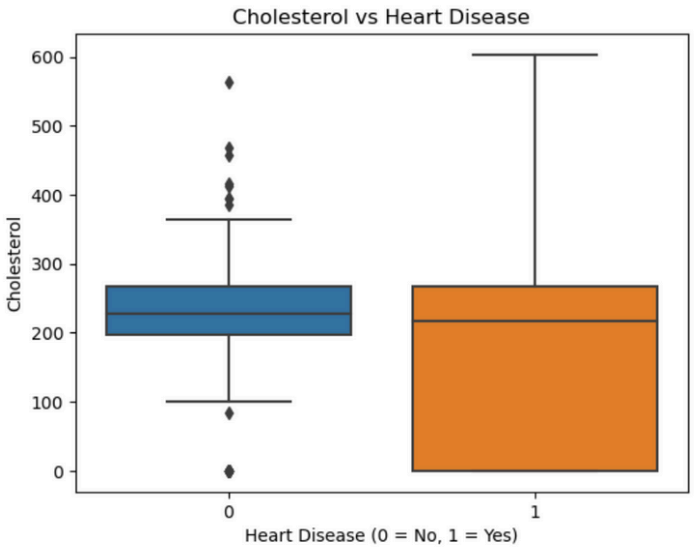


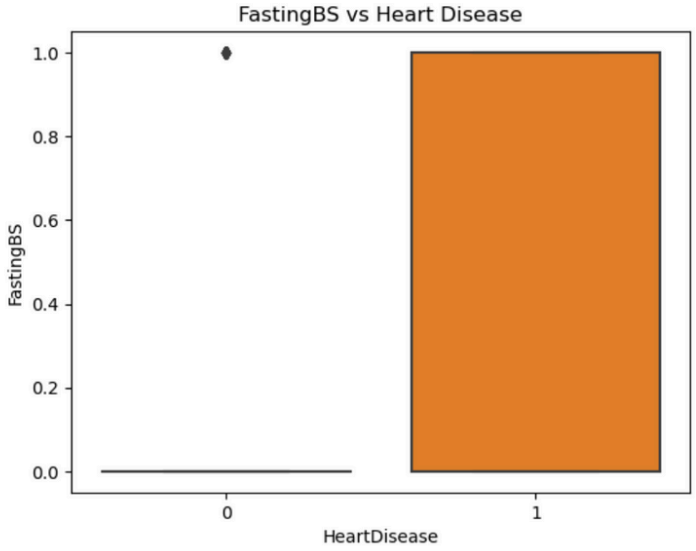


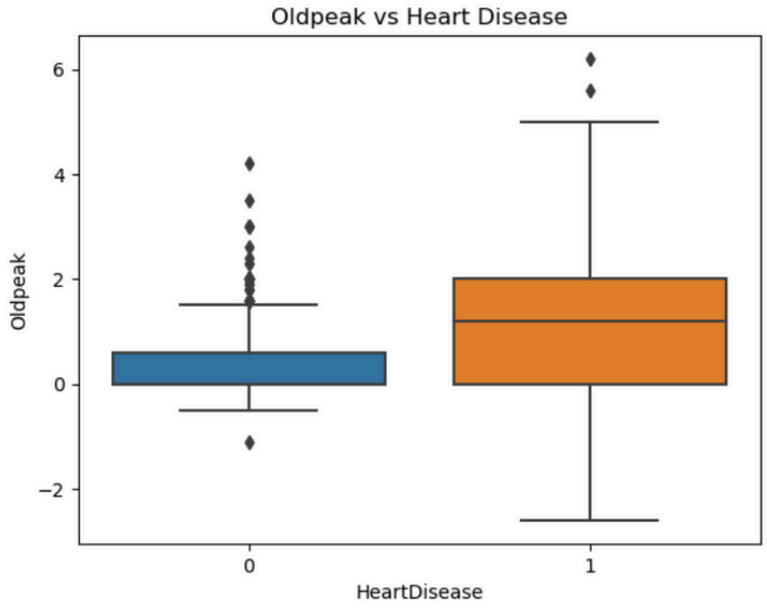

Heart Disease by ChestPainType
No Heart Disease
Heart Disease
Count
400
350
300
250
200
150
100
50
0
ASY
ATA
NAP
TA
ChestPainType

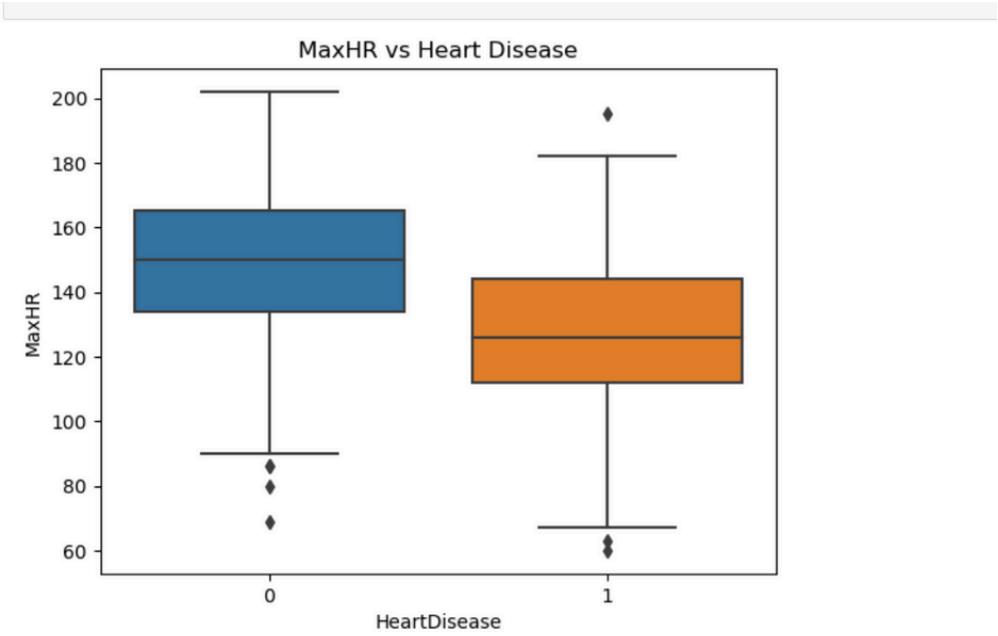

Heart Disease by RestingECG
No Heart Disease
Heart Disease
Count
250
200
150
100
50
0
LVH
Normal
ST
RestingECG


Age vs Heart Disease
Age
70
60
50
40
30
0
1
Heart Disease (0 = No, 1 = Yes)


Cholesterol vs Heart Disease
Cholesterol
600
500
400
300
200
100
0
0
1
Heart Disease (0 = No, 1 = Yes)

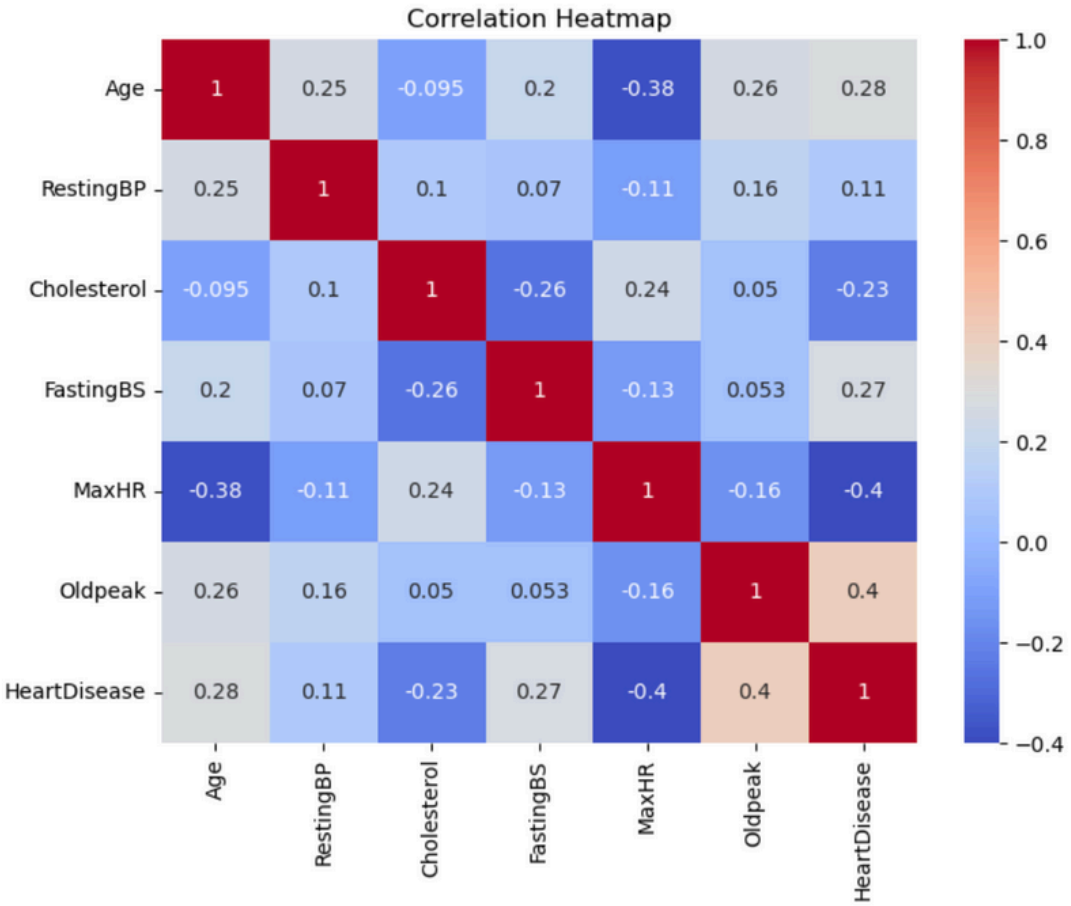

FastingBS vs Heart Disease
FastingBS
1.0
0.8
0.6
0.4
0.2
0.0
0
1
HeartDisease


Oldpeak vs Heart Disease
Oldpeak
6
4
2
0
−2
0
1
HeartDisease



3. Correlation Analysis

A correlation heatmap of numerical variables showed:

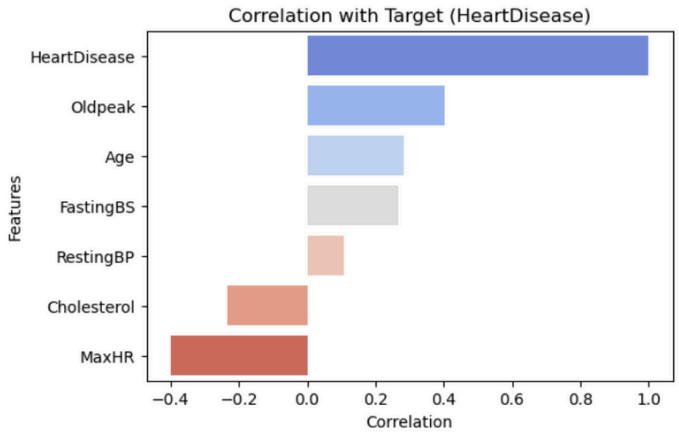


The heatmap reveals moderate correlations — HeartDisease is positively correlated with Oldpeak (0.40) and negatively correlated with MaxHR (-0.40), showing consistent clinical relationships.

4. Target Leakage Check

```
Correlation of features with HeartDisease:

HeartDisease    1.000000
Oldpeak         0.403951
Age             0.282039
FastingBS       0.267291
RestingBP       0.107589
Cholesterol    -0.232741
MaxHR          -0.400421
Name: HeartDisease, dtype: float64
```



Variables such as Oldpeak, MaxHR, and Age show moderate relationships with the target variable *HeartDisease*. These are clinically meaningful predictors rather than data leaks.
No variables appear to be derived directly from the outcome, suggesting that no evidence of target leakage is present in this dataset.

5. Draft Data Dictionary

	Column	Data Type	Missing Values	Unique Values
0	Age	int64	0	50
1	Sex	object	0	2
2	ChestPainType	object	0	4
3	RestingBP	int64	0	67
4	Cholesterol	int64	0	222
5	FastingBS	int64	0	2
6	RestingECG	object	0	3
7	MaxHR	int64	0	119
8	ExerciseAngina	object	0	2
9	Oldpeak	float64	0	53
10	ST_Slope	object	0	3
11	HeartDisease	int64	0	2

This dictionary provides context for understanding each variable and preparing them for modeling.

10 Key Insights from EDA

1. The majority of patients fall within the **40–60 age range**, suggesting midlife adults are most at risk.
2. **Males** represent most of the dataset and have a **higher heart disease rate** than females.

3. **High Oldpeak** and **low MaxHR** are the strongest numerical indicators of heart disease.
4. **Asymptomatic (ASY)** chest pain type is the most frequent among heart disease patients.
5. **ExerciseAngina = Yes** strongly correlates with disease presence.
6. **Flat ST_Slope** values are closely linked to heart disease, unlike **Up slopes**, which indicate lower risk.
7. **Cholesterol** shows wide variation and potential outliers above 350 mg/dl.
8. **Age** has a moderate positive correlation with heart disease, supporting known medical trends.
9. No variable shows evidence of **target leakage**, confirming data reliability for modeling.
10. The dataset overall demonstrates consistent clinical logic — risk increases with **age, low MaxHR, high Oldpeak, and presence of angina**.

Conclusion

The exploratory data analysis (EDA) highlights clear medical and demographic patterns influencing heart disease.Key predictors such as **Oldpeak**, **MaxHR**, **Age**, **Sex**, and **ExerciseAngina** show strong and clinically meaningful relationships with the target variable.

These features, along with carefully handled categorical data such as **ChestPainType** and **ST_Slope**, should play a central role in predictive modeling.

No signs of target leakage were found, confirming the dataset's reliability for further analysis.

This study establishes a solid foundation for the next phase — **data preprocessing, feature selection, and model development** to predict heart disease risk more accurately.