Exploratory Data Analysis Report

Heart Disease Prediction Dataset

By: Disha Sharma

1. Objective of the Analysis

The objective of this exploratory data analysis (EDA) is to systematically examine the distribution, relationships, and predictive relevance of clinical and demographic variables associated with heart disease. The analysis aims to identify patterns and risk indicators that can inform subsequent machine learning models for early detection of heart disease.

2. Dataset Overview

The dataset consists of **270 patient records** and **14 clinical attributes**, including demographic information, physiological measurements, symptom indicators, and diagnostic test results. The target variable represents the **presence (1) or absence (0) of heart disease**.

All features are numerical, either continuous or encoded categorical variables, enabling both statistical analysis and machine learning applications.

3. Data Quality Assessment

Initial inspection of the dataset revealed:

- No missing values across all features
- No duplicate records
- Consistent data types suitable for analysis

As a result, the dataset required **no imputation or cleaning**, allowing direct progression to exploratory analysis without introducing bias through preprocessing.

4. Target Variable Distribution

The dataset contains **150 patients without heart disease (55.56%)** and **120 patients with heart disease (44.44%)**, indicating a **mild class imbalance**. Although the imbalance is not severe, it is clinically significant.

In medical prediction tasks, this imbalance necessitates careful metric selection. **False negatives**, where a patient with heart disease is misclassified as healthy, pose greater risk than false positives. Consequently, evaluation metrics such as **recall, F1-score, and ROC-AUC** are more appropriate than accuracy alone for future modeling.

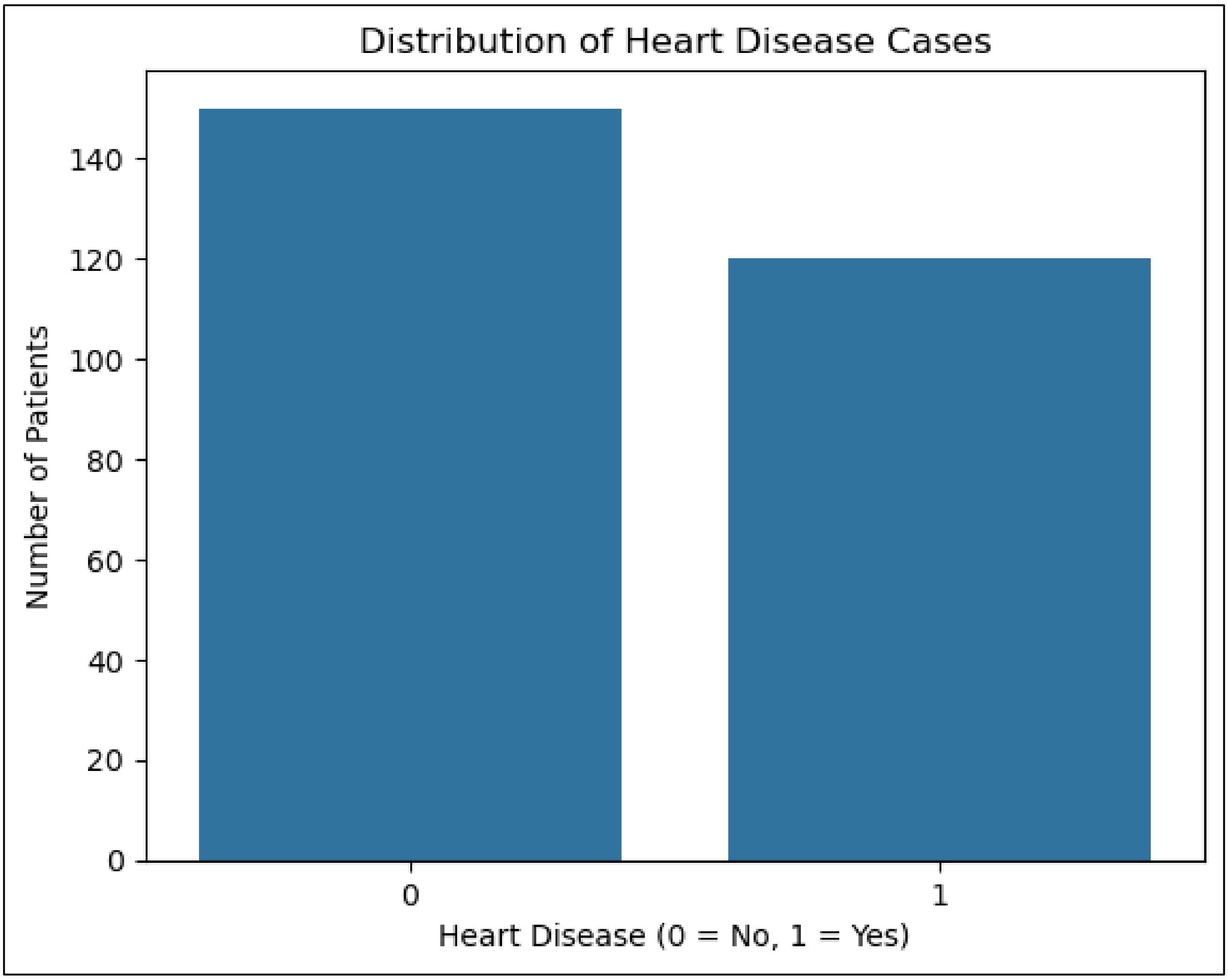
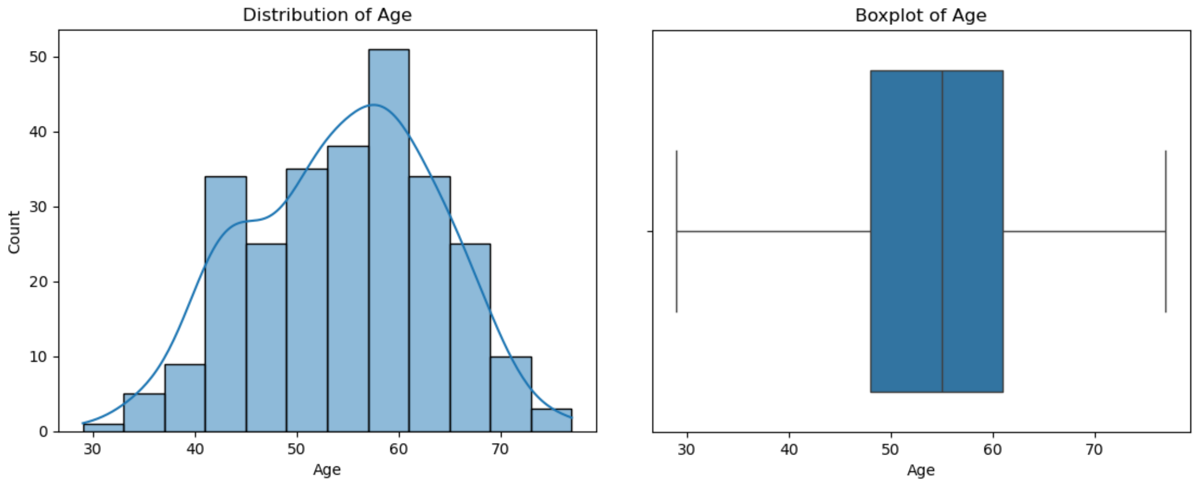


Figure 1: Distribution of Heart Disease Cases

5. Univariate Analysis of Numerical Features

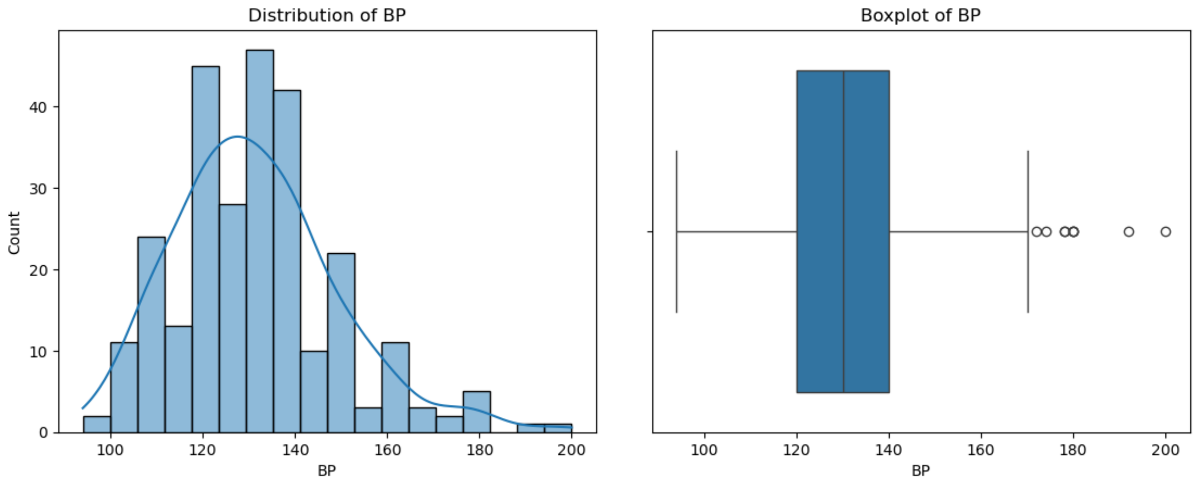
Age

The age distribution is approximately normal, with the majority of patients falling between **45 and 65 years**. The absence of extreme outliers suggests a relatively consistent age range among individuals undergoing cardiac evaluation.



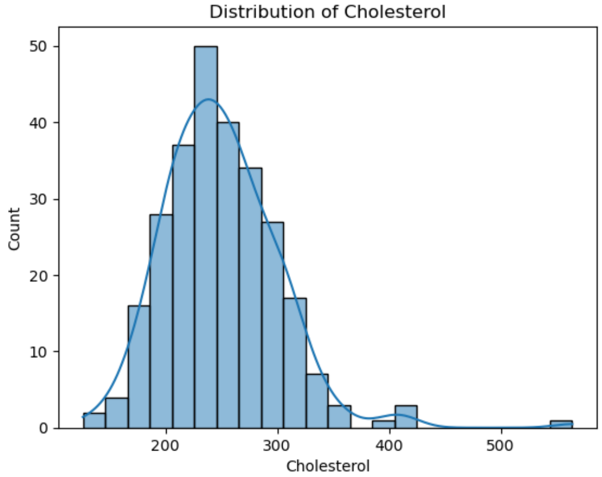
Resting Blood Pressure (BP)

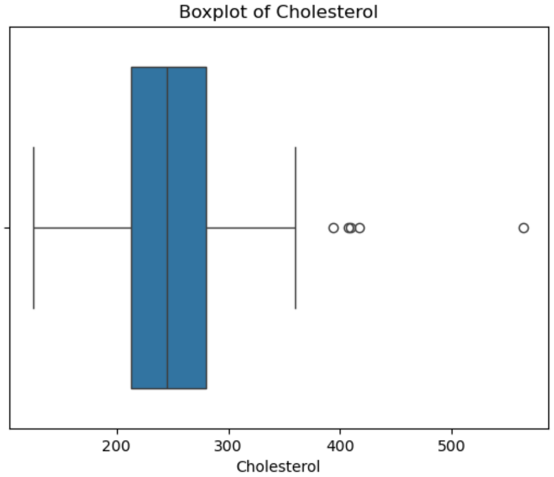
Resting blood pressure values are concentrated between **110 and 140 mm Hg**, with a right-skewed distribution. Several high-value outliers above **170 mm Hg** are observed, indicating the presence of patients with clinically elevated blood pressure.



Cholesterol

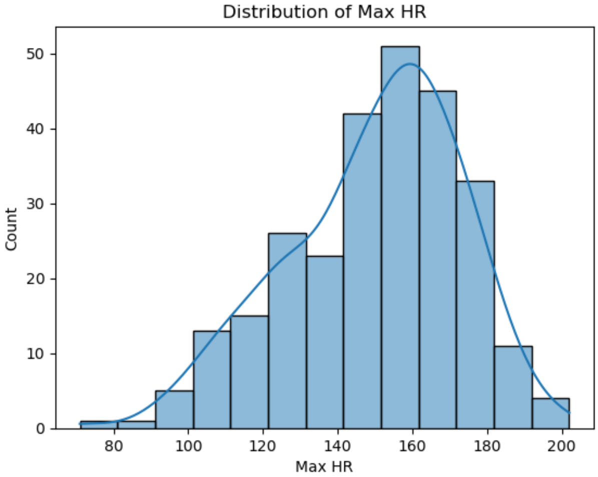
Cholesterol levels show a pronounced right skew, with most values between **200 and 300 mg/dL**. A small number of extreme outliers above **400 mg/dL** are present, suggesting unusually high cholesterol levels in certain individuals.

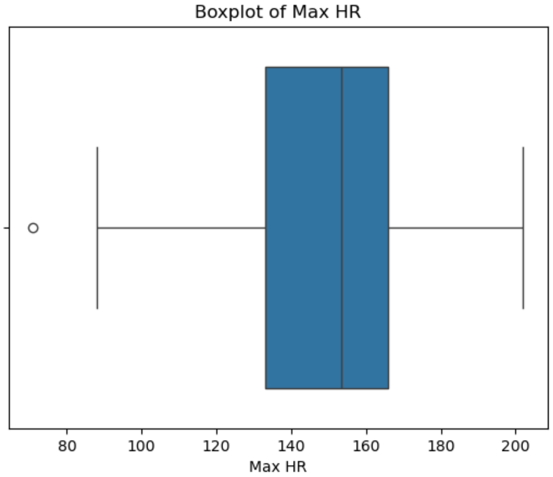




Maximum Heart Rate (Max HR)

Maximum heart rate values are moderately symmetric and widely distributed, primarily between **130 and 170 beats per minute**. Few outliers are present, indicating stable cardiovascular response across most patients.

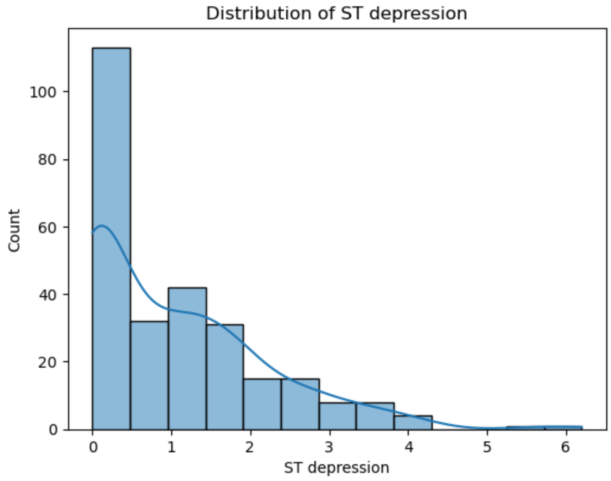


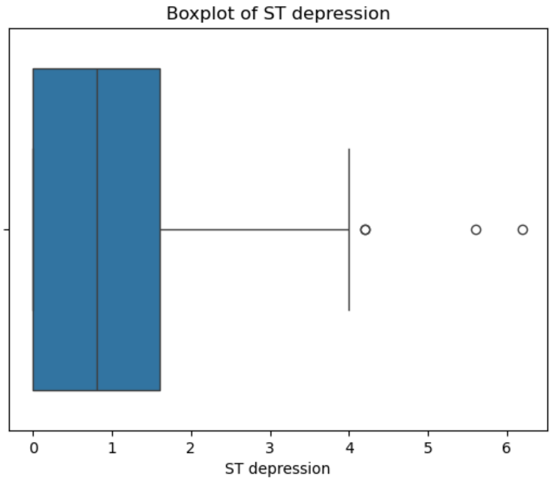


ST Depression

ST depression exhibits a heavily right-skewed distribution, with most values concentrated near **0–2**. Higher values extending beyond **4** are observed, representing patients with more pronounced ST segment depression, a known indicator of cardiac stress.

Overall, several numerical features demonstrate skewness and outliers, which should be considered during scaling and model selection.

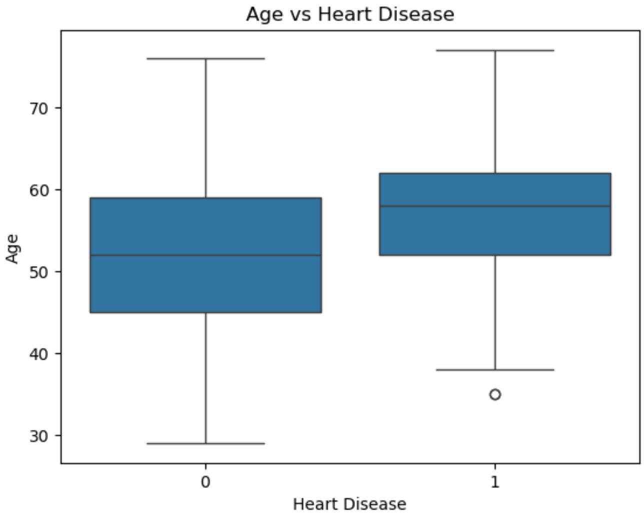




6. Bivariate Analysis: Feature Relationships with Heart Disease

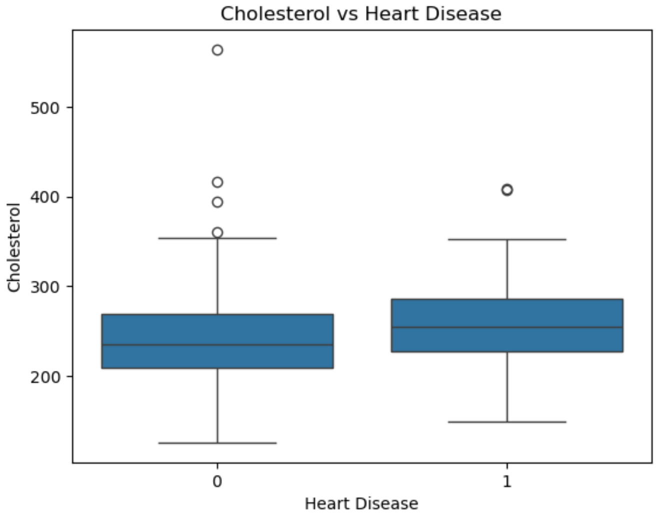
6.1 Age vs Heart Disease

Patients with heart disease exhibit a higher median age compared to those without the condition. Although substantial overlap exists between the two groups, the upward shift in the median and interquartile range indicates that age contributes meaningfully to heart disease risk, though it does not act as a standalone predictor.



6.2 Cholesterol vs Heart Disease

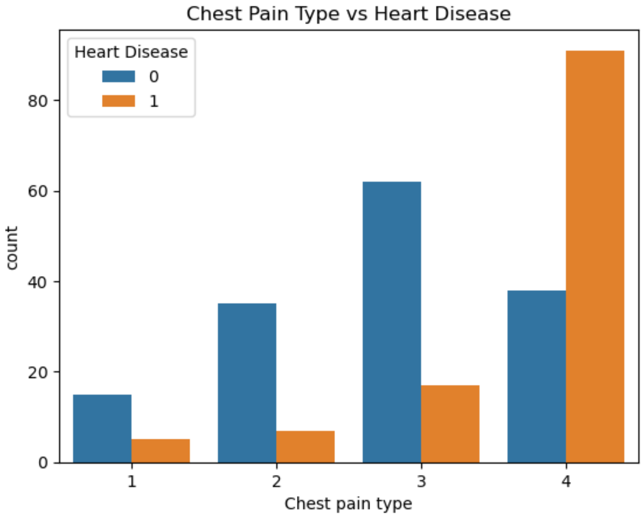
Patients diagnosed with heart disease show a slightly higher median cholesterol level. However, the extensive overlap between groups and the presence of high cholesterol outliers among non-diseased patients indicate that cholesterol alone has limited discriminative power and must be considered in combination with other clinical indicators.



6.3 Chest Pain Type vs Heart Disease

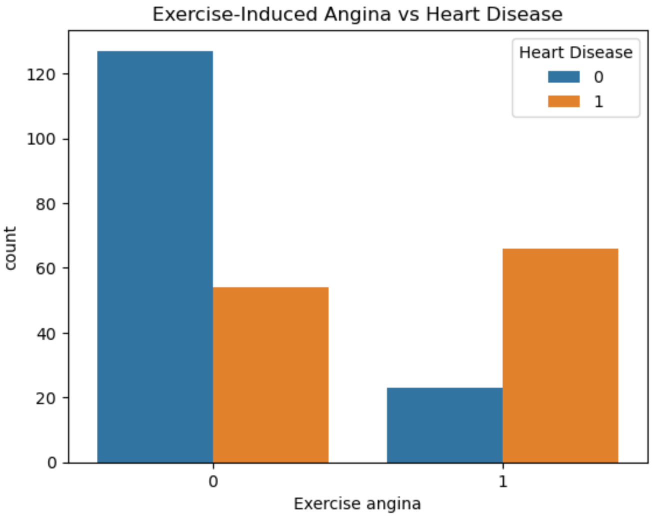
Chest pain type demonstrates one of the strongest associations with heart disease presence. Chest pain type 4 shows a markedly higher number of heart disease cases, while types 1 and 2 are predominantly associated with disease absence. Type 3 displays a mixed distribution.

This pronounced variation across categories highlights chest pain type as a highly informative categorical feature and a key contributor to predictive modeling.

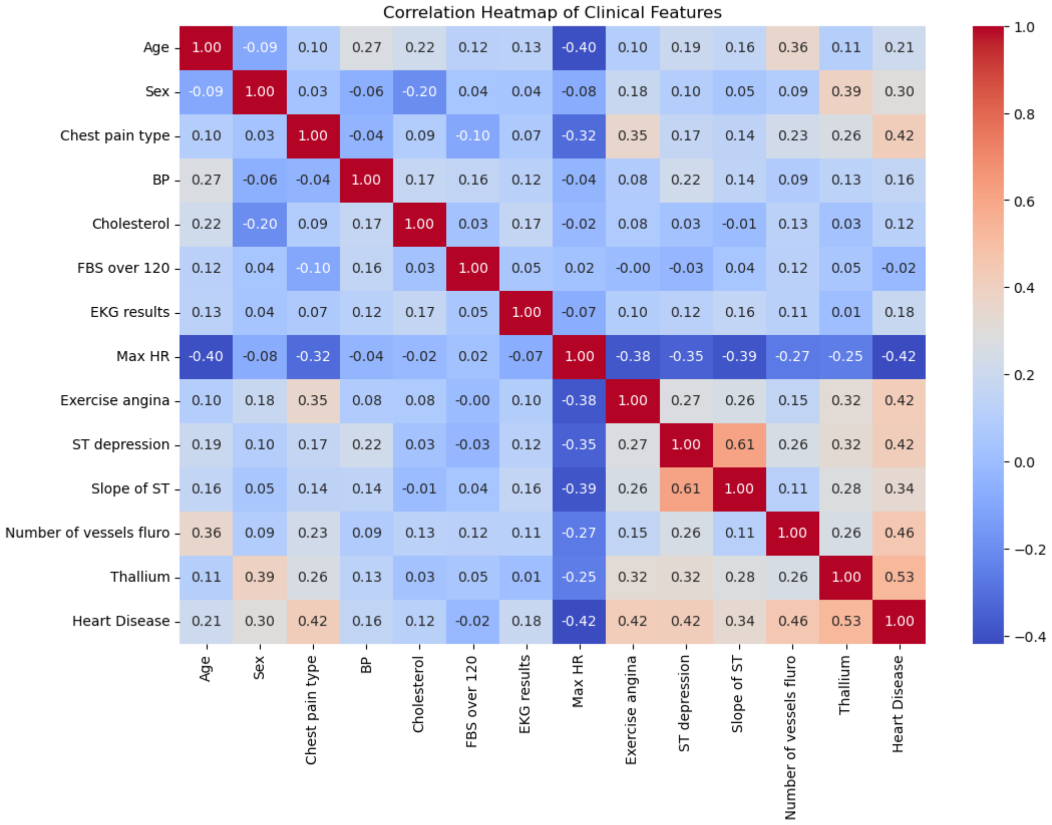


6.4 Exercise-Induced Angina vs Heart Disease

Patients experiencing exercise-induced angina show a substantially higher incidence of heart disease compared to those who do not. This clear separation indicates strong discriminative capability, making this feature one of the most predictive variables in the dataset.



7. Correlation Analysis



	Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluro	Thallium	Heart Disease
Age	1.00	-0.09	0.10	0.27	0.22	0.12	0.13	-0.40	0.10	0.19	0.16	0.36	0.11	0.21
Sex	-0.09	1.00	0.03	-0.06	-0.20	0.04	0.04	-0.08	0.18	0.10	0.05	0.09	0.39	0.30
Chest pain type	0.10	0.03	1.00	-0.04	0.09	-0.10	0.07	-0.32	0.35	0.17	0.14	0.23	0.26	0.42
BP	0.27	-0.06	-0.04	1.00	0.17	0.16	0.12	-0.04	0.08	0.22	0.14	0.09	0.13	0.16
Cholesterol	0.22	-0.20	0.09	0.17	1.00	0.03	0.17	-0.02	0.08	0.03	-0.01	0.13	0.03	0.12
FBS over 120	0.12	0.04	-0.10	0.16	0.03	1.00	0.05	0.02	-0.00	-0.03	0.04	0.12	0.05	-0.02
EKG results	0.13	0.04	0.07	0.12	0.17	0.05	1.00	-0.07	0.10	0.12	0.16	0.11	0.01	0.18
Max HR	-0.40	-0.08	-0.32	-0.04	-0.02	0.02	-0.07	1.00	-0.38	-0.35	-0.39	-0.27	-0.25	-0.42
Exercise angina	0.10	0.18	0.35	0.08	0.08	-0.00	0.10	-0.38	1.00	0.27	0.26	0.15	0.32	0.42
ST depression	0.19	0.10	0.17	0.22	0.03	-0.03	0.12	-0.35	0.27	1.00	0.61	0.26	0.32	0.42
Slope of ST	0.16	0.05	0.14	0.14	-0.01	0.04	0.16	-0.39	0.26	0.61	1.00	0.11	0.28	0.34
Number of vessels fluro	0.36	0.09	0.23	0.09	0.13	0.12	0.11	-0.27	0.15	0.26	0.11	1.00	0.26	0.46
Thallium	0.11	0.39	0.26	0.13	0.03	0.05	0.01	-0.25	0.32	0.32	0.28	0.26	1.00	0.53
Heart Disease	0.21	0.30	0.42	0.16	0.12	-0.02	0.18	-0.42	0.42	0.42	0.34	0.46	0.53	1.00

Correlation analysis reveals several features with moderate linear relationships with heart disease:

- **Thallium test results (0.53)**
- **Number of vessels fluro (0.46)**
- **Chest pain type (0.42)**
- **Exercise-induced angina (0.42)**
- **ST depression (0.42)**

Maximum heart rate shows a **moderate negative correlation (-0.42),** indicating lower achievable heart rates among patients with heart disease.

Inter-feature correlations, such as between **ST depression and slope of ST (0.61)**, are present but do not indicate severe multicollinearity. This suggests that most features can be retained without destabilizing predictive models.

8. Key Insights and Interpretations

- The dataset is clean, well-structured, and suitable for predictive modeling.
- Heart disease prevalence increases with age, though age alone does not fully explain disease presence.
- Chest pain type and exercise-induced angina exhibit the strongest associations with heart disease.
- Diagnostic indicators such as ST depression, number of affected vessels, and thallium test results provide substantial predictive value.
- Cholesterol and blood pressure contribute to risk but show limited standalone predictive capability.
- Maximum heart rate is inversely associated with heart disease, reflecting reduced cardiovascular performance.
- The absence of severe multicollinearity supports the inclusion of most variables in downstream machine learning models.

9. Conclusion

This exploratory data analysis identified several clinically meaningful patterns and predictors associated with heart disease. Features related to symptoms, exercise response, and diagnostic testing demonstrate stronger predictive relevance than basic demographic variables alone. These insights provide a strong foundation for feature selection, model design, and evaluation strategies in the subsequent machine learning phase aimed at early detection of heart disease.