Cardio Pulse Guardian

A MACHINE LEARNING MODEL FOR CARDIOVASCULAR RISK DETECTION AND MANAGEMENT

COURSE PROJECT | TTDS | 2324
WAQAS DS019
EMMAD DS005

Introduction & Background

- Cardiovascular diseases (CVDs) are a leading cause of global mortality, accounting for approximately 17.9 million deaths annually

Cardiovascular Diseases (CVDs) Statistics

- CVDs contribute to 31% of all deaths worldwide.
- Four out of 5 CVD deaths result from heart attacks and strokes.
- Onethird of these deaths occur prematurely in individuals under 70 years of age.

Cardio Pulse Guardian

Cardio pulse Guardian is our response to this critical issue, presenting a machine learning model for early detection and effective management of high cardiovascular risk individuals.

Project Possible Outcomes

- **Early Detection**: Apply classification models (Logistic Regression, Gradient Boosting, Random Forest) for early identification of high cardiovascular risk individuals.
- **Effective Management**: Utilize both supervised and unsupervised learning. Supervised learning tailors personalized management, while unsupervised learning (clustering) reveals patterns for targeted interventions.
- **Global Impact**: Employ scalable solutions for largescale early risk identification (classification) and diverse health profile understanding (clustering)
- **Research Advancement**: Use advanced techniques to contribute insights into cardiovascular risk factors and management strategies. Combining supervised and unsupervised learning enhances research comprehensiveness.

Dataset Overview

- The Cardiopulse Guardian model is trained on a comprehensive dataset available at Kaggle.
- This dataset comprises 11 crucial features used to predict the likelihood of heart disease.

Key Features To Predict Likelihood of Heart Disease

1. **Age:** Age of the patient [years]
2. **Sex:** Sex of the patient [M: Male, F: Female]
3. **ChestPainType:** Chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: NonAnginal Pain, ASY: Asymptomatic]
4. **RestingBP:** Resting blood pressure [mm Hg]
5. **Cholesterol:** Serum cholesterol [mm/dl]
6. **FastingBS:** Fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. **RestingECG:** Resting electrocardiogram results [Normal: Normal, ST: having STT wave abnormality]
8. **MaxHR:** Maximum heart rate achieved [Numeric value between 60 and 202]
9. **ExerciseAngina:** Exercise induced angina [Y: Yes, N: No]
10. **Oldpeak:** Oldpeak = ST [Numeric value measured in depression]
11. **ST_Slope:** The slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]

Key Features To Predict Likelihood of Heart Disease

1. **Chest Pain Type**: Describes the kind of chest pain someone feels.

2. **Resting Blood Pressure (RestingBP):** Shows how hard the blood pushes against the vessel walls when a person is at rest.

3. **Serum Cholesterol:** Indicates the level of fat in the blood, affecting heart health.

4. **Fasting Blood Sugar (FastingBS):** Checks if blood sugar is high after not eating for a while.

5. **Resting Electrocardiogram Results (RestingECG):** Records how the heart works at rest.

6. **Maximum Heart Rate Achieved (MaxHR):** Shows the fastest heart rate during intense activity.

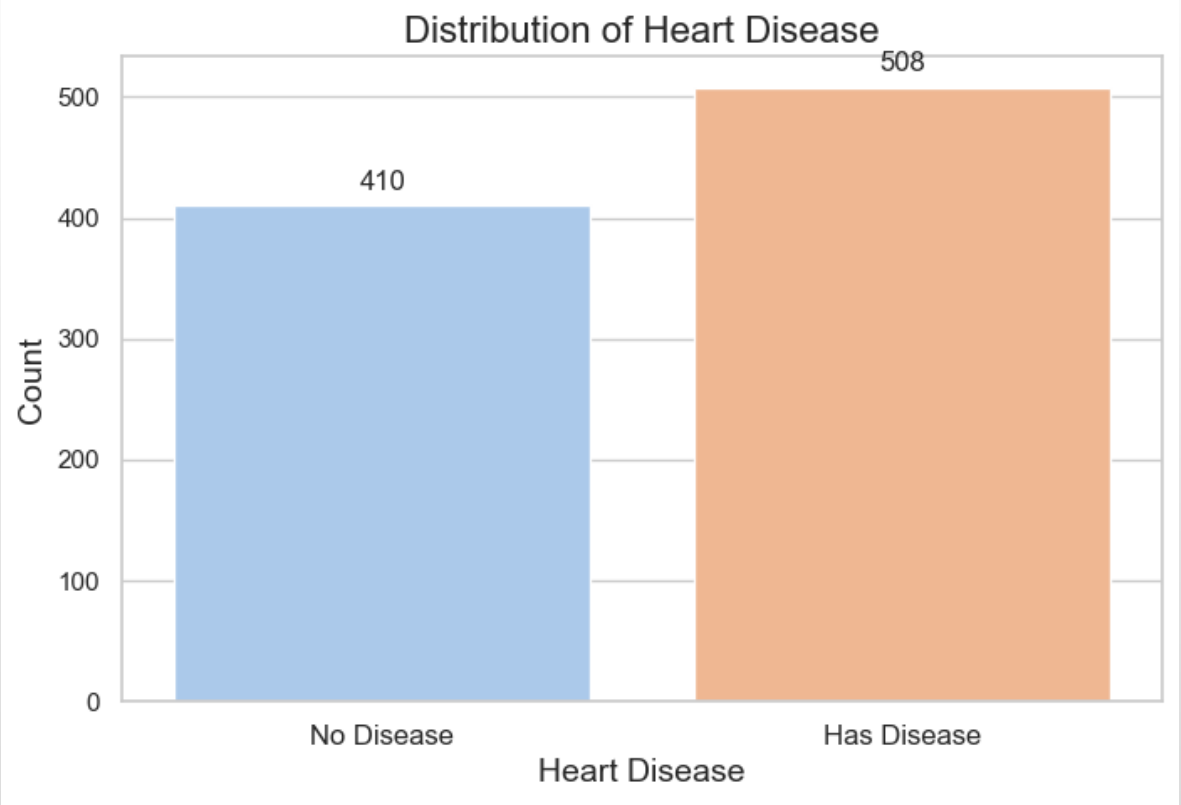
7. **Exercise-Induced Angina (ExerciseAngina):**Indicates if chest pain occurs during exercise.

8. **Oldpeak:** Describes the decrease in a specific part of the heart's electrical signal.

9. **The Slope of the Peak Exercise ST Segment (ST_Slope):** Describes the shape of the heart's electrical signal during intense activity.

Target Variable

- Our goal is to predict the likelihood of heart disease (HeartDisease) based on the provided dataset.



Data Preprocessing

DATA CLEANING AND FEATURE TRANSFORMATION

Dataset Composition

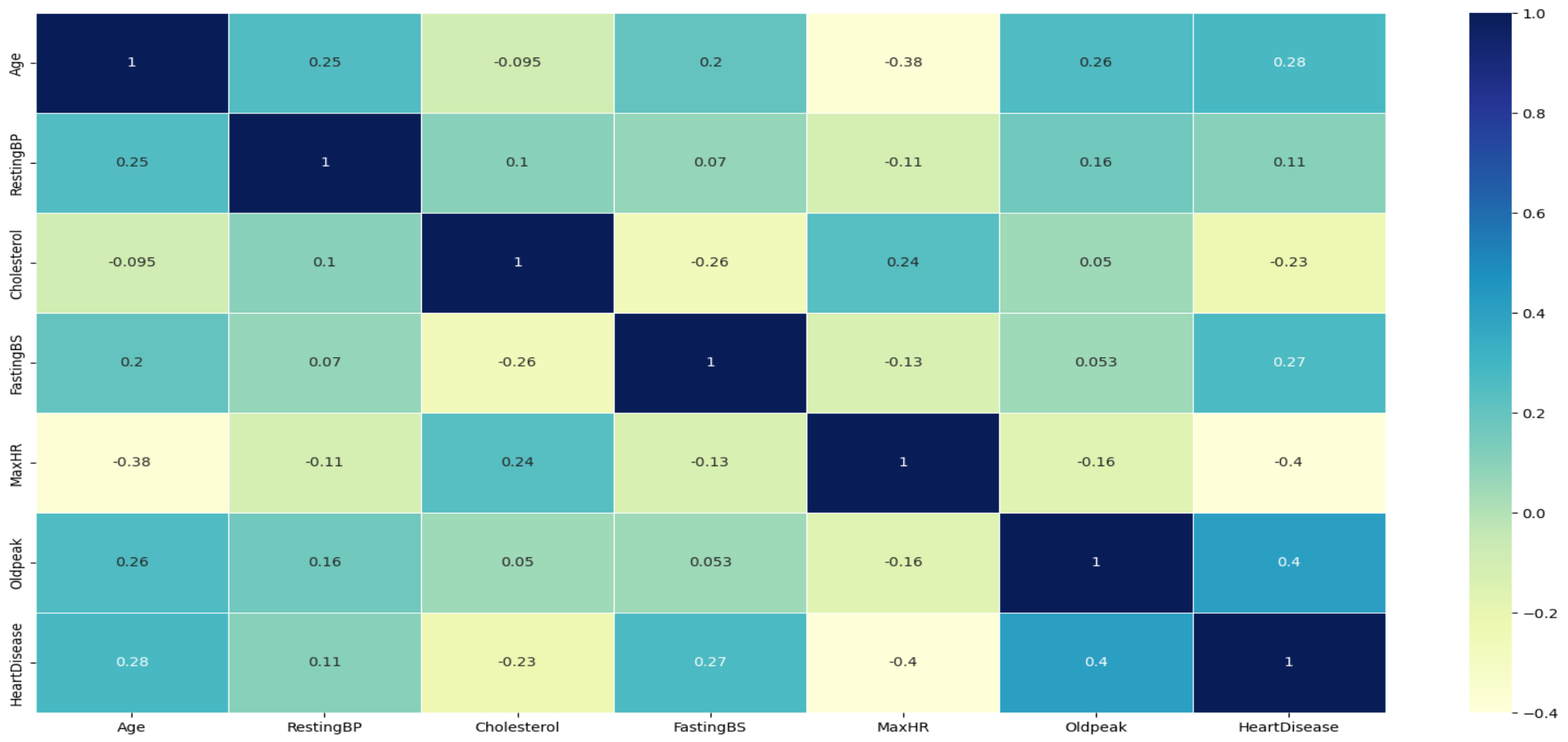
- 12 variables in total (1 dependent, 11 independent)
- 7 numerical features: Age, RestingBP, Cholesterol, MaxHR, Oldpeak
- 5 categorical features: Gender, ChestPainType, FastingBS, RestingECG, ExerciseAngina, ST_Slope
- No apparent missing values under normal circumstances.
- Caution: Potential hidden missing values.

Steps For Data Cleaning

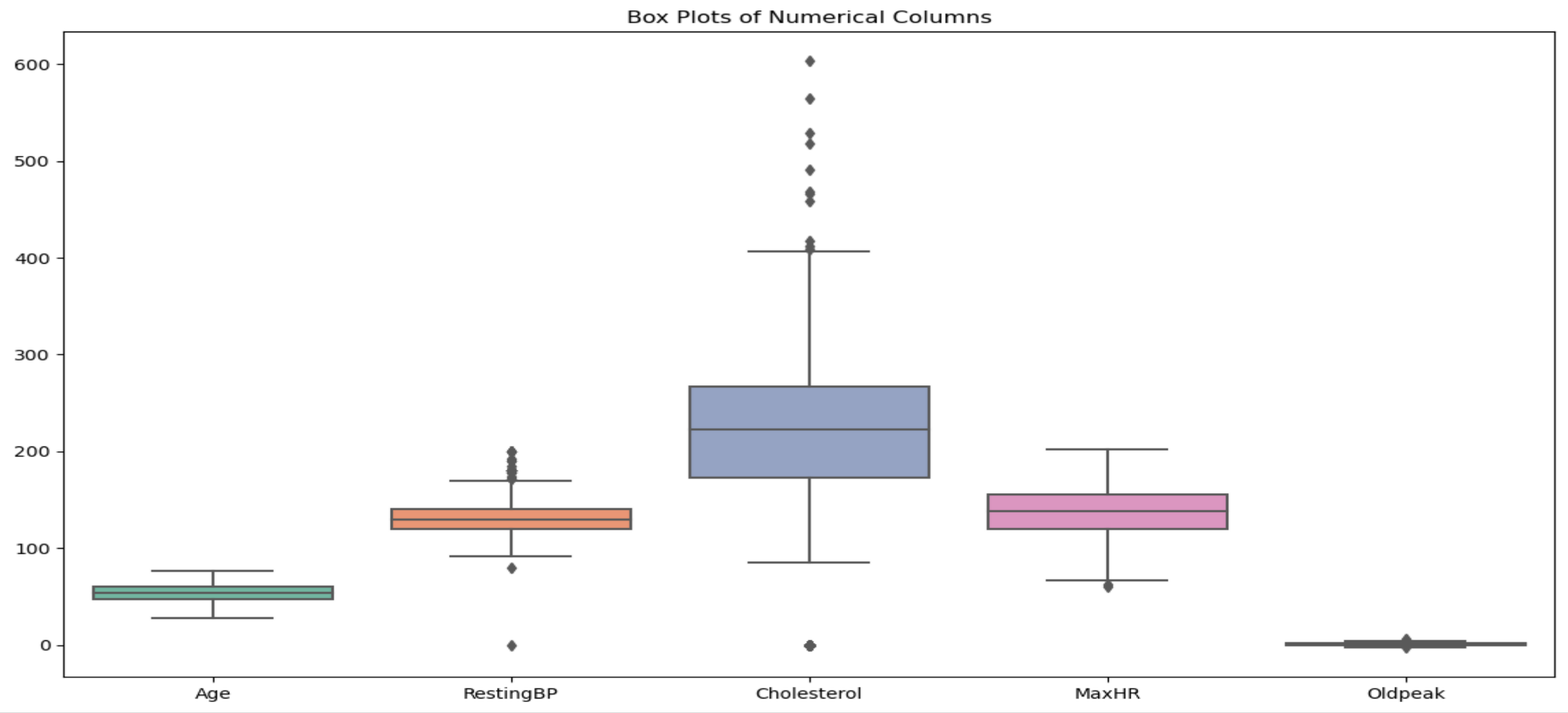
Performing data cleaning to get the data ready for the next step. It involves the following steps:

- Fix missing values (Imputation can be considered)
- Remove duplicate values (where needed)
- Check Data Types beforehand

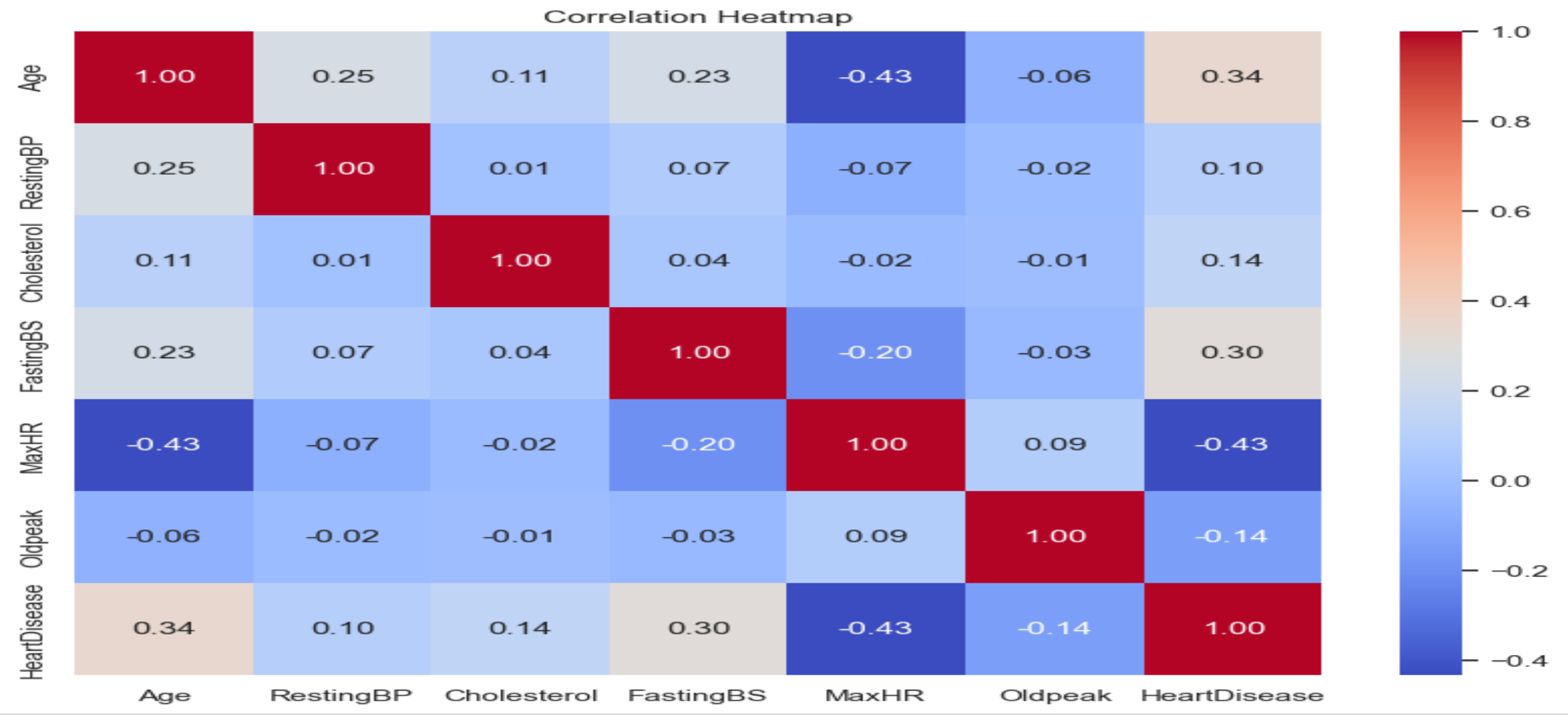
Correlation Heatmap



Box Plot of Numerical Feature Indicating Outliers



Correlation Heatmap After Removing Outlier



Observation

- We have not found any significant changes in correlation before and after removing outlier.
- We prefer to go with our complete dataset.

Imputation of Missing Values

- Initial data check revealed no missing values.
- Noticed 0 values in numerical columns (["Age," "RestingBP," "Cholesterol," "MaxHR"]).
- Treated 0 values as potential null values within practical ranges.
- Replaced 0 with NaN and identified three columns with null values.
- Utilized imputation techniques to address missing values in these columns.

Exploratory Data Analysis

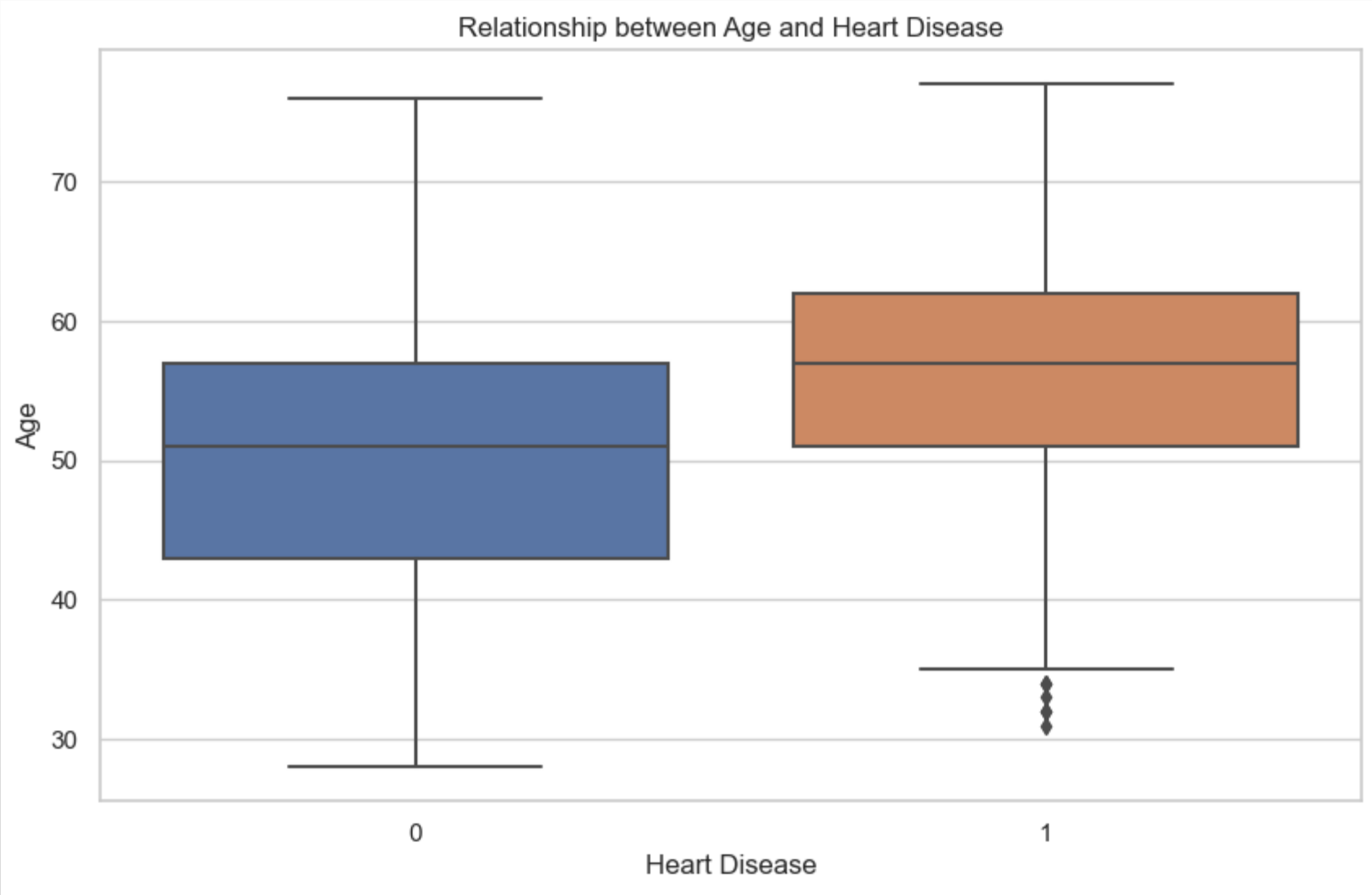
VISUALIZING KEY PARAMETER AND FINDING THEIR TRENDS AND INSIGHT

Exploring Age Group and Heart Disease Methodology

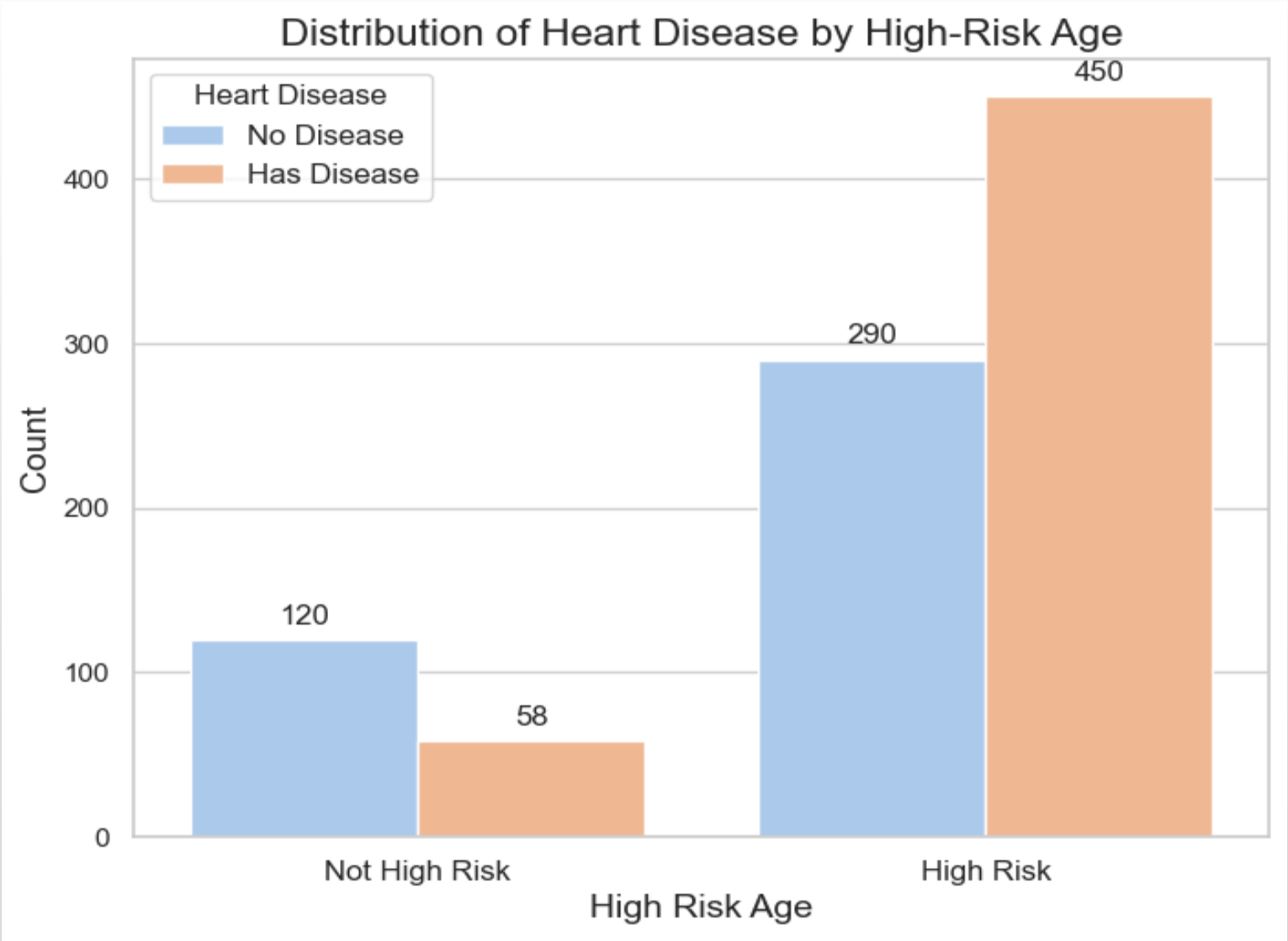
The relation between the age groups and cardiovascular disease will be found by :

- investigating the impact of age on cardiovascular risk by categorizing individuals into different age groups.
- Analyzing age distributions, prevalence of cardiovascular risk factors, and correlations with other health indicators to gain insights into how age influences the likelihood of heart disease in the dataset.

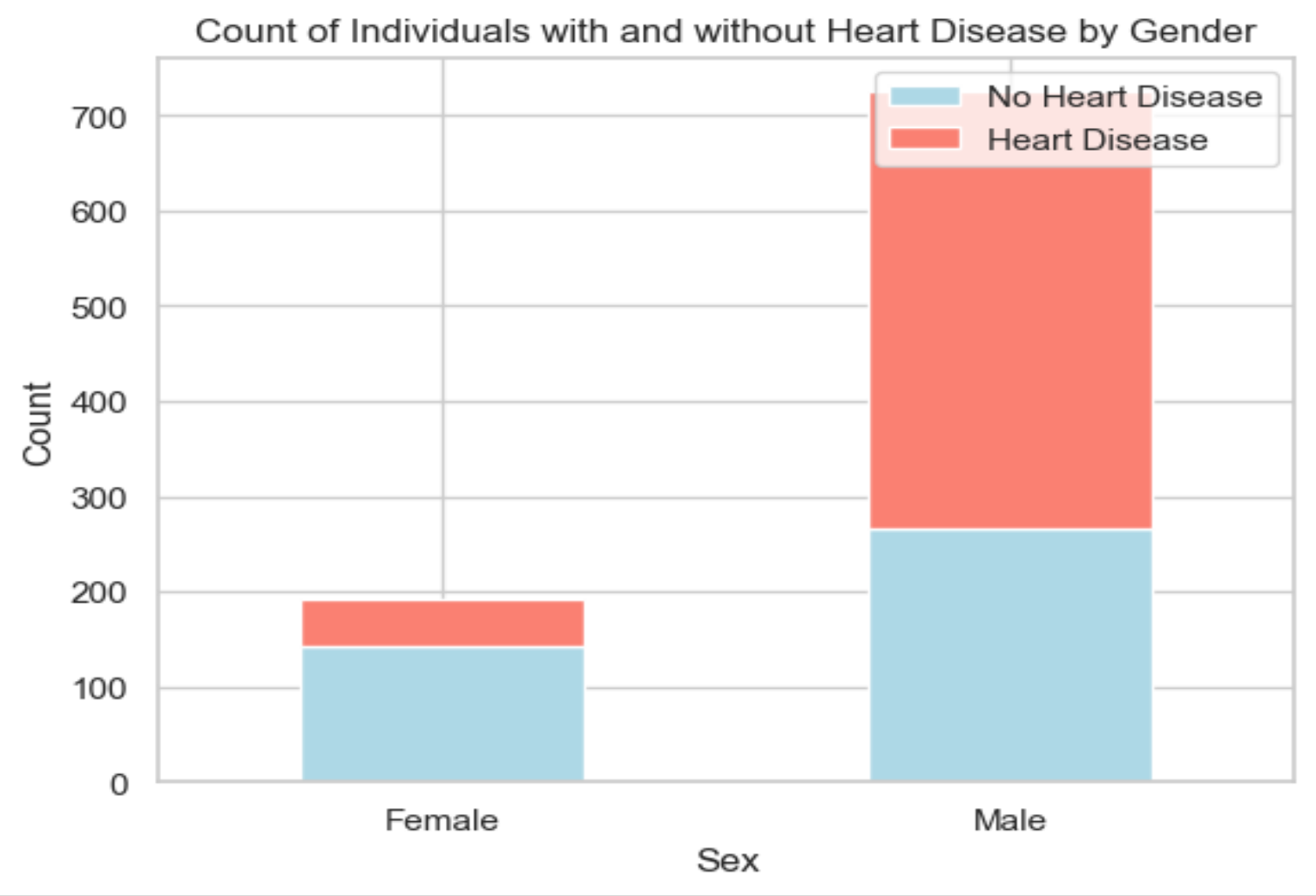
Exploring Age Group and Heart Disease



Exploring High Risk Age Group and Heart Disease



Heart Disease Relation with Gender

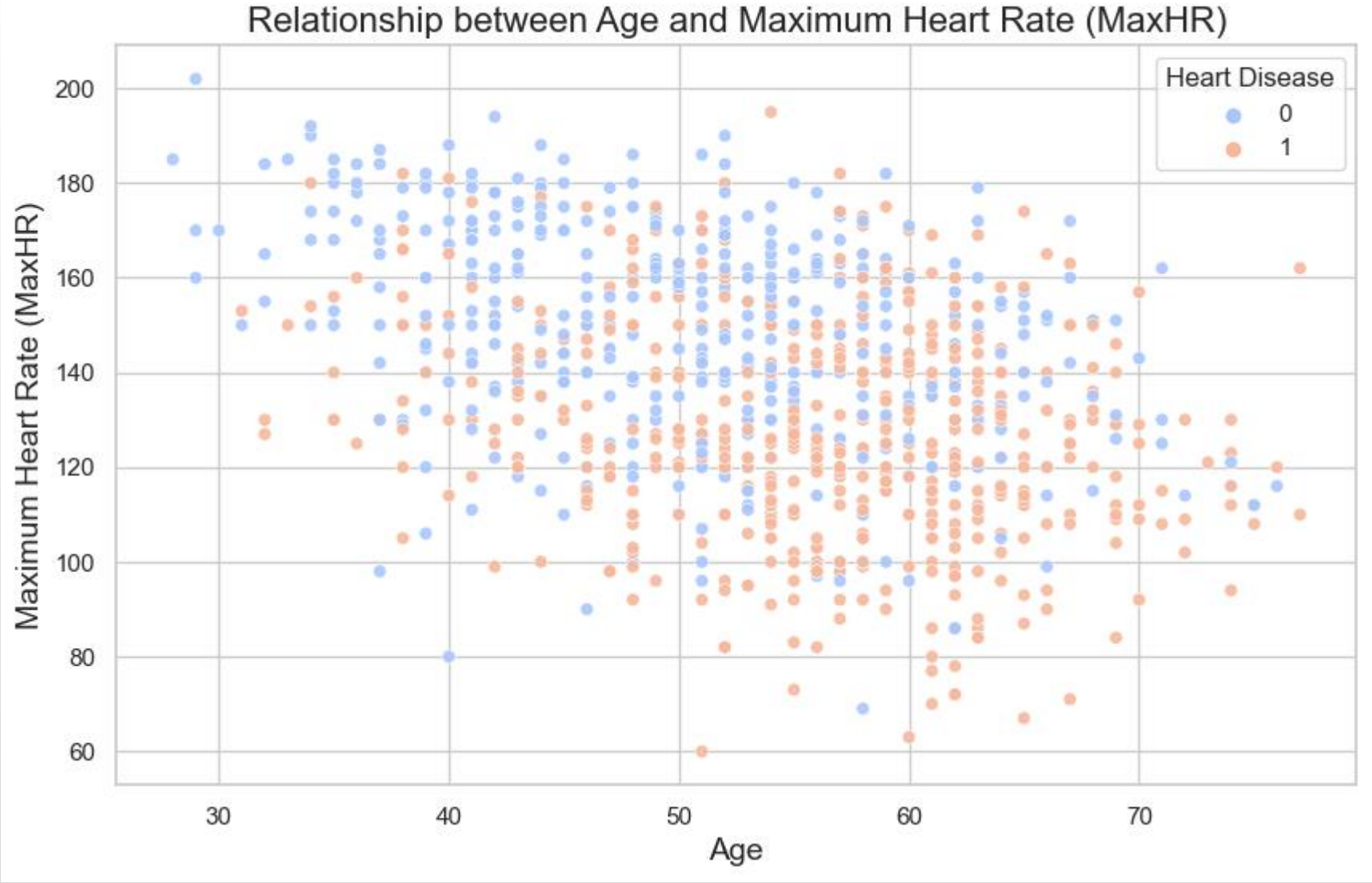


Outcome Of Heart Disease Incidence in Male

- Despite a substantial representation of males in the dataset, the observed percentage of individuals with heart disease among males remains high.
- This suggests a potential gender-related pattern in heart disease incidence, warranting further investigation and consideration in the overall analysis.

Relationship Between Cholesterol and Age

Relationship Between Max Heart Rate and Age



Insights

Age vs. Cholesterol Levels:

- The scatter plot reveals a varied distribution of cholesterol levels across different age groups. Individuals with heart disease tend to exhibit a diverse range of cholesterol levels, suggesting that cholesterol alone may not be a decisive factor in cardiovascular risk.

Age vs. Maximum Heart Rate (MaxHR):

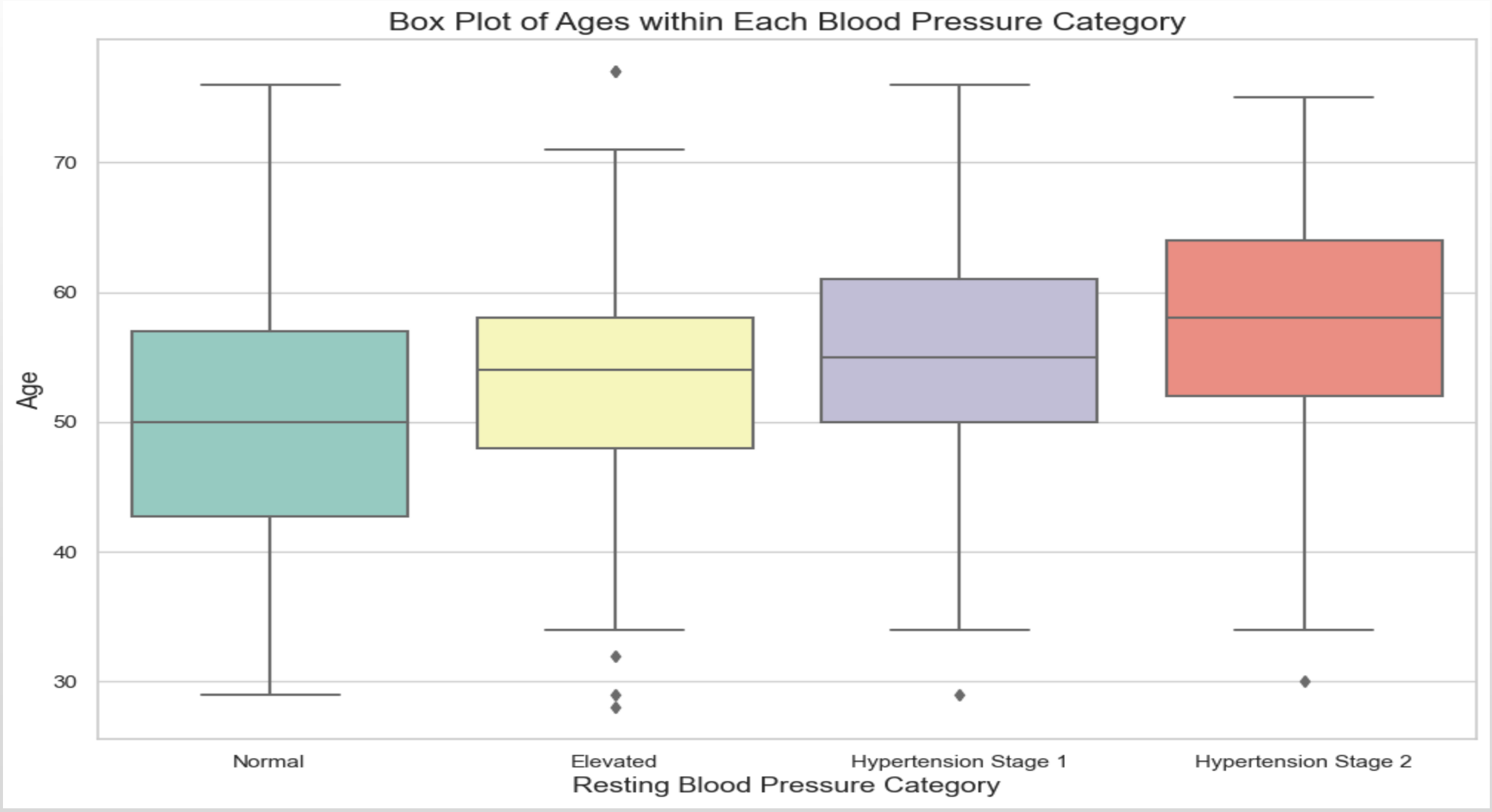
- The scatter plot illustrates the relationship between age and maximum heart rate (MaxHR) for individuals with and without heart disease. There seems to be a subtle negative correlation, with older individuals generally showing a lower maximum heart rate, particularly among those with heart disease. The distribution of MaxHR varies significantly, emphasizing the need for further investigation into factors influencing cardiovascular fitness with age.

Analyzing Resting Blood Pressure with Age

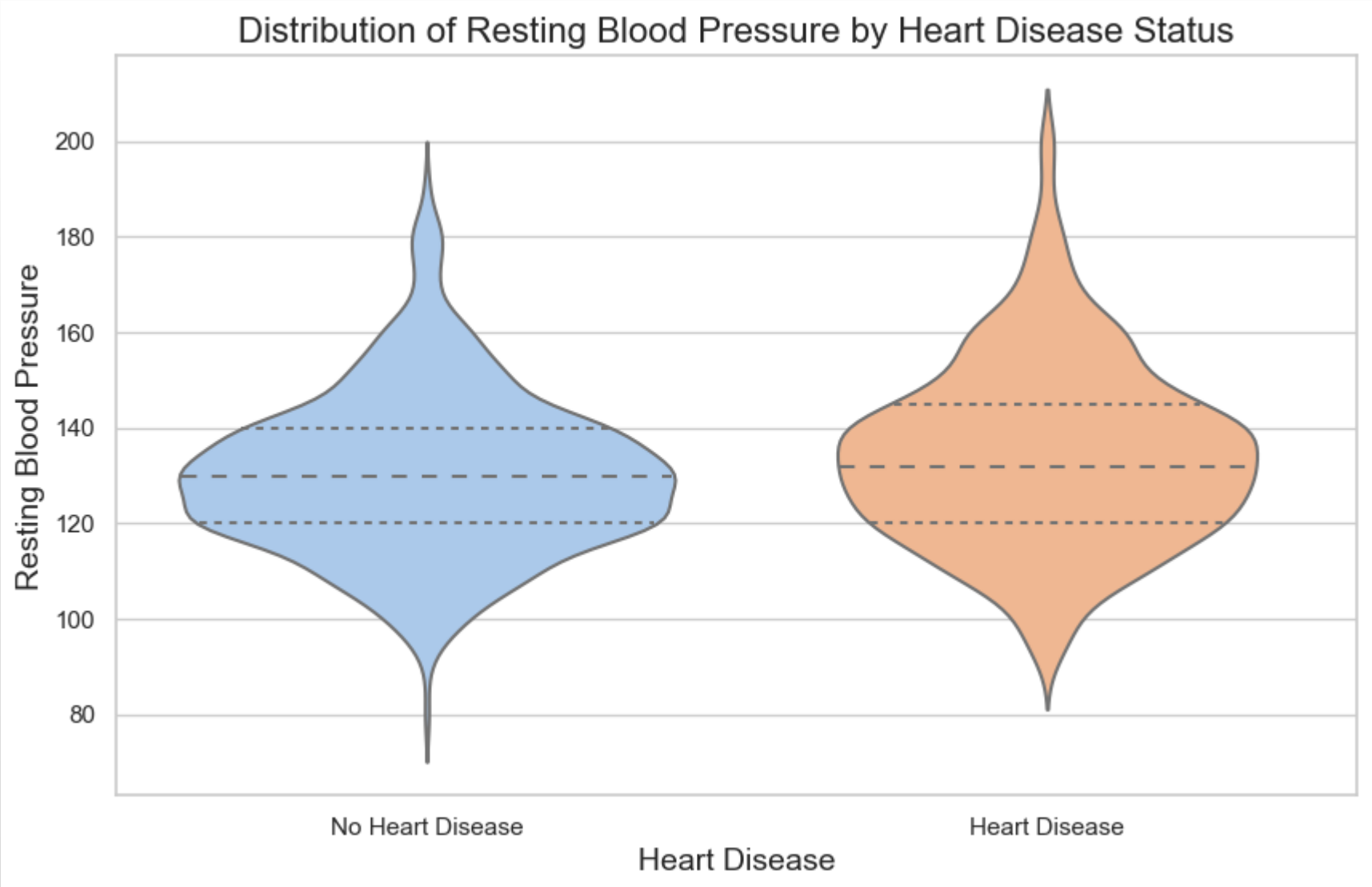
In order to analyze resting blood pressure with age, following procedure was followed:

- Exploring the distribution and impact of resting blood pressure on cardiovascular health.
- Investigating statistical measures, visualize the relationship between blood pressure and heart disease, and identify patterns that may contribute to risk assessment.

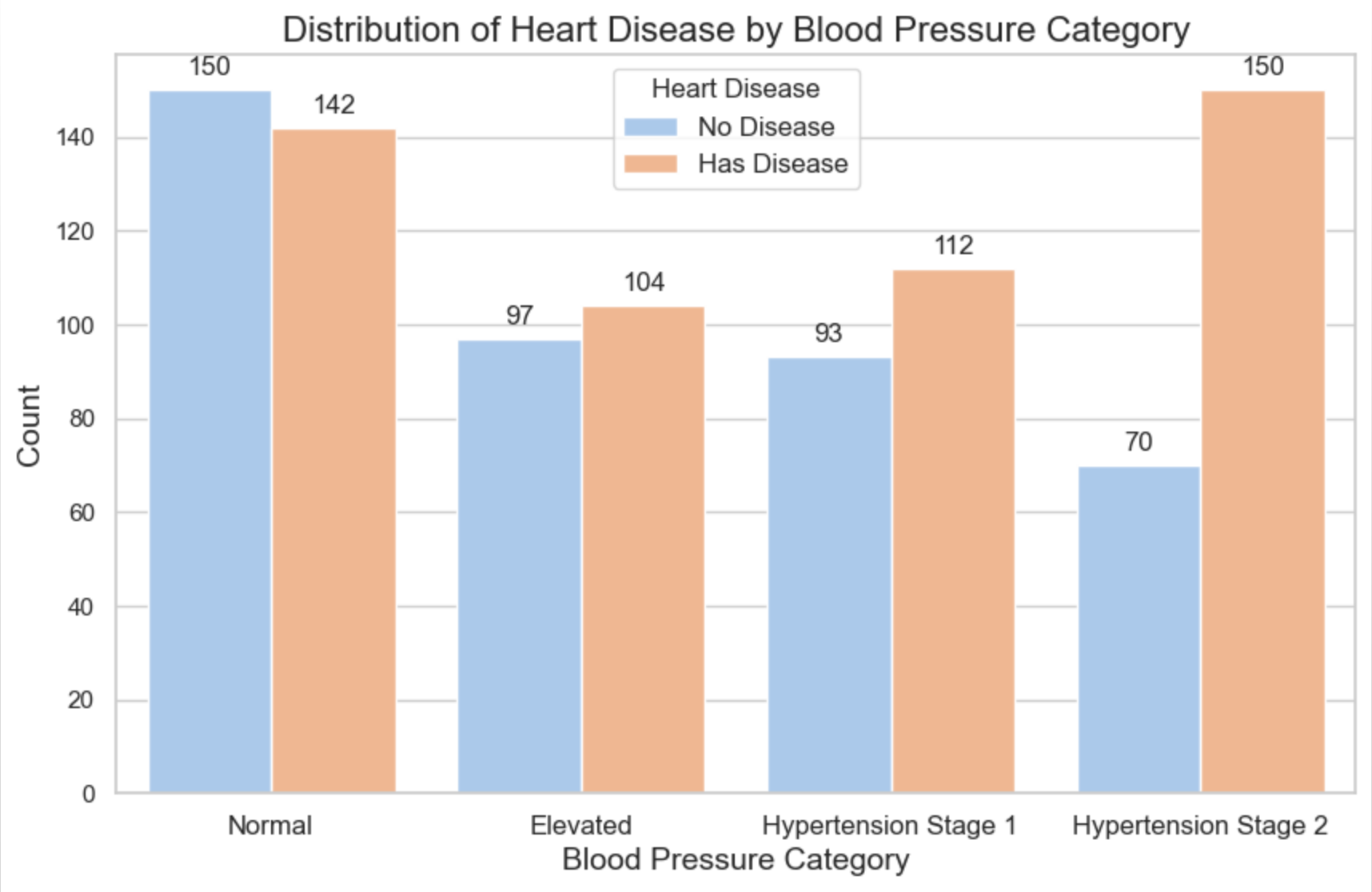
Analyzing Resting Blood Pressure with Age



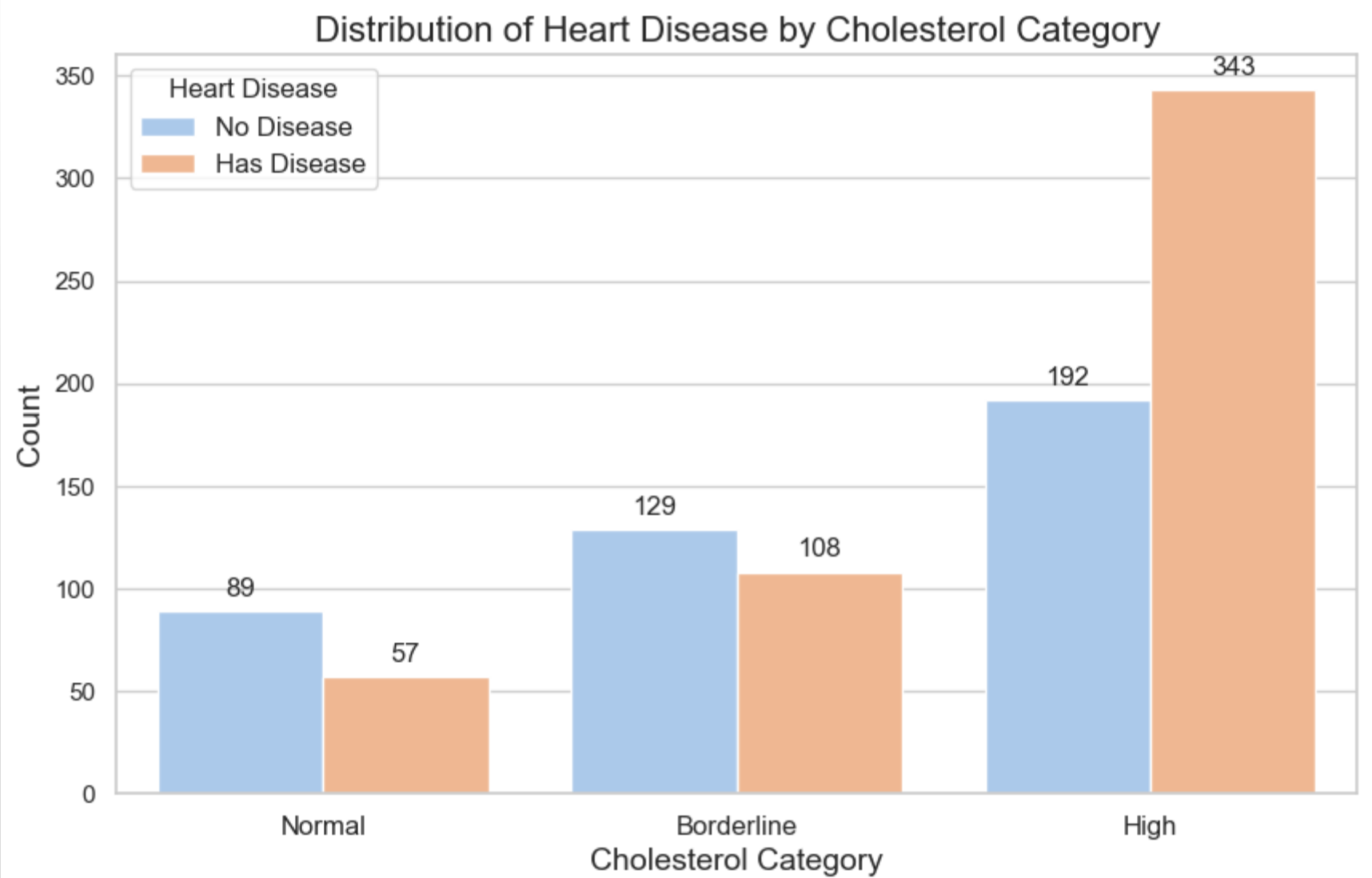
Distribution of Resting Blood Pressure And Heart Disease



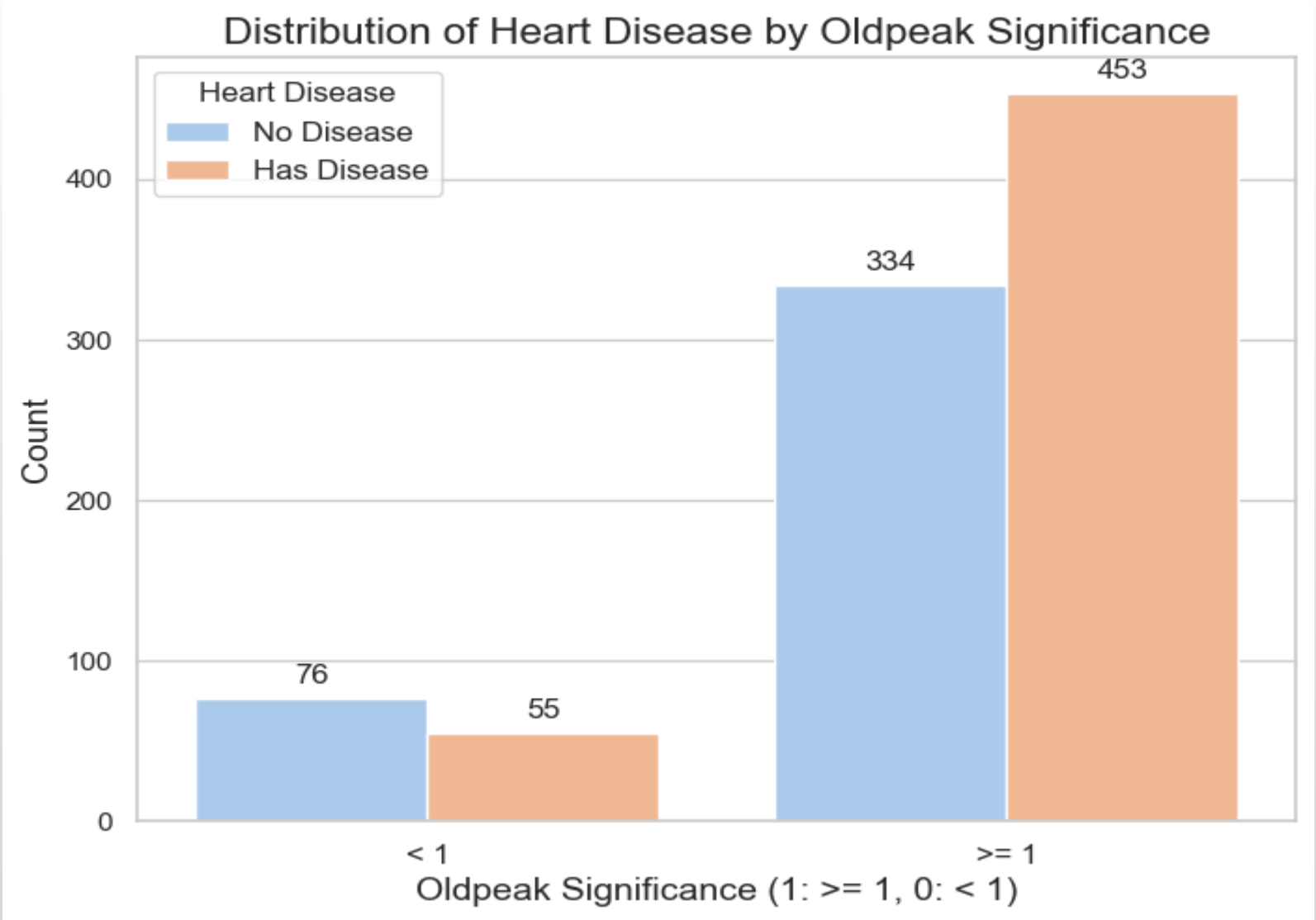
Distribution of Heart Disease By Blood Pressure Category



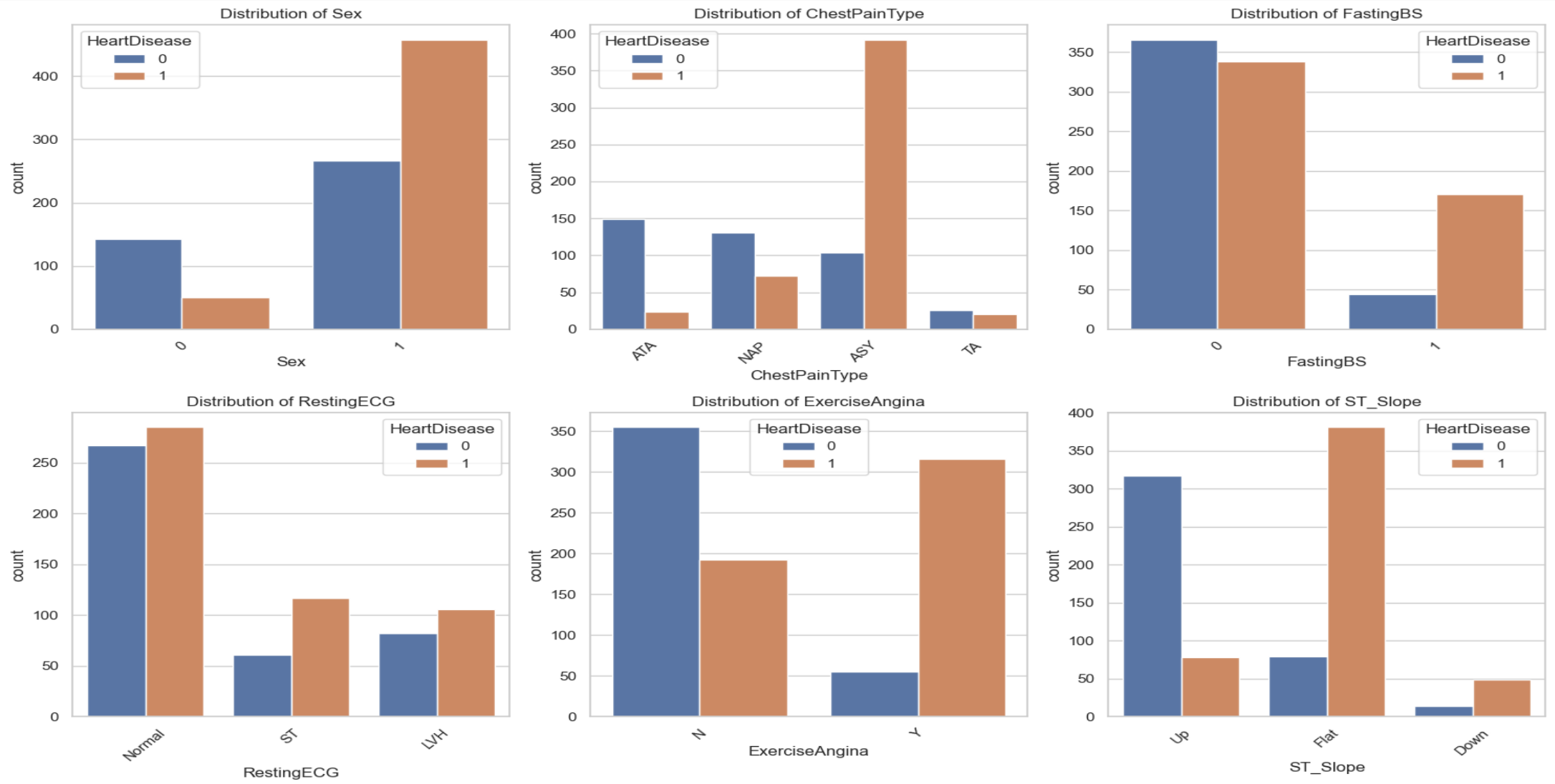
Distribution of Heart Disease By Cholestrol Category



Distribution of Heart Disease By Old peak



Analyzing Categorical Features



Correlation of Categorical Feature By Cramér's V Method

	Categorical Feature	Correlation with HeartDisease
5	ST_Slope	0.621249
1	ChestPainType	0.537639
4	ExerciseAngina	0.491208
0	Sex	0.301114
2	FastingBS	0.262775
3	RestingECG	0.098679

ChiSquare Test for Independence

```
Chi-square test for independence between Sex and HeartDisease:
Chi2: 84.15, p-value: 4.5976174508091635e-20
There is a significant association.
========================================
Chi-square test for independence between ChestPainType and HeartDisease:
Chi2: 268.07, p-value: 8.08372842808765e-58
There is a significant association.
========================================
Chi-square test for independence between FastingBS and HeartDisease:
Chi2: 64.32, p-value: 1.0573018731809955e-15
There is a significant association.
========================================
Chi-square test for independence between RestingECG and HeartDisease:
Chi2: 10.93, p-value: 0.0042292328167544925
There is a significant association.
========================================
Chi-square test for independence between ExerciseAngina and HeartDisease:
Chi2: 222.26, p-value: 2.907808387659878e-50
There is a significant association.
========================================
Chi-square test for independence between ST_Slope and HeartDisease:
Chi2: 355.92, p-value: 5.167637689470128e-78
There is a significant association.
========================================
```

Machine Learning Model

- We selected three machine learning model for our heart disease predictive model.

1. Random Forest
2. Gradient Boosting
3. Logistic Regression

Machine Learning Model

- We applied our ML model in three different ways.

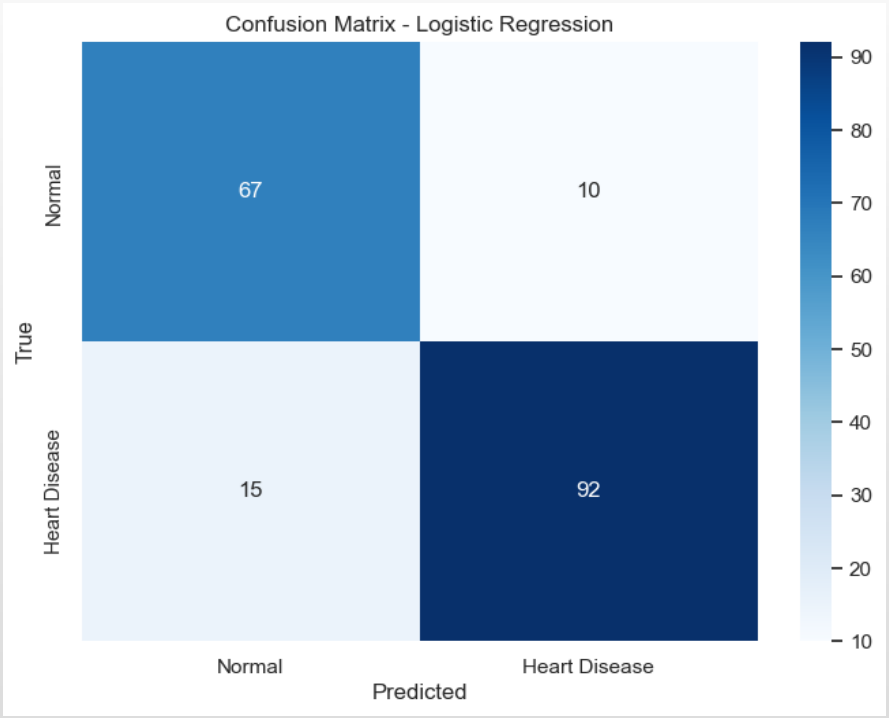
1. On over all data
2. By selecting feature
3. By selecting feature with cross validation

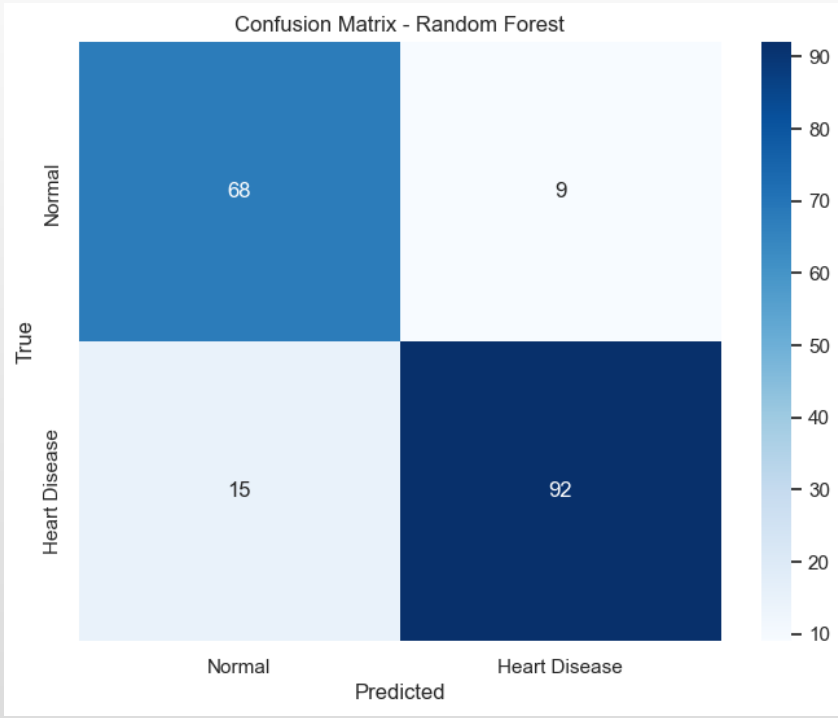
Let’s discuss performance of each method one by one.

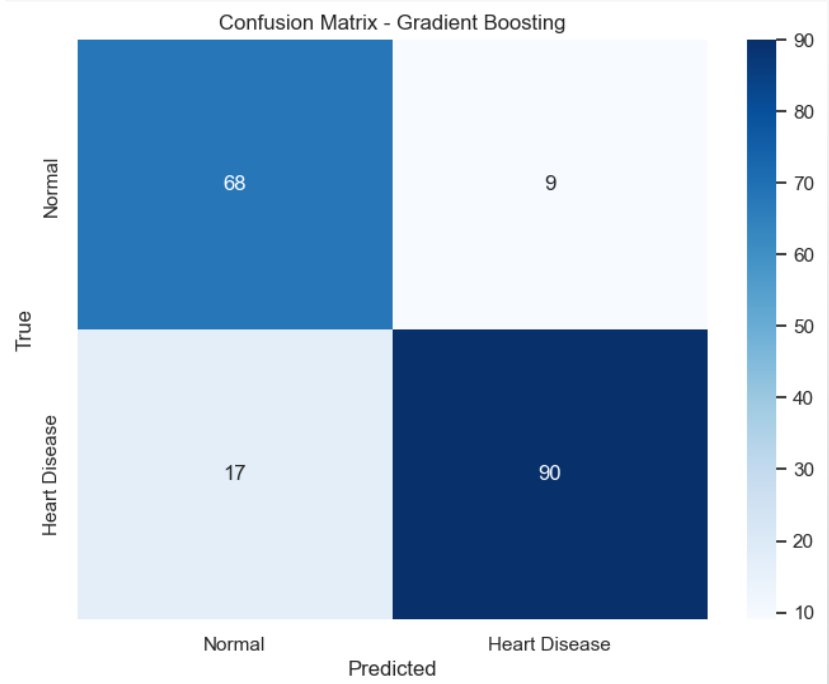
Machine Learning Model On Over All Data

Machine Learning Model	Model Accuracy
Logistic Regression	86.4 %
Random Forest	86.5%
Gradient Boosting	85.86%

Confusion Matrices







Machine Learning Model On Selected Feature

Following are the categorical feature we selected for our model.

1. ChestPainType
2. RestingECG
3. ExerciseAngina
4. ST_Slope
5. RestBP_Category
6. CholesterolCategory
7. MaxHRCategory

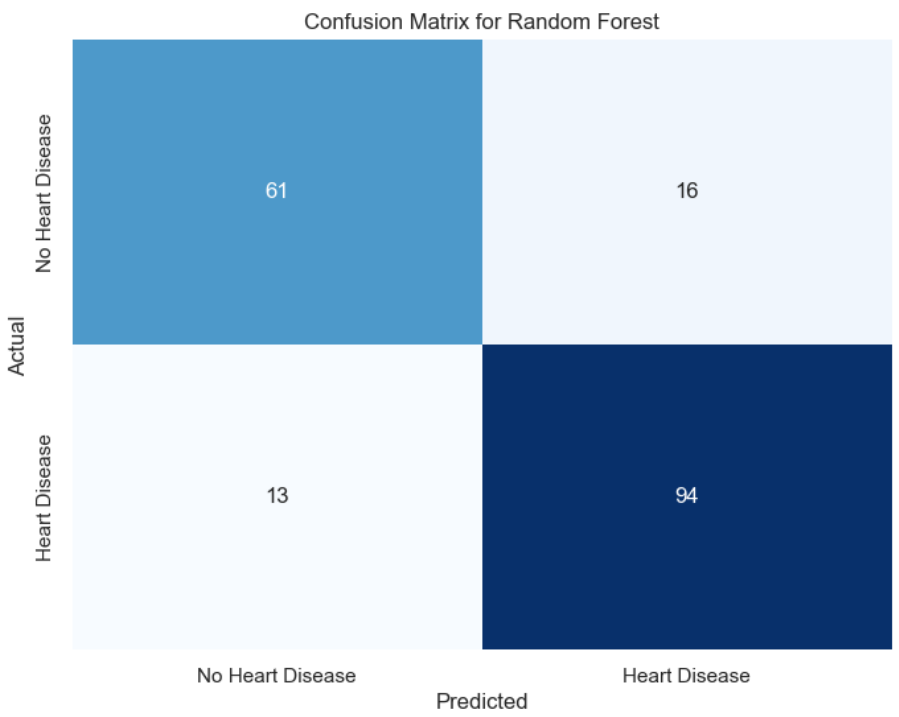
Following are the numerical feature we selected for our model.

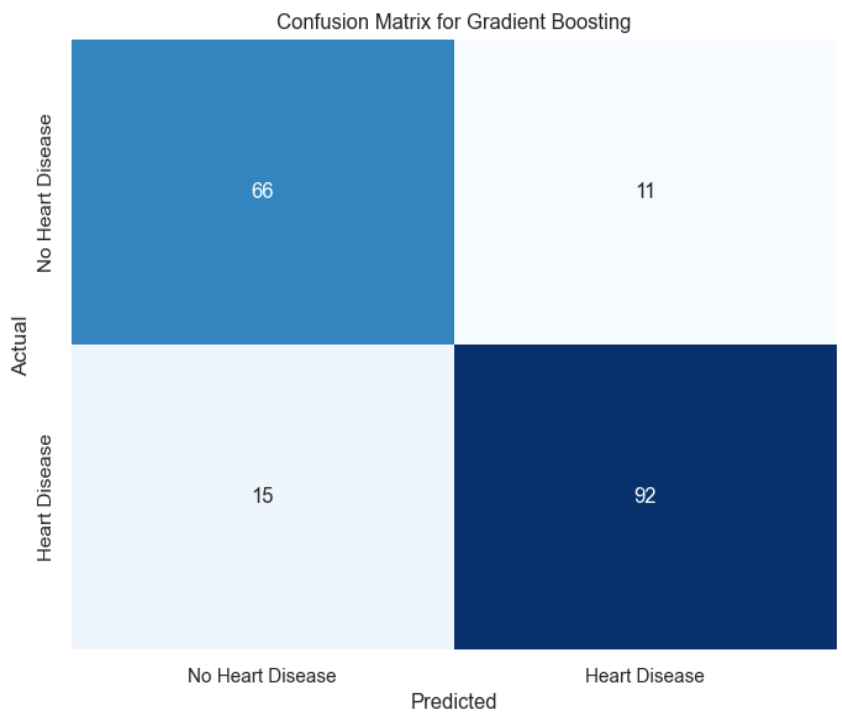
1. Gender
2. FastingBS
3. Oldpeaksignificant

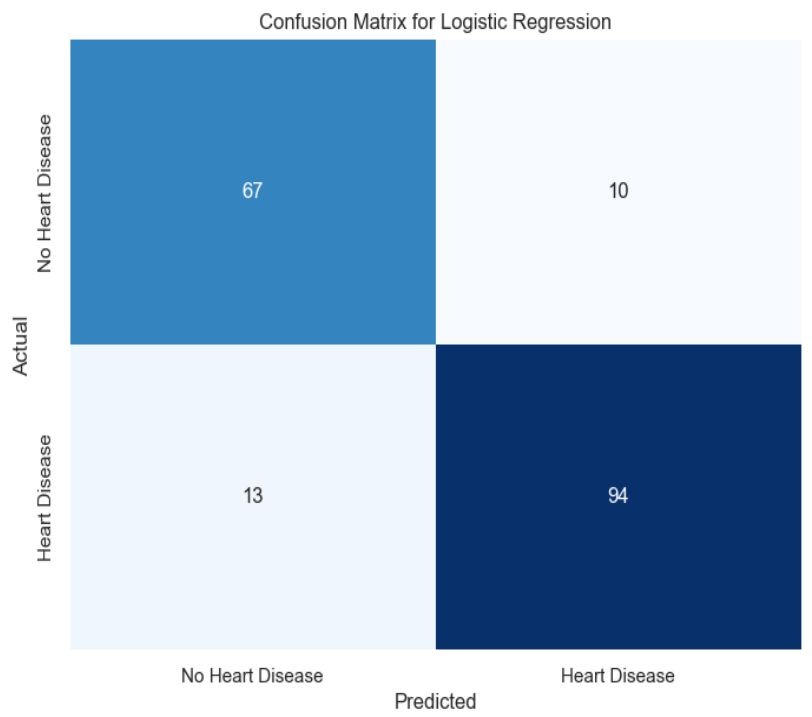
Machine Learning Model On Selected Features

Machine Learning Model	Model Accuracy
Logistic Regression	87.5 %
Random Forest	84.24%
Gradient Boosting	85.87%

Confusion Matrices



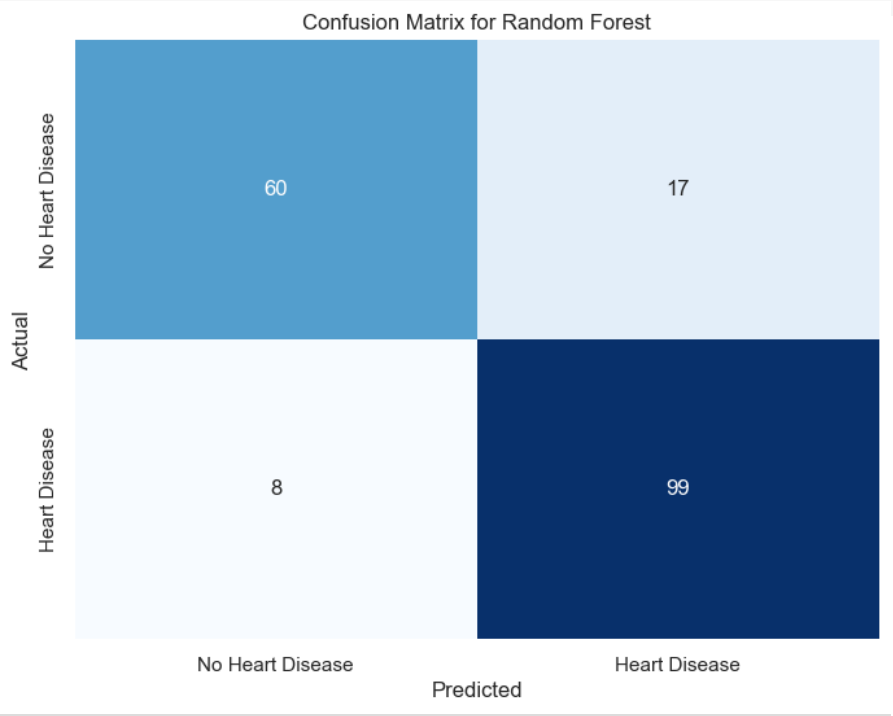


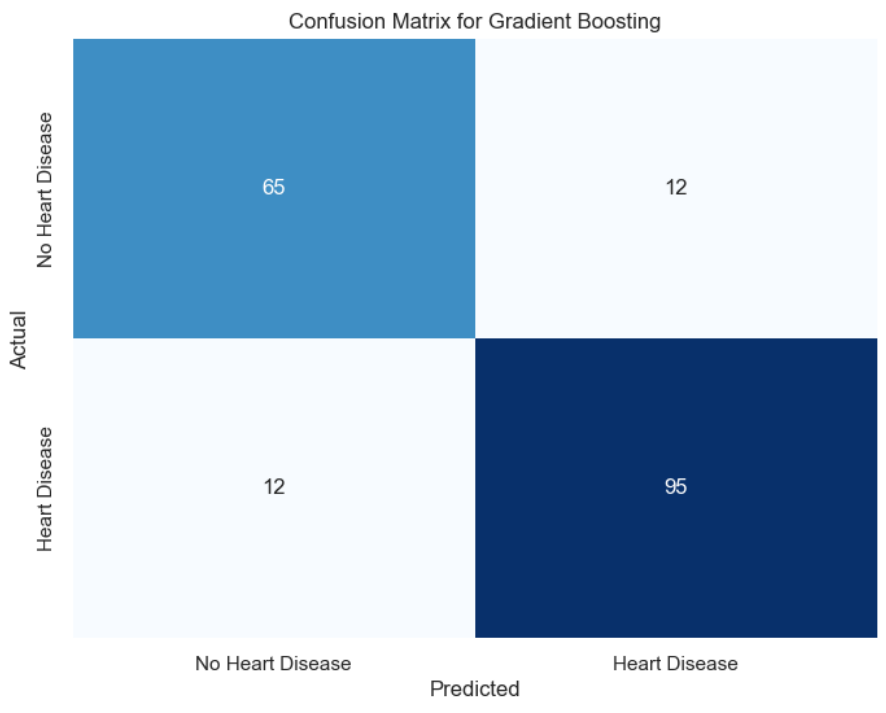


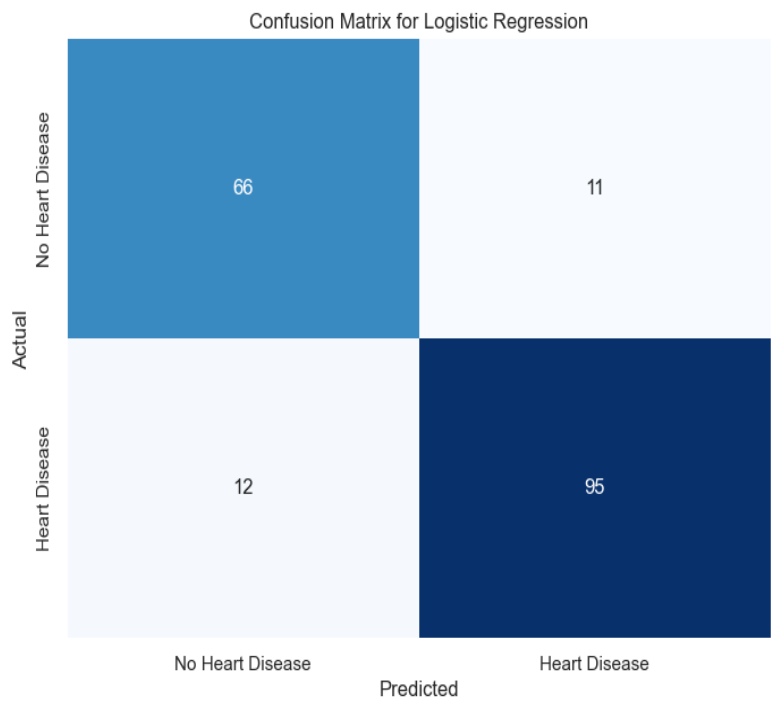
Machine Learning Model On Selected Features With Cross Validation

Machine Learning Model	Model Accuracy
Logistic Regression	87.5 %
Random Forest	86.41%
Gradient Boosting	86.96%

Confusion Matrices







Conclusion

- As recall value of Random Forest with feature selection and cross validation is 93% for predicting presence of heart disease in heart patient, which is also greater than any other model. Hence, we chose this model as our final model.

Future Work

- We will try to gather more data to increase accuracy of model.
- We will acquire more domain knowledge for further feature selection.
- We will consider disease to incorporate in our machine learning model.


THANK YOU
ANY QUESTIONS?