**Exploratory Data Analytics of Heart Disease Dataset**

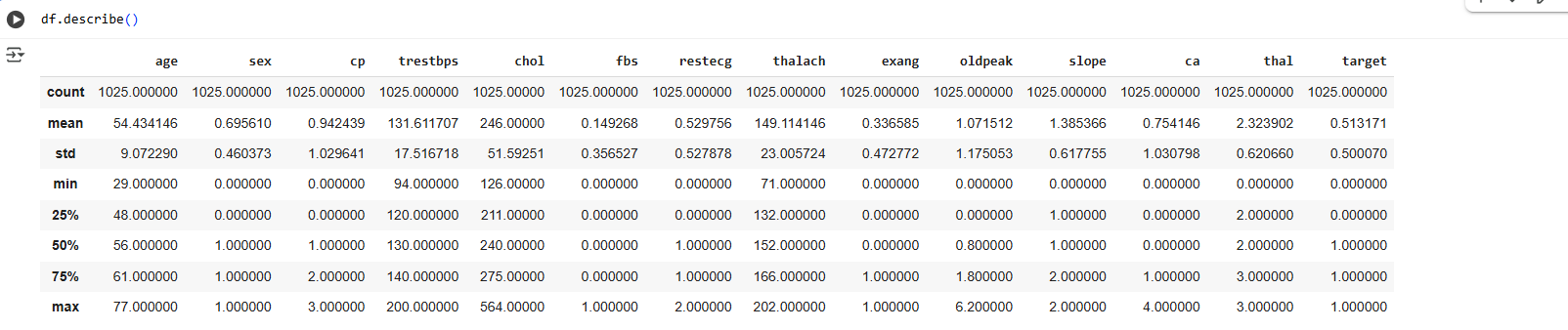
**Members: Shayan Khan (2244554), Kelvin Asah (2311635)**

**Data Analytics of Heart Disease Dataset**

**Dataset:**

* Source: Heart Disease Dataset (heart.csv) from Kaggle
* Historical data collected since 1988 from four databases: Cleveland, Hungary, Switzerland, and Long Beach V.
* Labeled dataset with 1,025 rows and 14 columns, where the 14th column contains the classification output for supervised learning.
  + An integer value of 0 indicates no heart disease.
  + An integer value of 1 indicates heart disease.
* Attributes:
  + Age
  + Sex
  + Chest paint type (4 values)
  + Resting blood pressure
  + Serum cholesterol in mg/dl
  + Fasting blood sugar > 120 mg/dl
  + Resting electrocardiographic results (values 0,1,2)
  + Maximum heart rate achieved
  + Exercise-induced angina
  + Oldpeak = ST depression induced by exercise relative to rest
  + The slope of the peak exercise ST segment
  + Number of major vessels (0-3) colored by fluoroscopy
  + Thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

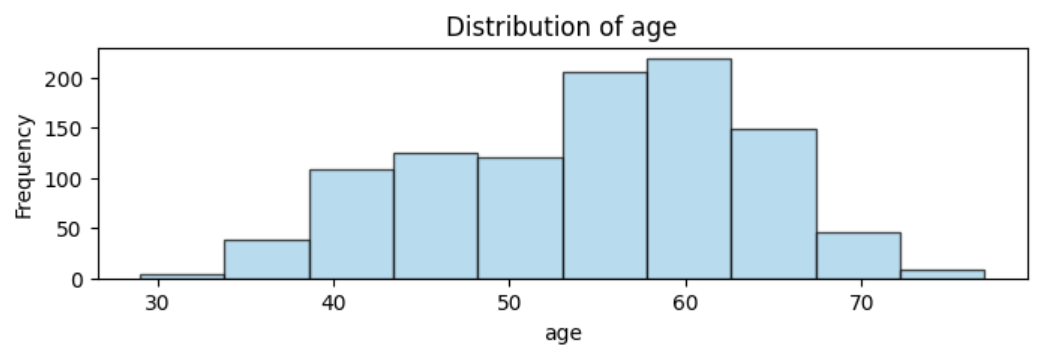
**Summary Statistics:**



The descriptive statistics sum up the dataset's numerical variables' central tendency and distribution. Patients range in age from 29 to 77 years, with an average age of approximately 54. The range of resting blood pressure is 94 mmHg to 200 mmHg, with an average of 131 mmHg. There are significant variations in the lipid levels of patients, with cholesterol levels ranging from 126 mg/dL to 564 mg/dL. The "oldpeak" variable, which measures ST depression brought on by activity, has an average of 1.07 beats per minute, and the average maximum heart rate attained (thalach) is 149 beats per minute.

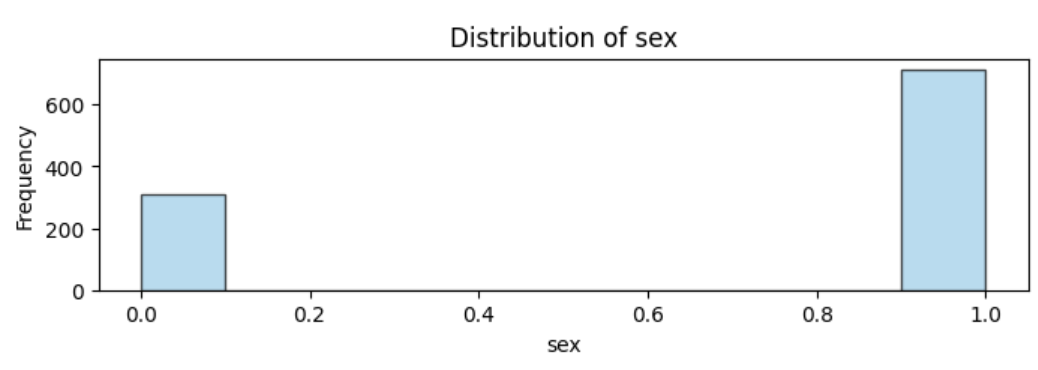
These numbers demonstrate that the dataset contains patients with a wide range of health statuses, from those with substantial cardiovascular risk factors to those who are generally healthy. The significant differences in blood pressure and cholesterol suggest that certain people are far more likely than others to experience heart-related problems.

**Visualize Univariate Distributions:**

****

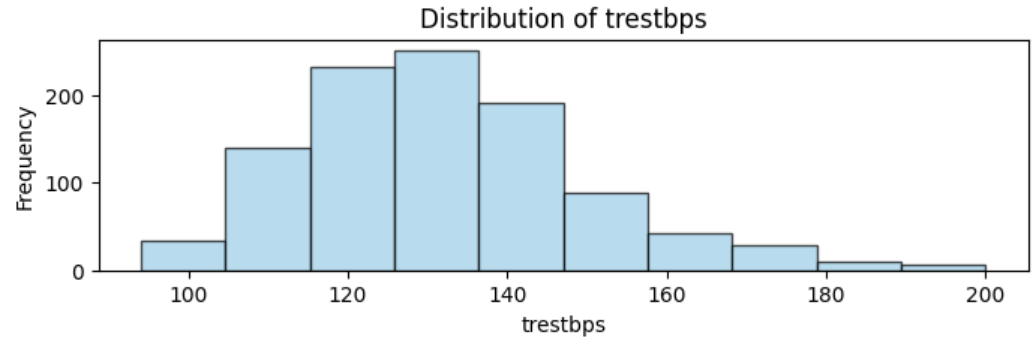
**1. Histogram of Age**

The distribution of patient ages throughout the sample is shown by the age histogram. The majority of patients are between the ages of 45 and 65, giving the distribution an approximately bell-shaped appearance. This suggests that the majority of the sample consists of middle-aged adults. Given that the risk of heart disease rises sharply with age, the concentration of patients in this age group is to be expected. Middle-aged people are the main group in this dataset impacted by heart-related diseases, as evidenced by the lower numbers of younger patients under 40 and older people over 70.



**2. Histogram of Sex**

The gender distribution in the dataset is shown by the sex histogram. In this case, males are represented by the number 1, and females by the number 0. The graph indicates that men make up the bulk of the patients. This disparity mirrors actual trends in which males are more likely than women to develop cardiac disease and to do so earlier in life. The outcome highlights the significance of gender in assessing and forecasting the risk of heart disease.

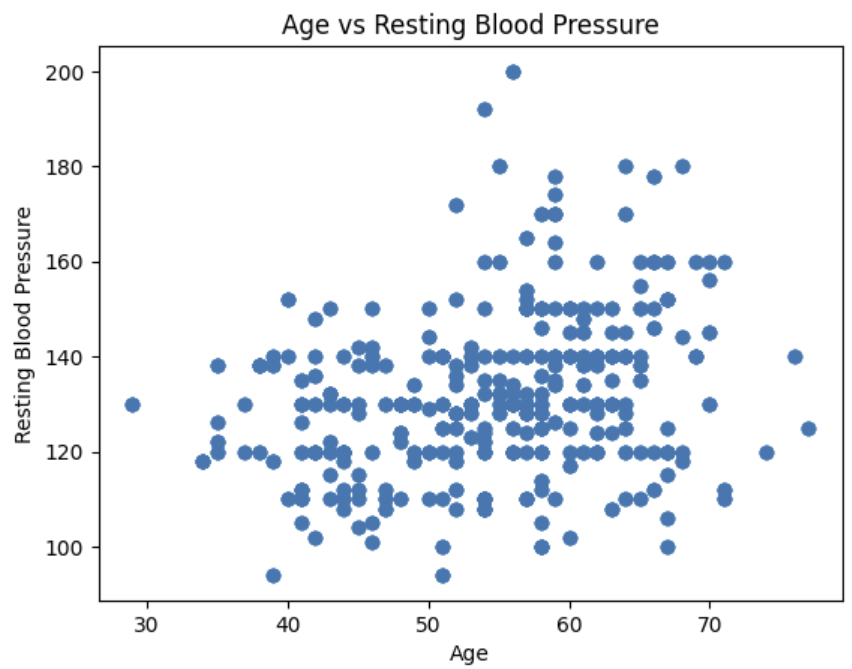


**3. Histogram of Resting Blood Pressure**

The distribution of patients' blood pressure values is shown by the resting blood pressure histogram (trestbps). A lesser percentage of patients have resting blood pressure readings above 160 mmHg, whereas the majority have readings between 120 and 140 mmHg.This implies that moderate hypertension, a prevalent and important risk factor for heart disease, is present in a significant proportion of individuals. Some patients may have severe hypertension and be at higher risk for cardiovascular disease, as indicated by the few extreme results.

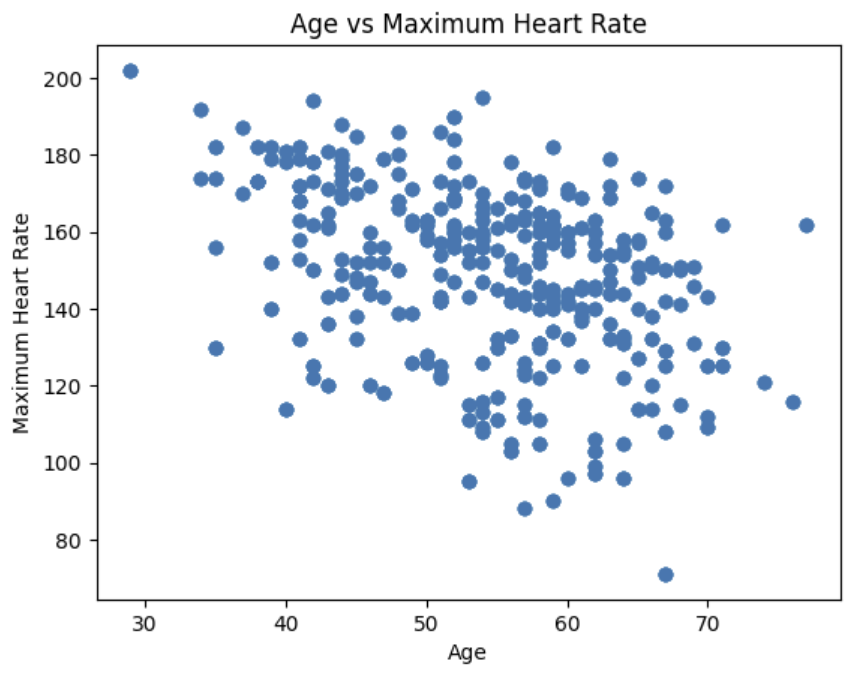
**Relationships Between Variables:**

Out of the 13 features, only 4 were numerical values (age, trestbps, chol, thalach), while the rest were categorical features with discrete values represented by small numbers. This made it harder to identify clear linear relationships, since scatterplots are not always meaningful for categorical data. So, for the exploratory analysis, scatterplots were used for numerical features, while countplots, cross-tabs, or barplots were used for categorical features to show clear linear relationships.

****

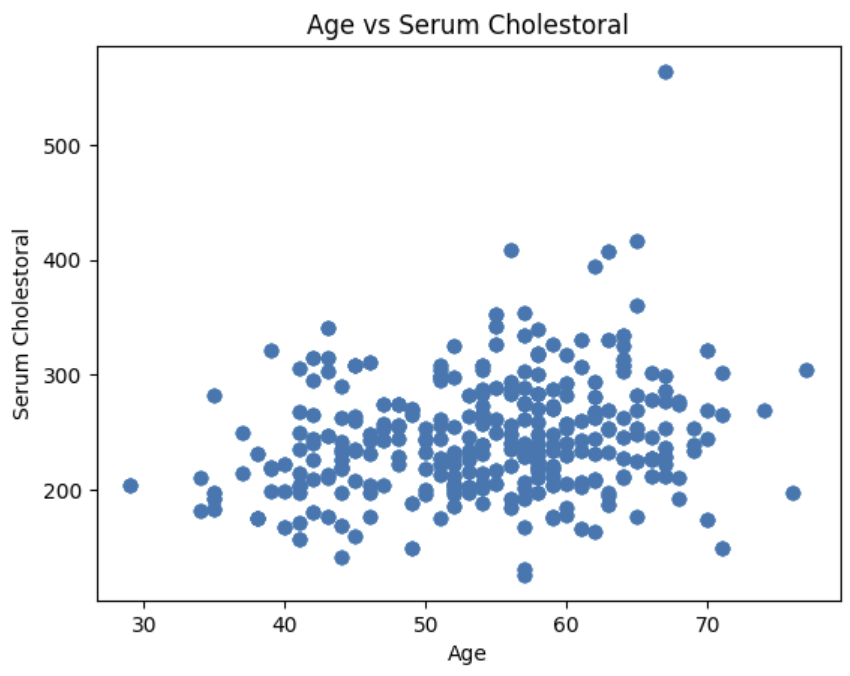
**1. Scatterplot of Age Vs Resting Blood Pressure**

The first graph, Age vs Resting Blood Pressure, suggests a slight positive relationship. In general, older patients tend to have higher resting blood pressure, although there is still a wide spread of values, and some can be viewed as outliers. For example, a few points lie well above the range of 180 mmHg, and there are also a few points below the range of 100 mmHg, which suggests unusual blood pressure compared to the bulk of the data. Despite this spread, the main pattern is a positive relationship, which makes sense, since blood pressure tends to increase with age.

****

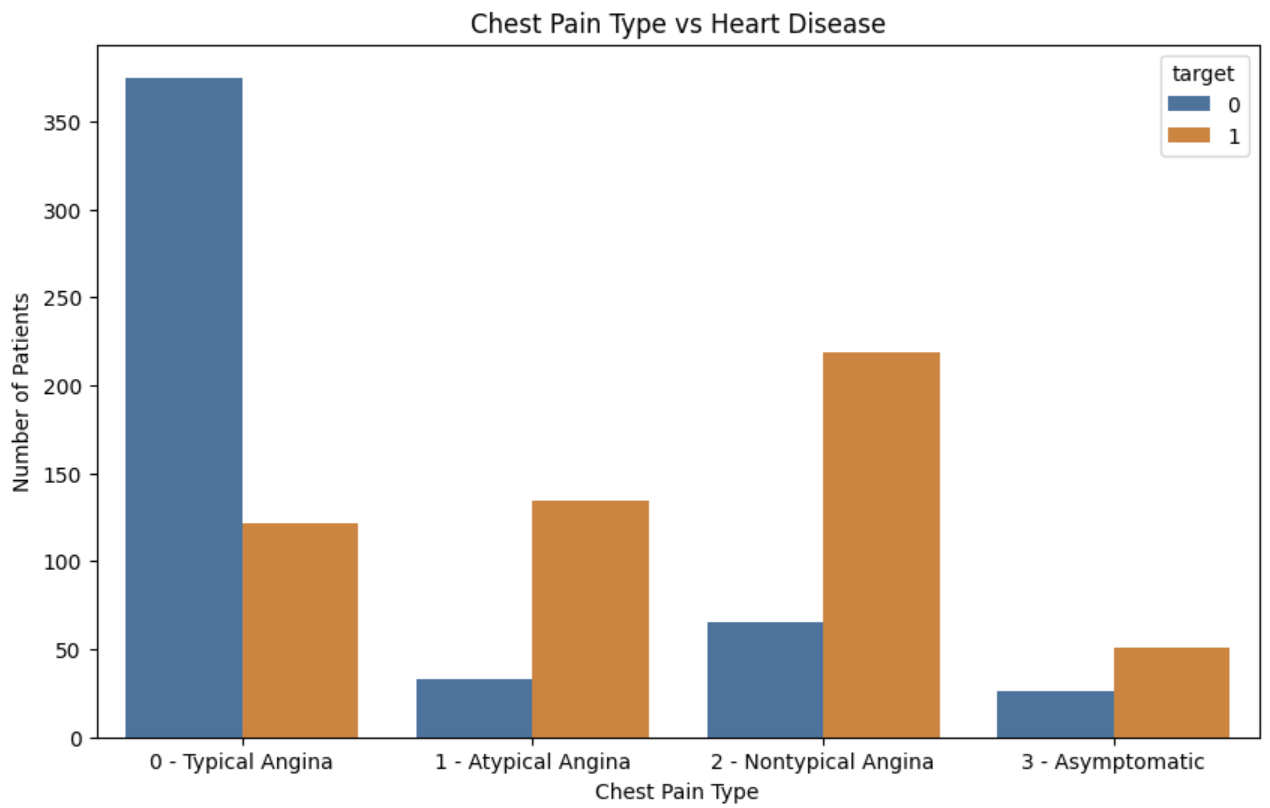
**2. Scatterplot of Age Vs Maximum Heart Rate**

The second graph, Age vs Maximum Heart Rate, suggests a clear negative relationship. From this, it was observed that older patients tend to have a lower heart rate while younger patients tend to have a higher heart rate. This makes sense, as younger individuals are typically more active and exercise more, which allows their hearts to reach higher rates during activity. The data also show some variability, where a few older patients still achieve relatively high heart rates above 170 bpm at their age of 65, while some younger patients fall below 120 bpm at their age of 40, both of which stand out as potential outliers. Overall, the main pattern supports the well-established negative relationship between age and maximum heart rate.

****

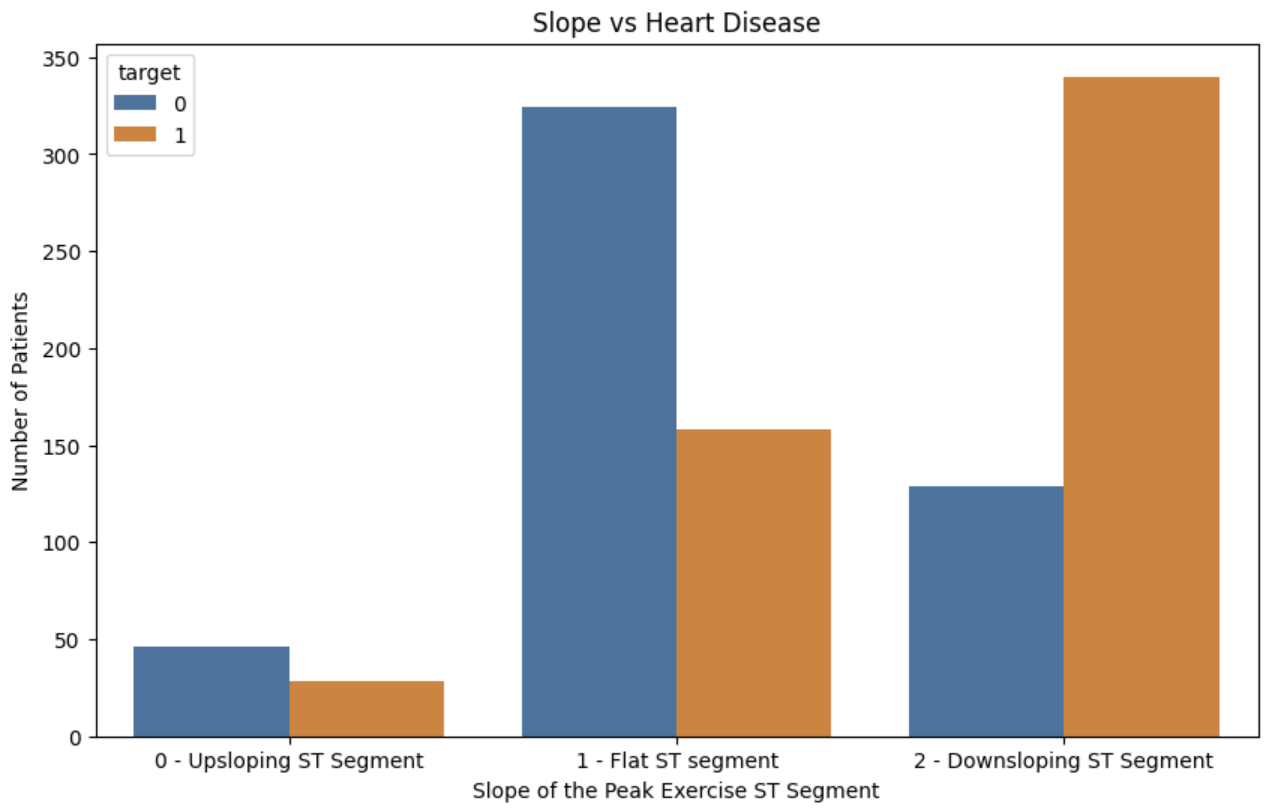
**3. Scatterplot of Age Vs Serum Cholesterol**

The third graph, Age vs. Cholesterol, shows no clear trends, patterns, or relationships. Most patients, aged 30 to 70, have cholesterol levels between 100 and 400 mg/dL, forming a large cluster. There is only one notable outlier at around 550 mg/dL at age 65. This indicates that cholesterol does not have a direct relationship with age and suggests that other factors may play a more significant role in influencing cholesterol levels.

****

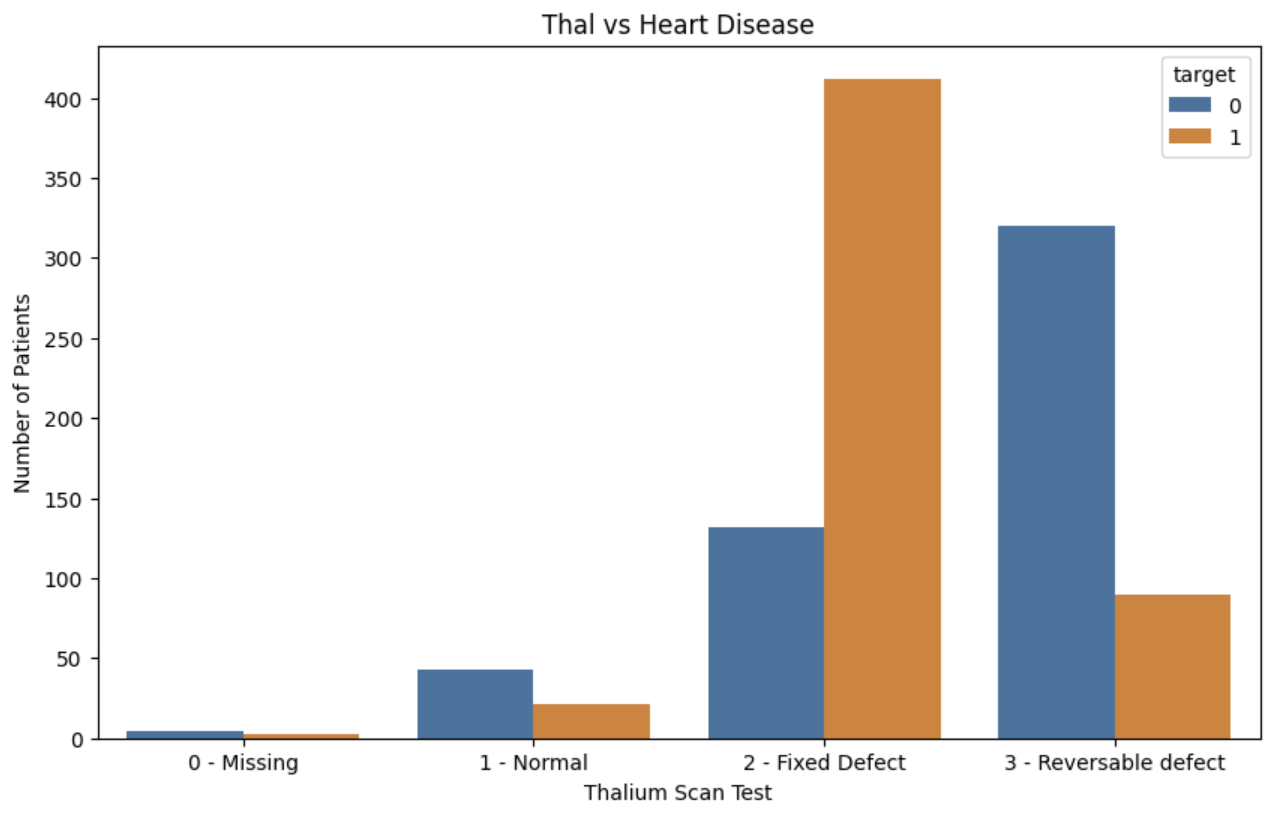
**4. Countplot of Chest Pain Type Vs. Target Label**

The fourth graph, Chest Pain Type vs Target, suggests a clear positive relationship between chest pain type and the presence of heart disease. From this, it was observed that patients with chest pain type 0 (typical angina) were more likely to have no heart disease, as the majority of patients fell under the target value 0. In contrast, chest pain type 1 (atypical angina) and chest pain type (non-typical angina) are strongly linked with heart disease, as a majority of patients fell under the target value 1. For chest pain type 3 (asymptomatic), there are fewer patients observed, but it still suggests that asymptomatic chest pain is also associated with heart disease. Overall, this indicates that chest pain type is a strong predictive feature in determining the likelihood of heart disease.

****

**5. Countplot of Slope Vs. Target Label**

The fifth graph, Slope vs Target, suggests a clear positive relationship between the slope of the peak exercise ST segment and the presence of heart disease. Patients with a slope of 0, representing a normal upsloping ST segment, have the lowest chance of heart disease. Those with a slope of 1, corresponding to a flat ST segment, show a moderate increase in heart disease. Finally, patients with a slope of 2, indicating a downsloping ST segment, have the highest incidence of heart disease, exceeding the number of patients without the condition at this slope. This countplot highlights that as the ST segment slope increases, the likelihood of heart disease also increases, suggesting that this feature is a strong indicator of cardiac risk.



**6. Countplot of Thal Vs. Target Label**

The sixth graph, Thal vs Target, suggests a clear relationship between the thalium scan test results and the presence of heart disease. Thal in this dataset refers to the results of a blood flow test that assesses whether parts of the heart muscle are receiving adequate blood supply. Patients with a thal value of 1, representing a normal heart muscle without defects, have the lowest chance of heart disease. Patients with a thal value of 2, indicating a fixed defect, show the highest incidence of heart disease due to permanent damage to the heart muscle. Those with a thal value of 3, representing a reversible defect, have a lower incidence than fixed defects, likely because their blood flow issues can improve with treatments. This countplot highlights that the type of thal defect is strongly associated with heart disease risk, where the thal value of 2 has the most influence on heart disease.