Heart Disease Exploratory Analysis

Group Members

Binaya Kumar Chaudhary C0913554

Hemant Raj Singh C0904955

Rohan Aryan C0912902

Sani Asnain C0906772

Shreya Baral C0913115

Table of Content

Introduction

Collection of health-related attributes and their association with the presence or absence of heart disease

Obtained from the UCI dataset repository

Investigate factors contributing to heart disease

https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Dataset Overview

The dataset comprised 1190 individual records. It included 11 features:

1. **age:** Numerical variable representing the age of individuals.
2. **sex**: Categorical variable indicating gender, with values 0 for female and 1 for male.
3. **chest pain type**: Categorical variable representing the type of chest pain with values 1, 2, 3, and 4.
   1 = typical angina: related to decrease blood supply in a heart
   2 = atypical angina: not related to heart,
   3 = non-anginal pain: typically, esophageal spasms (not related to heart),
   4 = not showing signs of chest pain
4. **resting bp s**: Numerical variable representing resting blood pressure.
5. **cholesterol**: Numerical variable representing cholesterol levels.
6. **fasting blood sugar**: Categorical variable indicating whether fasting blood sugar is greater than 120 mg/dl (1) or not (0).
7. **resting ecg**: Categorical variable representing resting electrocardiographic results, with values 0, 1, and 2.
8. **max heart rate**: Numerical variable representing the maximum heart rate achieved.
9. **exercise angina**: Categorical variable indicating the presence (1) or absence (0) of exercise-induced angina.
10. **oldpeak**: Numerical variable measuring the depression induced by exercise relative to rest.
11. **ST slope**: Categorical variable representing the slope of the peak exercise ST segment, with values 0, 1, 2, and 3.
12. **target**: Binary variable indicating the presence (1) or absence (0) of heart disease, which is the target variable for classification tasks.

Descriptive Summary

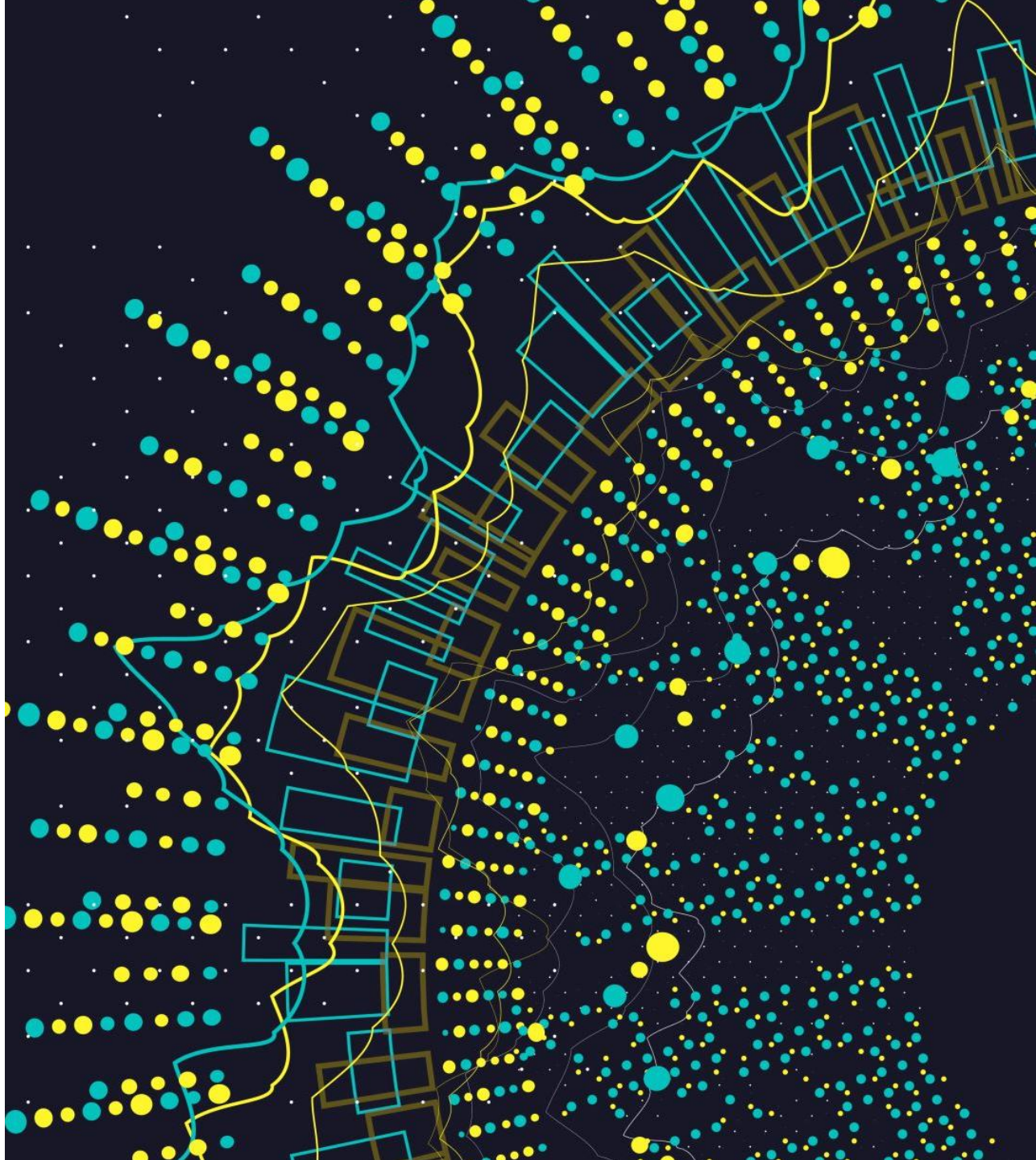
- Concise and informative overview of the count, mean, standard deviation, maximum, minimum and Interquartile range
- Offer clear and easily understandable account of the essential aspects of the dataset

	count	mean	std	min	25%	50%	75%	max
age	1190.0	53.720168	9.358203	28.0	47.0	54.0	60.00	77.0
sex	1190.0	0.763866	0.424884	0.0	1.0	1.0	1.00	1.0
chest pain type	1190.0	3.232773	0.935480	1.0	3.0	4.0	4.00	4.0
resting bp s	1190.0	132.153782	18.368823	0.0	120.0	130.0	140.00	200.0
cholesterol	1190.0	210.363866	101.420489	0.0	188.0	229.0	269.75	603.0
fasting blood sugar	1190.0	0.213445	0.409912	0.0	0.0	0.0	0.00	1.0
resting ecg	1190.0	0.698319	0.870359	0.0	0.0	0.0	2.00	2.0
max heart rate	1190.0	139.732773	25.517636	60.0	121.0	140.5	160.00	202.0
exercise angina	1190.0	0.387395	0.487360	0.0	0.0	0.0	1.00	1.0
oldpeak	1190.0	0.922773	1.086337	-2.6	0.0	0.6	1.60	6.2
ST slope	1190.0	1.624370	0.610459	0.0	1.0	2.0	2.00	3.0
target	1190.0	0.528571	0.499393	0.0	0.0	1.0	1.00	1.0

Missing Values

No Missing Values

```
[ ] #checking if there are any missing values in a dataframe
    df.isnull().sum().any()

    False
```



Handling Duplicates

```
[23] #checking if there are any duplicate values in a dataframe
     df.duplicated().sum()

     272
```

```
     # Removing duplicates
     df = df.drop_duplicates()
     df.reset_index(drop=True)
     # Resettig the index after removing duplicates
```

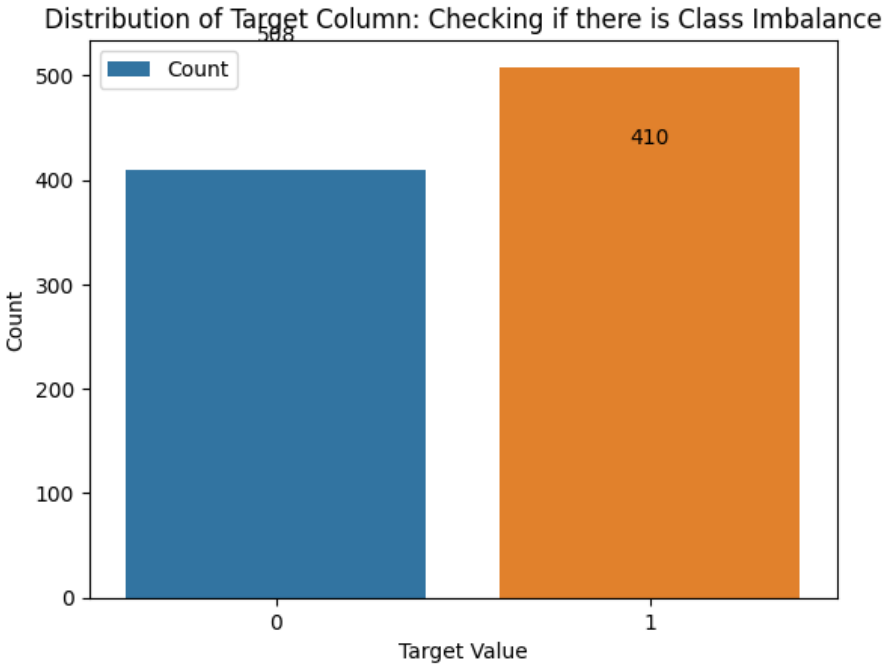
```
[27] df.duplicated().sum()

     0
```

```
[29] df.shape

     (918, 12)
```

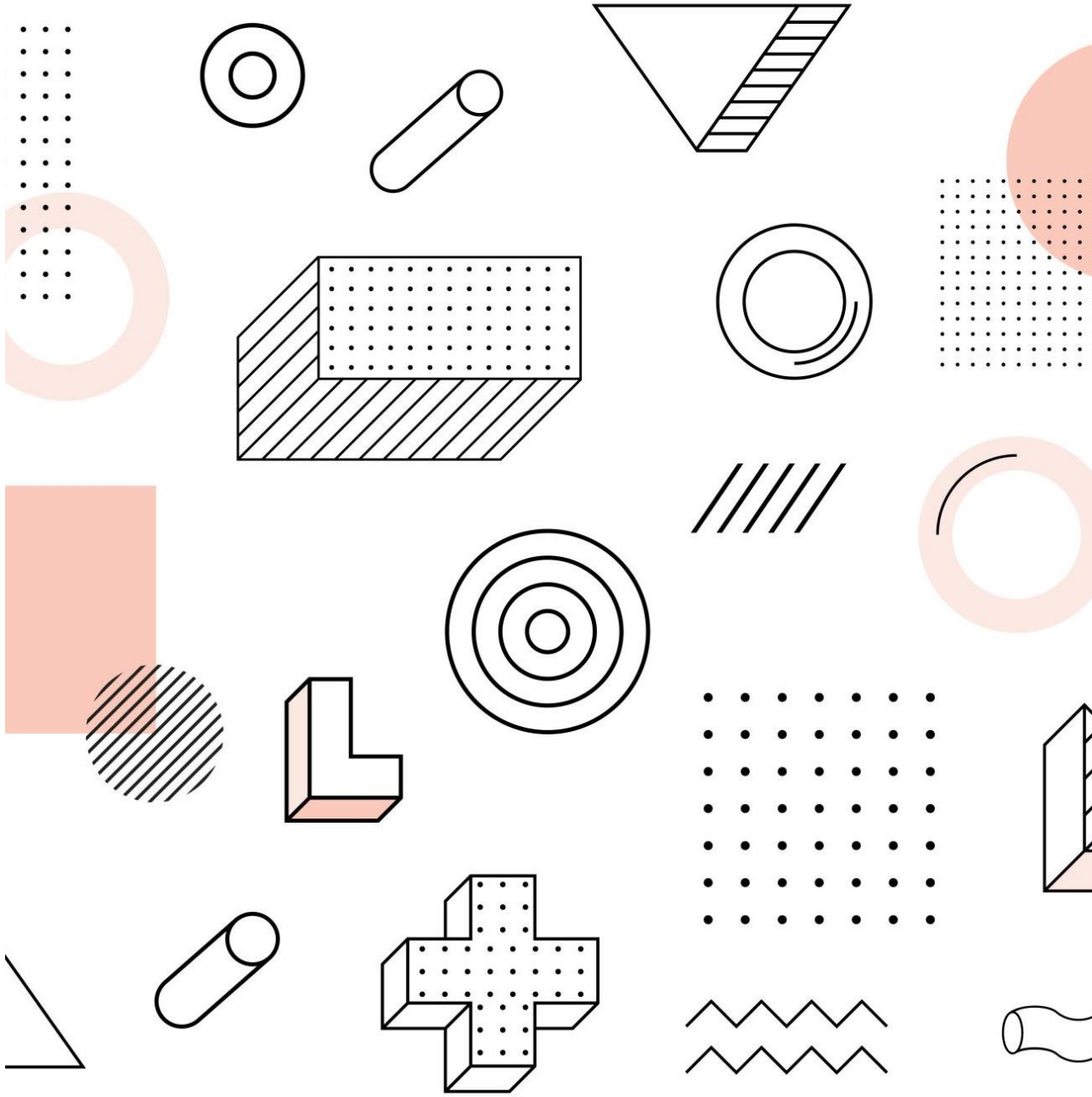
Understanding Target Variable

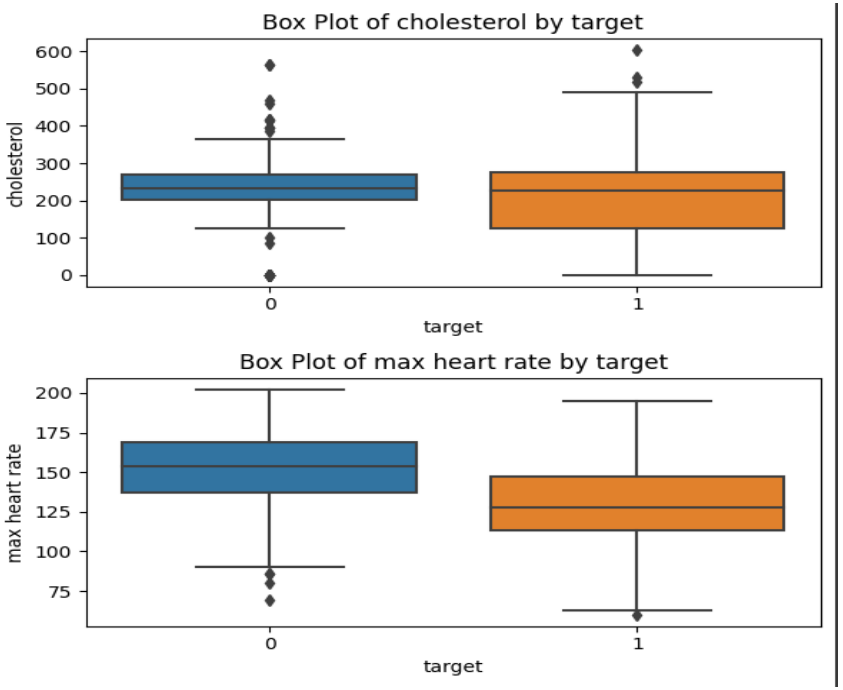
- It is a Categorical Value
- Presence (1) or absence (0) of heart disease

```
[30] #count of target value
     df['target'].value_counts()

     1    508
     0    410
     Name: target, dtype: int64
```



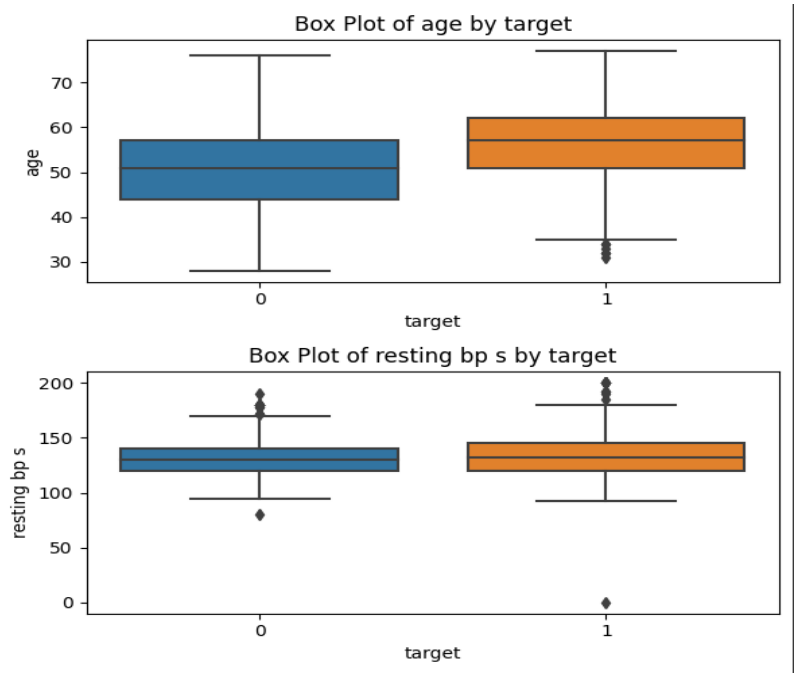
Outlier Detection

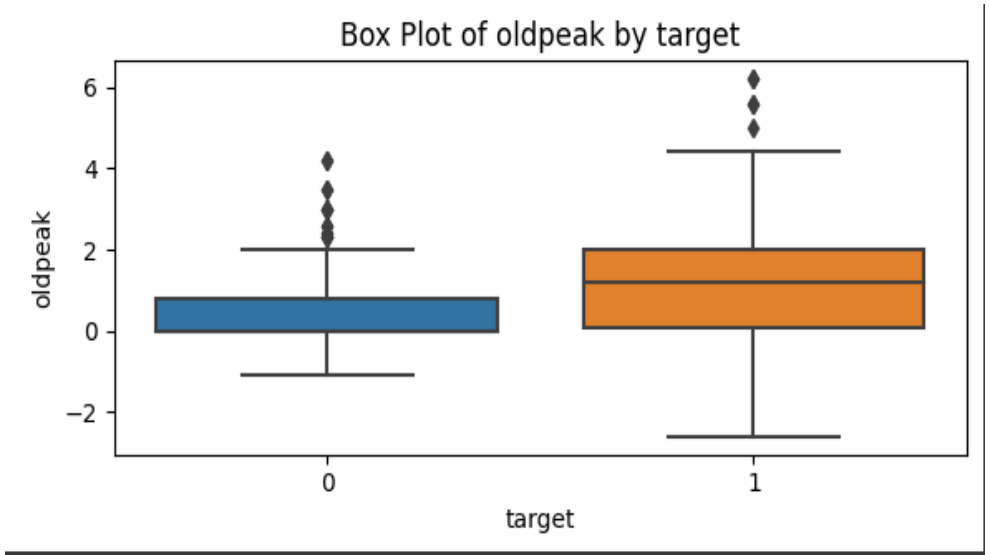




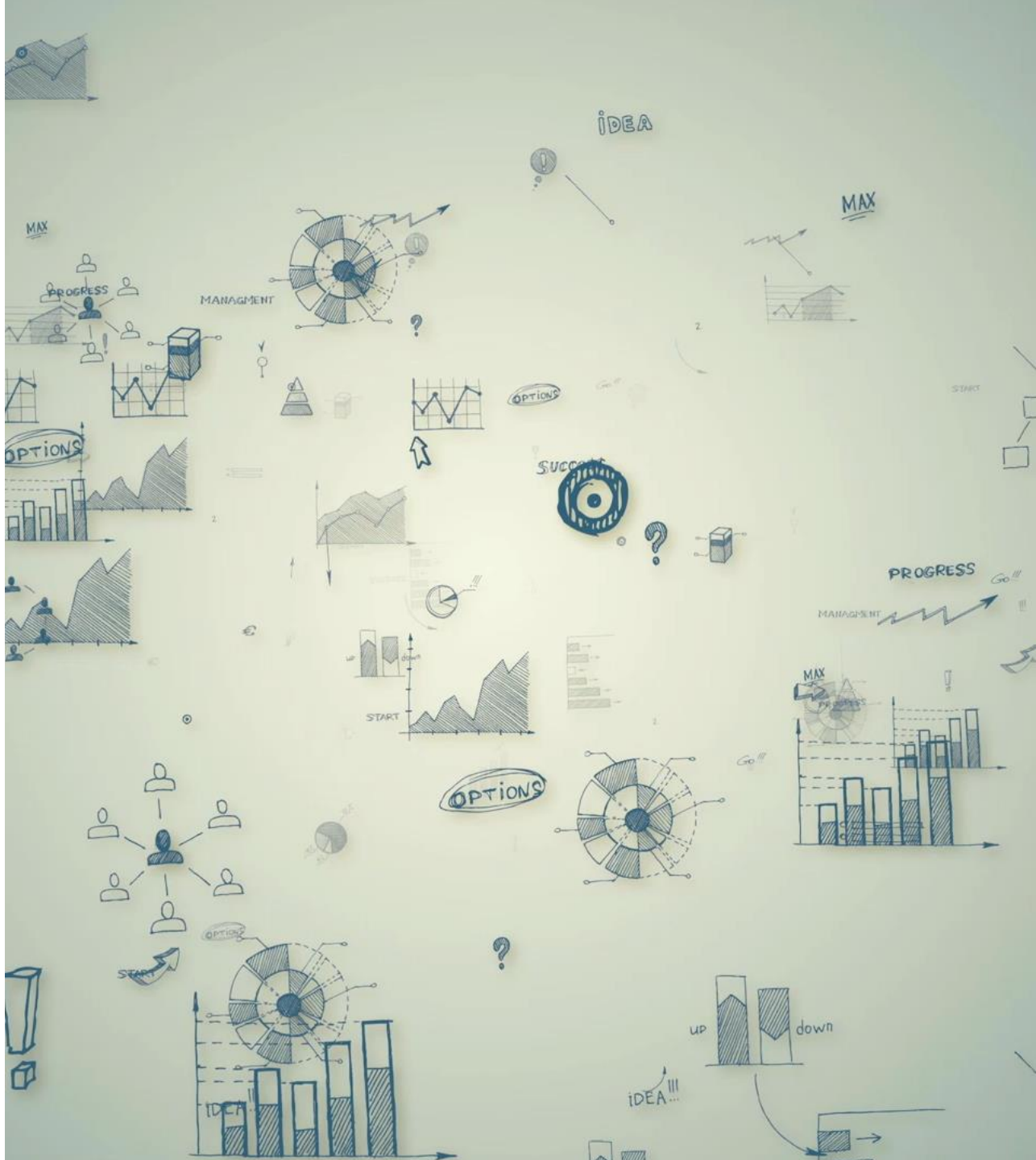
Box Plot

- Data points that are different from the general patterns of the dataset.
- Extreme high or the extreme low values.
- Influence the statistical measures like the mean, median and standard deviations of the dataset.
- Represents the Interquartile Range. The line in the middle is median, the whiskers extend to the maximum and minimum range of values and anything outside the whiskers are the potential outliers.



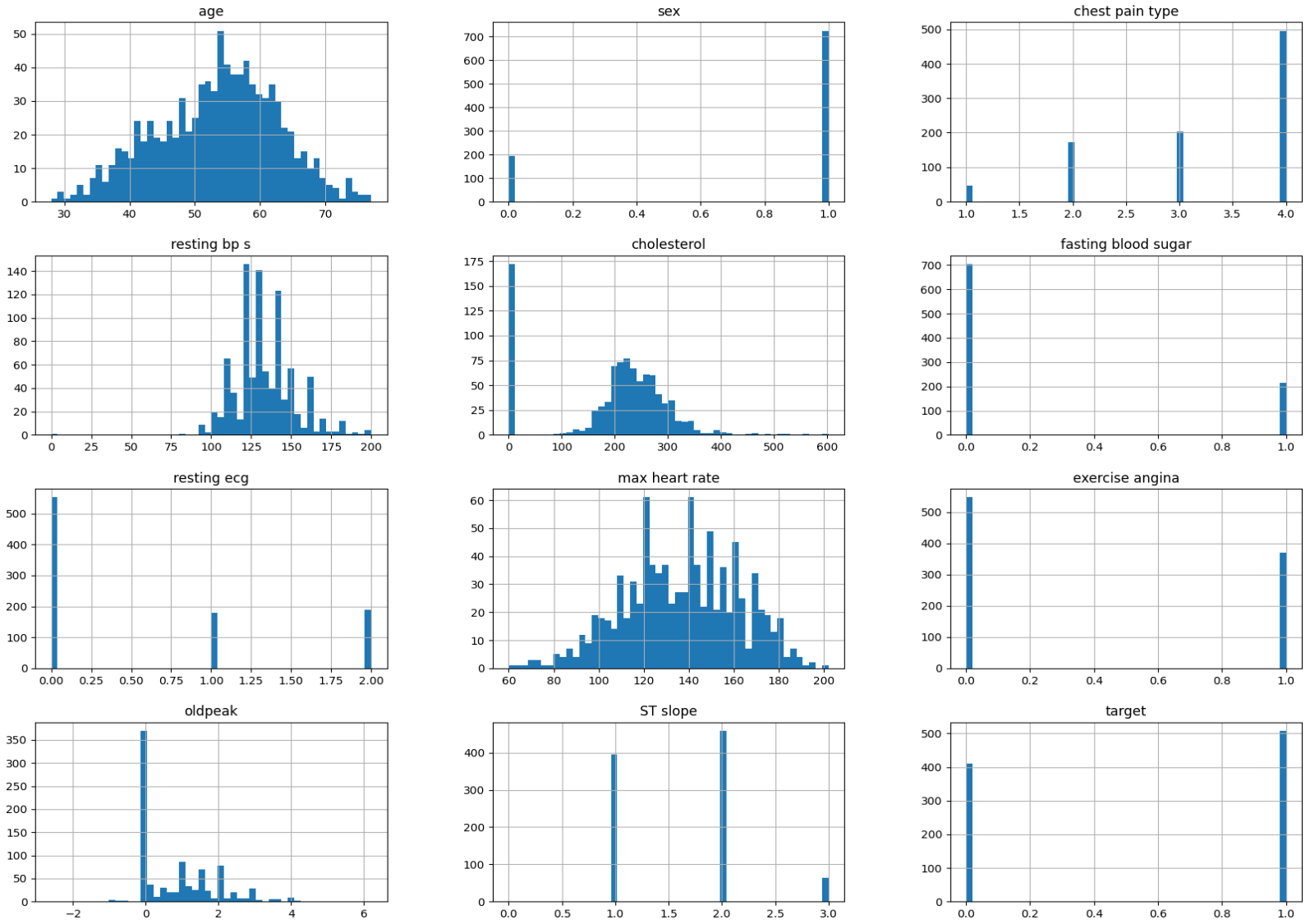


Data Visualizations



Histogram

- Histogram to understand if variables are categorical or continuous.
- Display of distribution of our categorical and continuous variables.
- Continuous variables: age, resting bp s, cholesterol, max heart rate, oldpeak
- Categorical variables = sex, chest pain type, fasting blood sugar, resting ecg, exercise angina, ST slope and target



Distribution of Categorical Variables

Sex: distribution, indicating a predominance of male participants in the dataset.

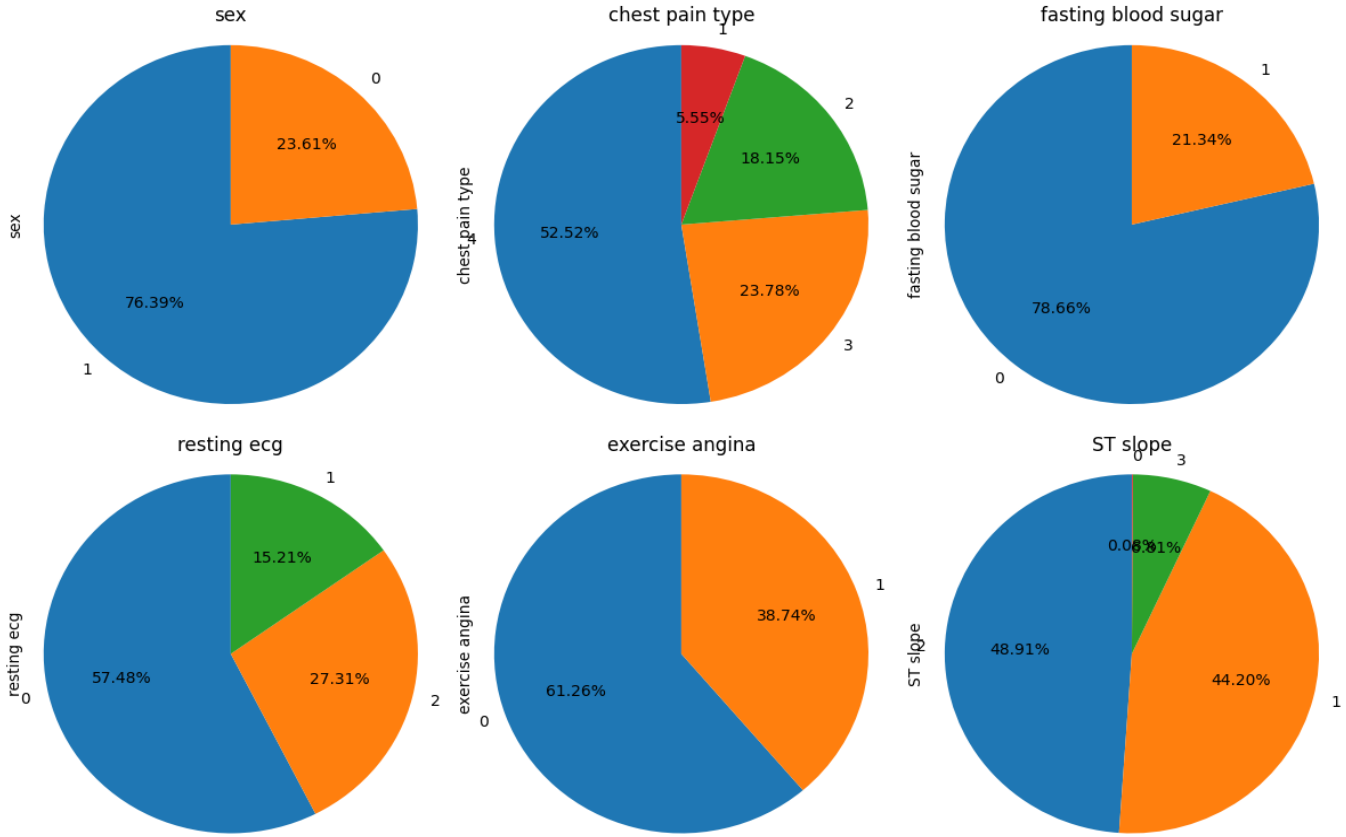
Chest Pain Type: Only 5.55% of the participants has type 1 chest pain which is chest pain related to heart disease

Fasting Blood Sugar: 76.66 % of 0 shows majority of people did not have fasting blood sugar

Resting ecg: 27.31% of people shows enlarged heart's main pumping chamber, 15.21% shows mild symptoms/ signals of non-normal heart beat and 57.59 % of population has the normal heart beat.

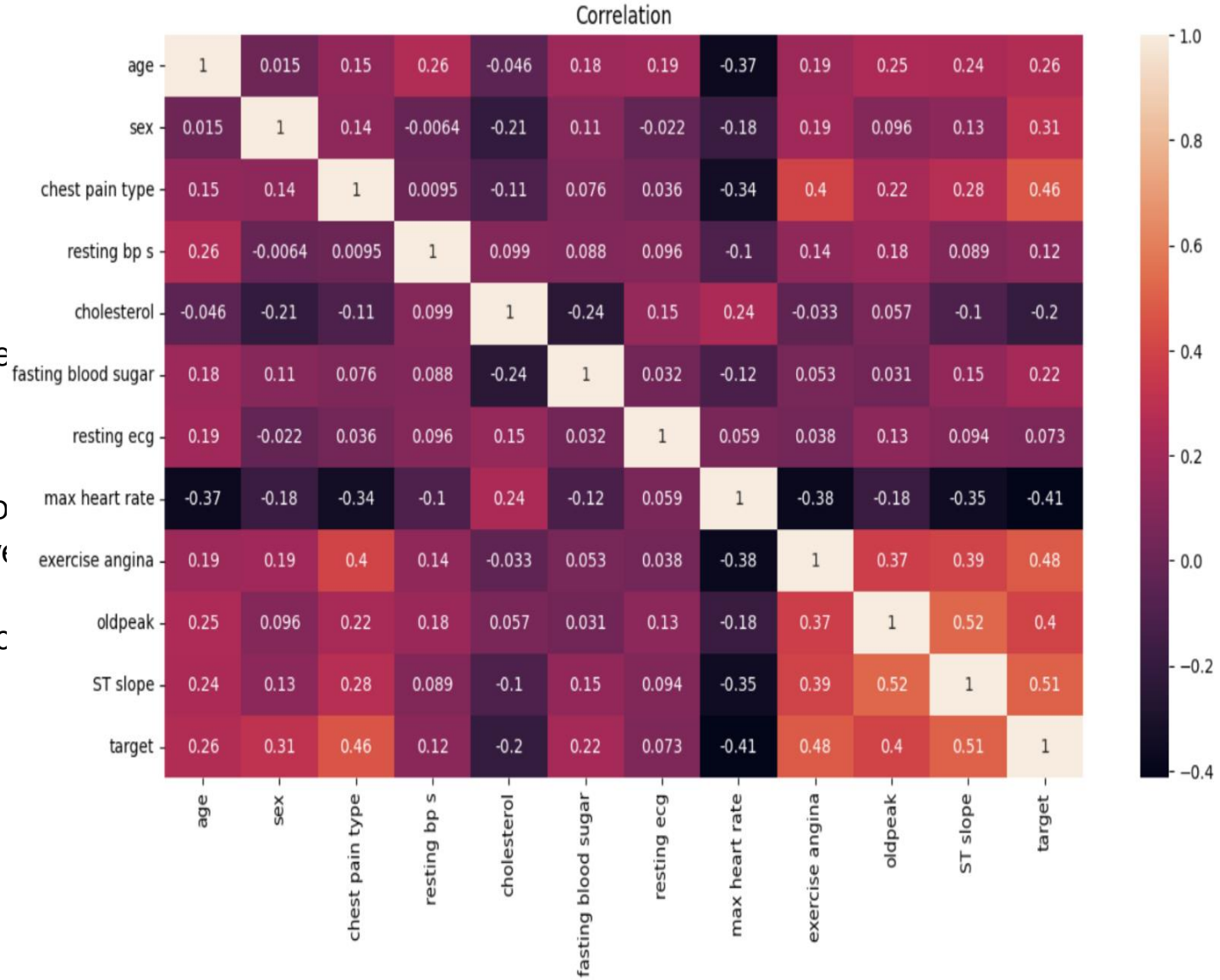
Exercise angina: 61.26% did not experience any exercise induced angina

ST slope: Approximately 92% exhibit an upsloping or flat slope of the peak exercise ST segment, yet certain errors lead to the emergence of a new classification group as 0 hence it doesn't have any data.



Correlation

- Describes the relationship between the two variables and the strength and direction of those relationships.
- Ranges from –1 to 1. 1 indicates the positive correlation i.e. as one value increases other also increases. The negative values show the negative correlation i.e. as one value increases the other decreases. Similarly, 0 shows no linear correlatic between the variables.



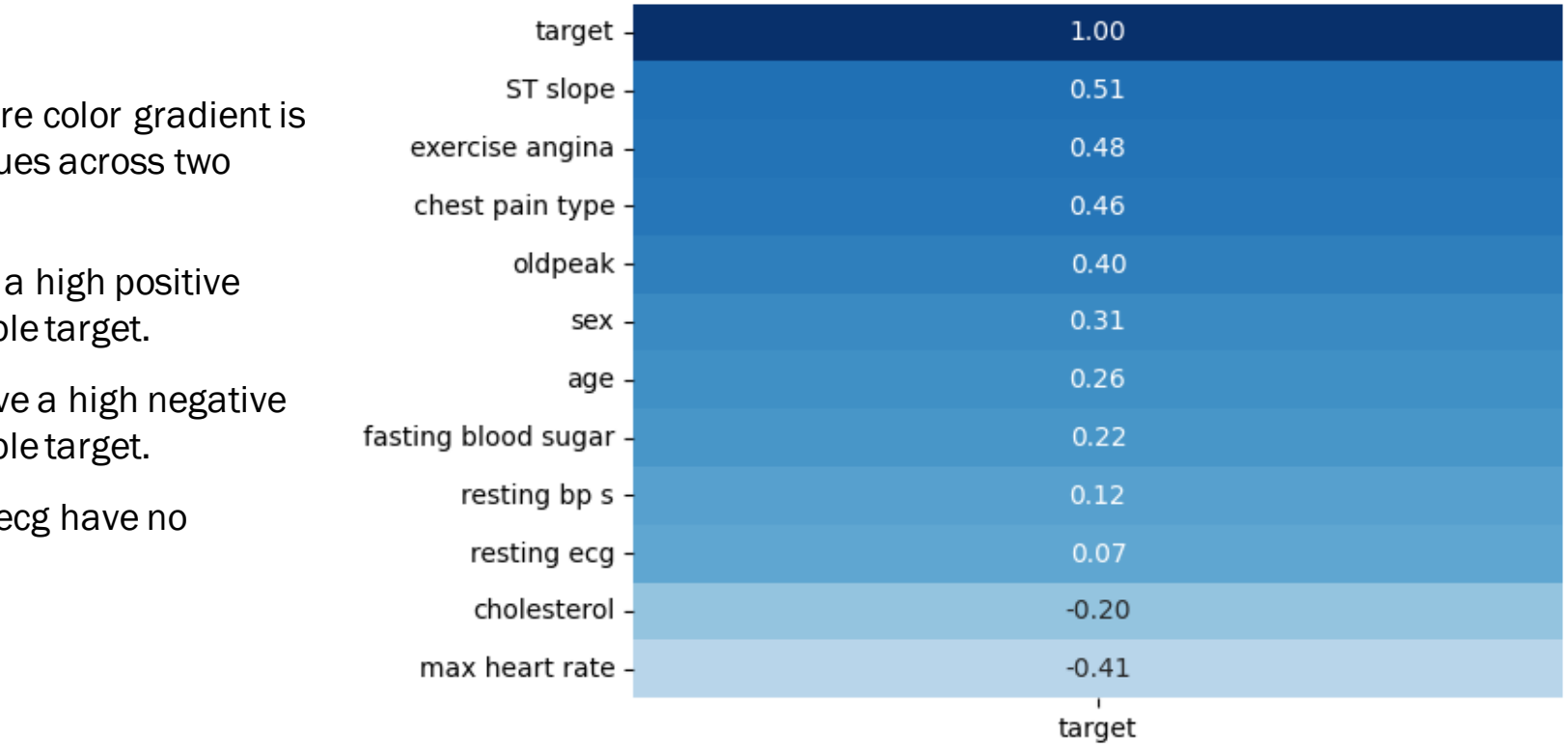
Heat Map

Graphical representation of data where color gradient is used to represent the intensity of values across two dimensions.

1. ST slope and exercise angina have a high positive correlation with the dependent variable target.

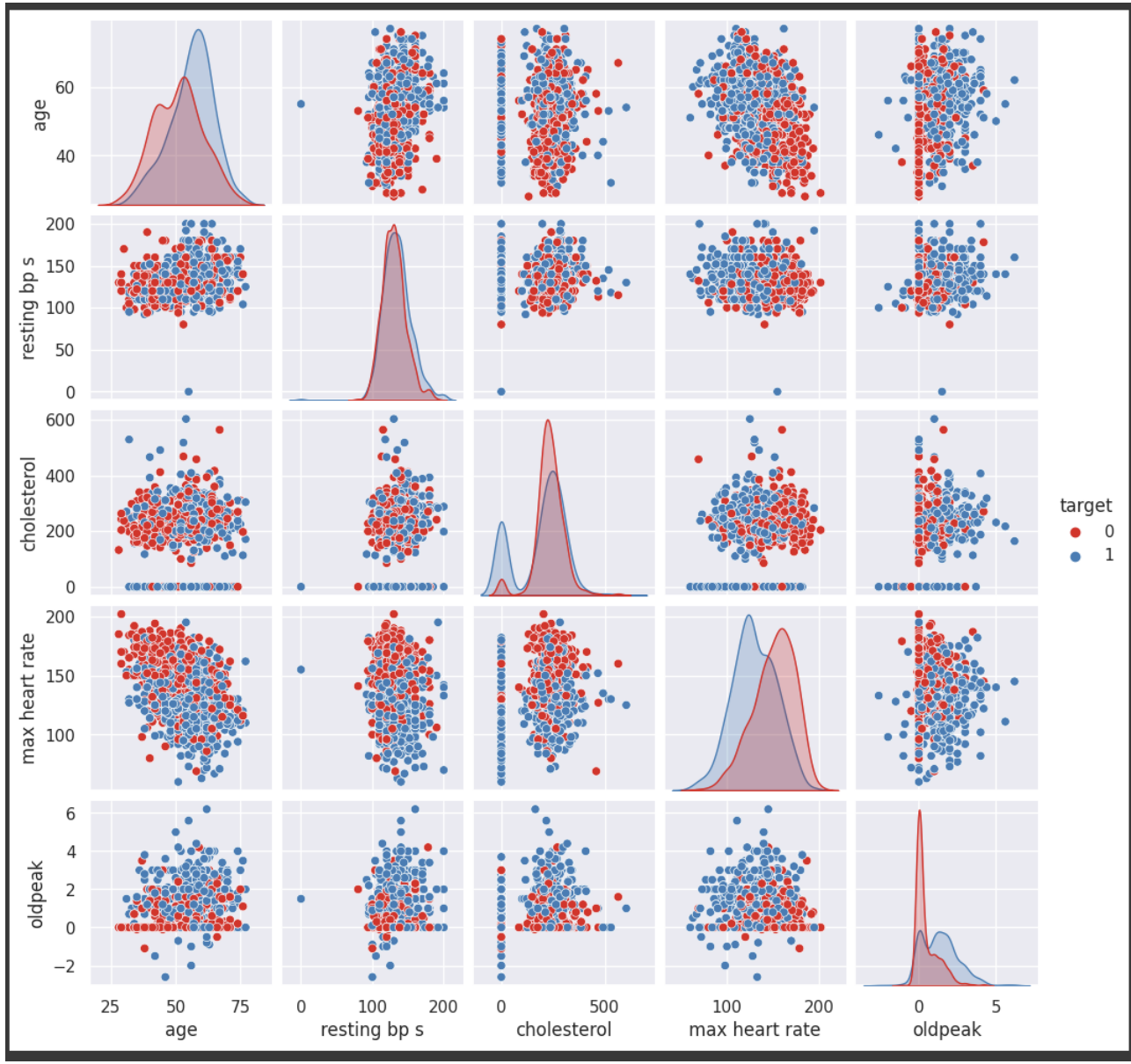
2. Cholesterol and max heart rate have a high negative correlation with the dependent variable target.

3. Variables like resting bp s, resting ecg have no correlation at all with price.



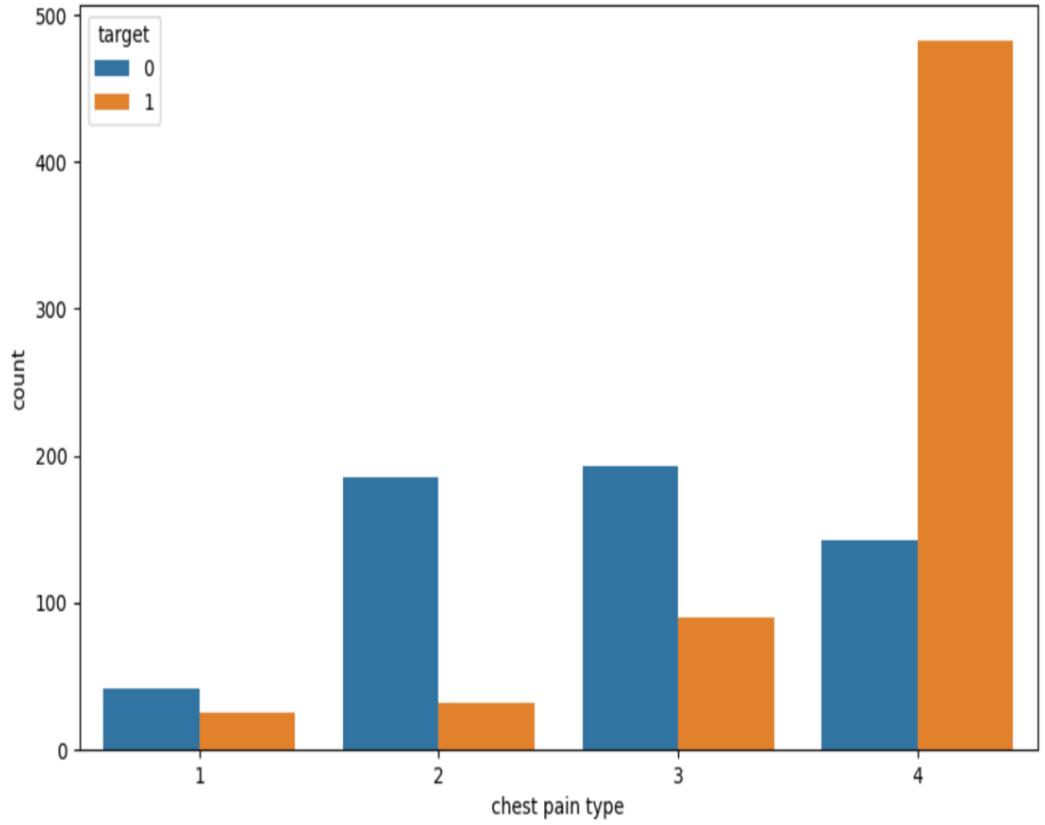
Pair Plot

- Function provided by the seaborn library to create matrix of scatterplot for multiple variables.
- Supports both continuous and categorical variables.
- Each scatter plot represents the correlation between two variables, with blue and red dots indicating the absence (0) and presence (1) of heart disease, respectively.
- The diagonal histograms show the distribution of each variable for both categories.

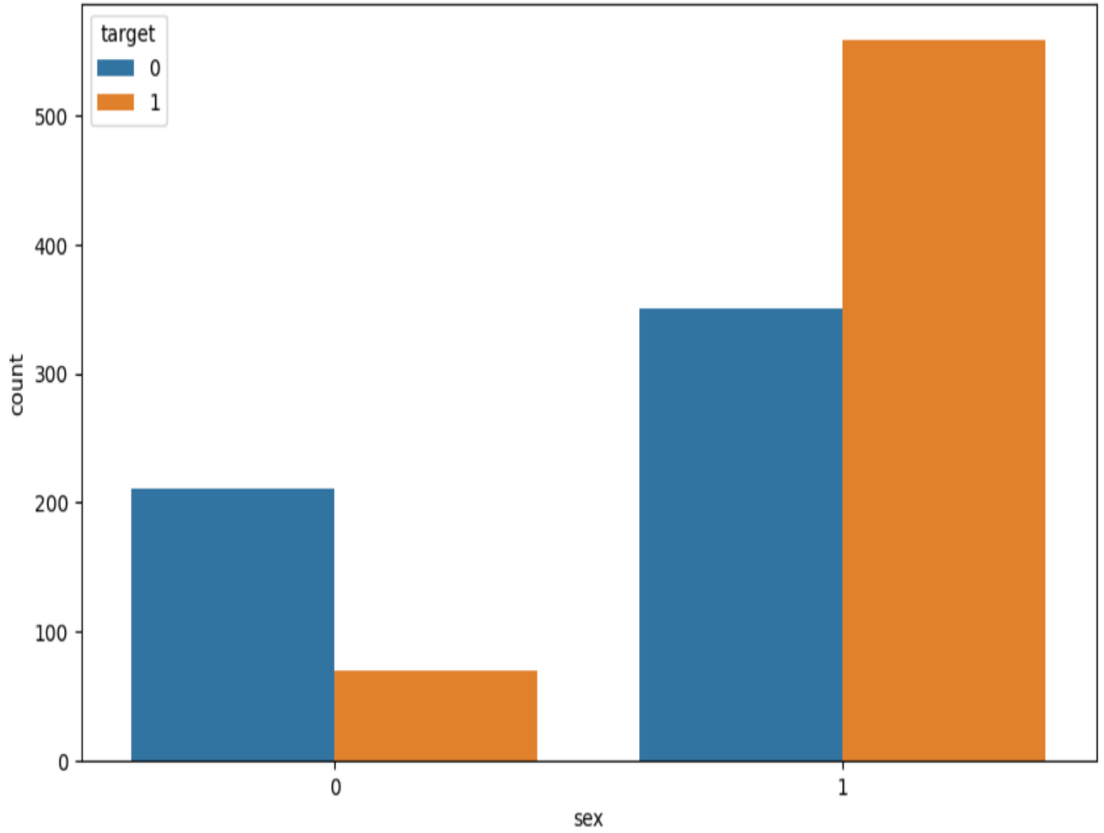


Visualizing the frequency distribution

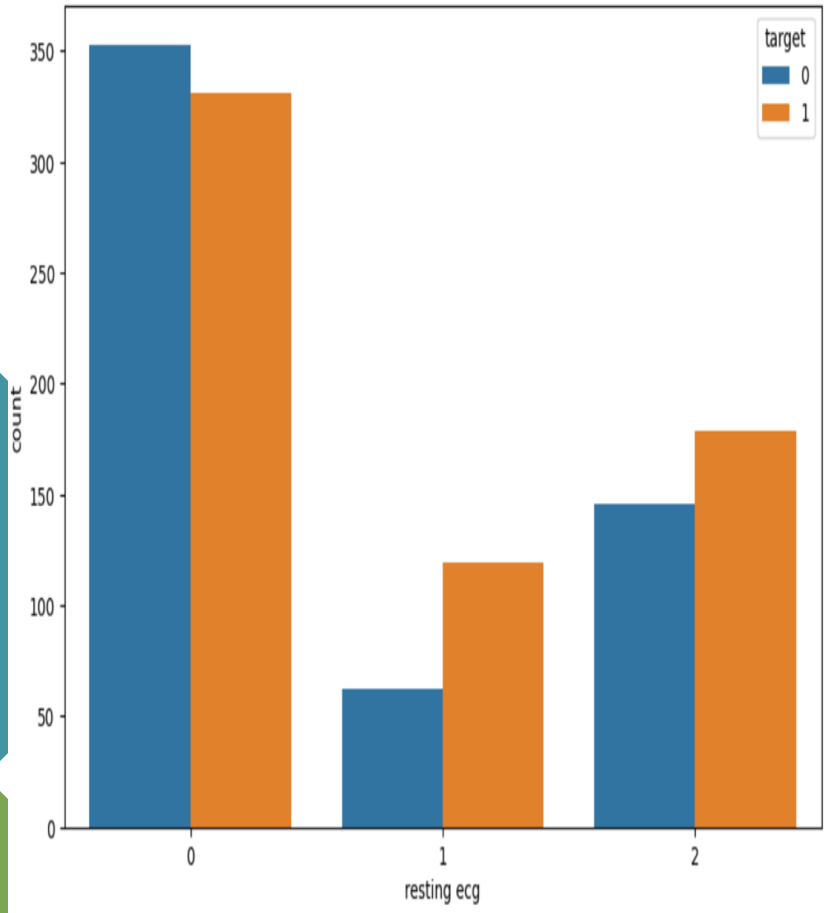
This analysis shows amongst the people who have heart disease ,majority of them do not show any signs of chest pain.



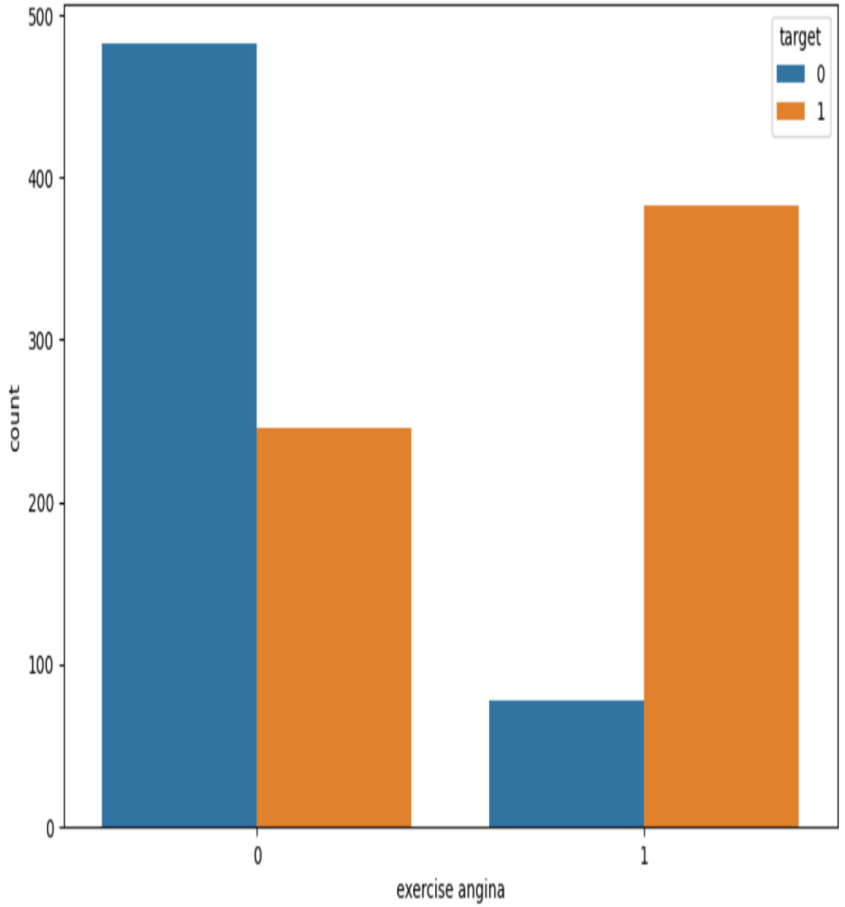
From this chart, we can see that the chances of heart disease are much higher in the male population (1) compared to female (0).



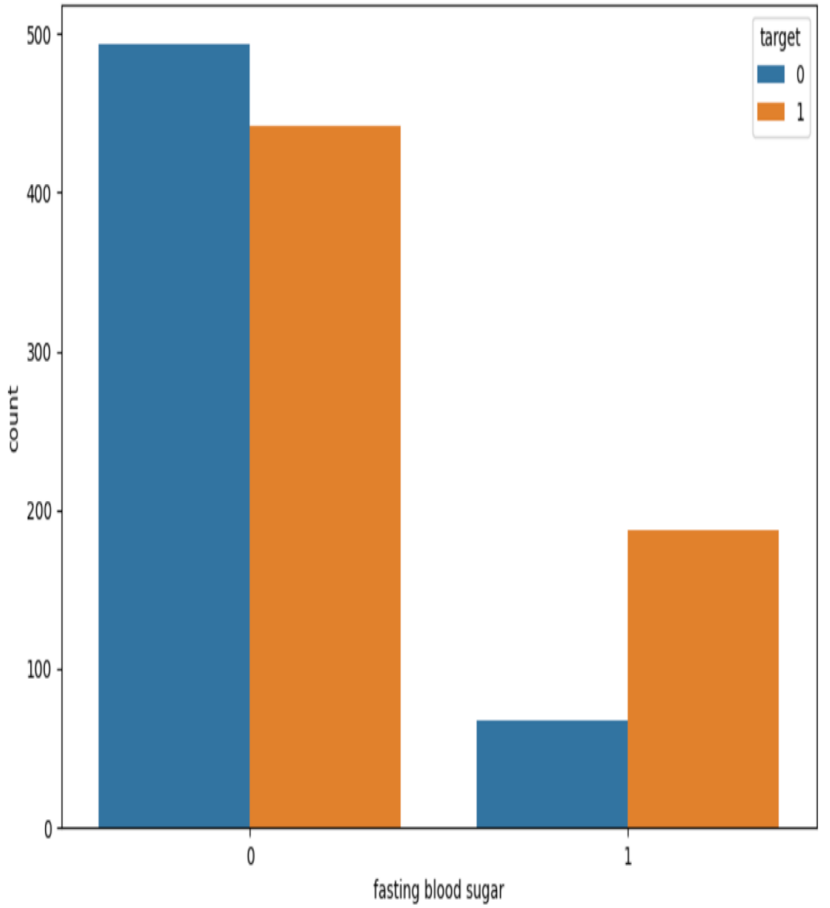
Patients who have higher resting ecg have more risk to heart disease.



Majority of participants who experience exercise induced angina have a higher chance of getting heart disease.

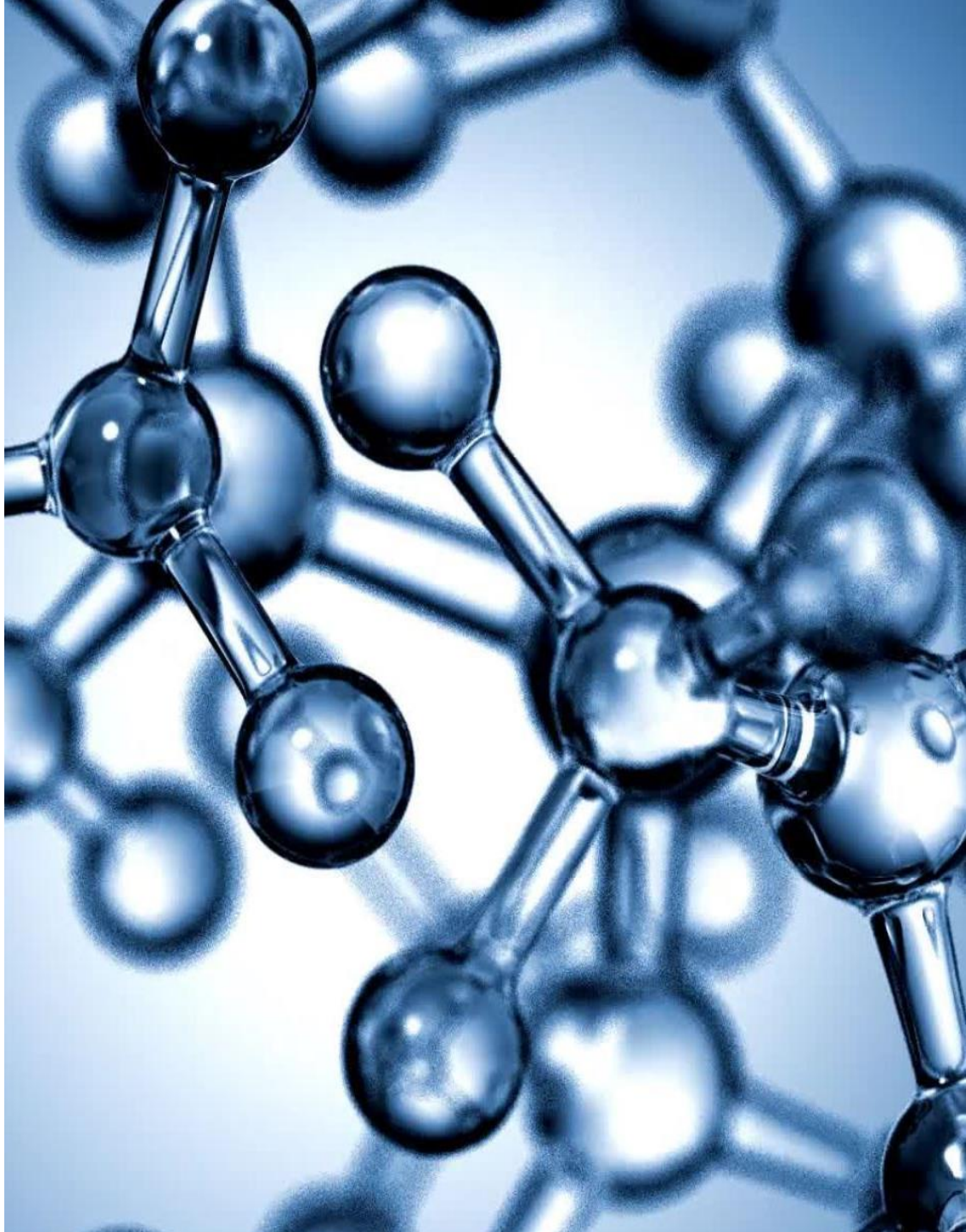


High number of patients having normal fasting sugar level will have a moderate chance of having heart disease. But the small number of patients who have an above normal cholesterol will have a higher chance of having heart disease.



Results

- The analysis confirmed several factors as significant in the prediction of heart disease, such as age, cholesterol levels, and resting blood pressure.
- The findings underscore the importance of these variables in the development of predictive models for heart disease.
- ST slope and exercise angina have a high positive correlation with the dependent variable target.
- Cholesterol and max heart rate have a high negative correlation with the dependent variable target.
- Variables like resting bp s, resting ecg have no correlation at all with price.



Further Analysis

- Refining data cleaning processes
- Applying machine learning models
- Validating the results against an independent dataset to confirm the predictive power and reliability of the identified risk factors.
- Implementation of Classification since the target is categorical to predict the presence or absence of heart disease.

Thank you

Any Questions?