Part I:

Exploratory Data Analysis on Heart Disease

Agenda

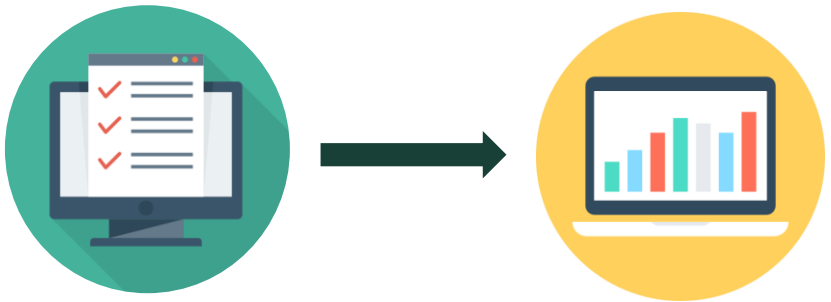
01 Background & dataset

02 Data Preprocessing

03 Univariate Analysis

04 Bivariate Analysis & Multivariate Analysis

05 Conclusion



Project Objective

- In Hong Kong, heart diseases rank ***fourth*** in mortality, with an average of 10.9 deaths per day in 2022, primarily affecting males.

- Globally, heart diseases (medically known as cardiovascular diseases), including conditions affecting the **heart** and **blood vessels**, are the ***leading cause*** of death according to the WHO, with 17.9 million deaths in 2019.

We aim to

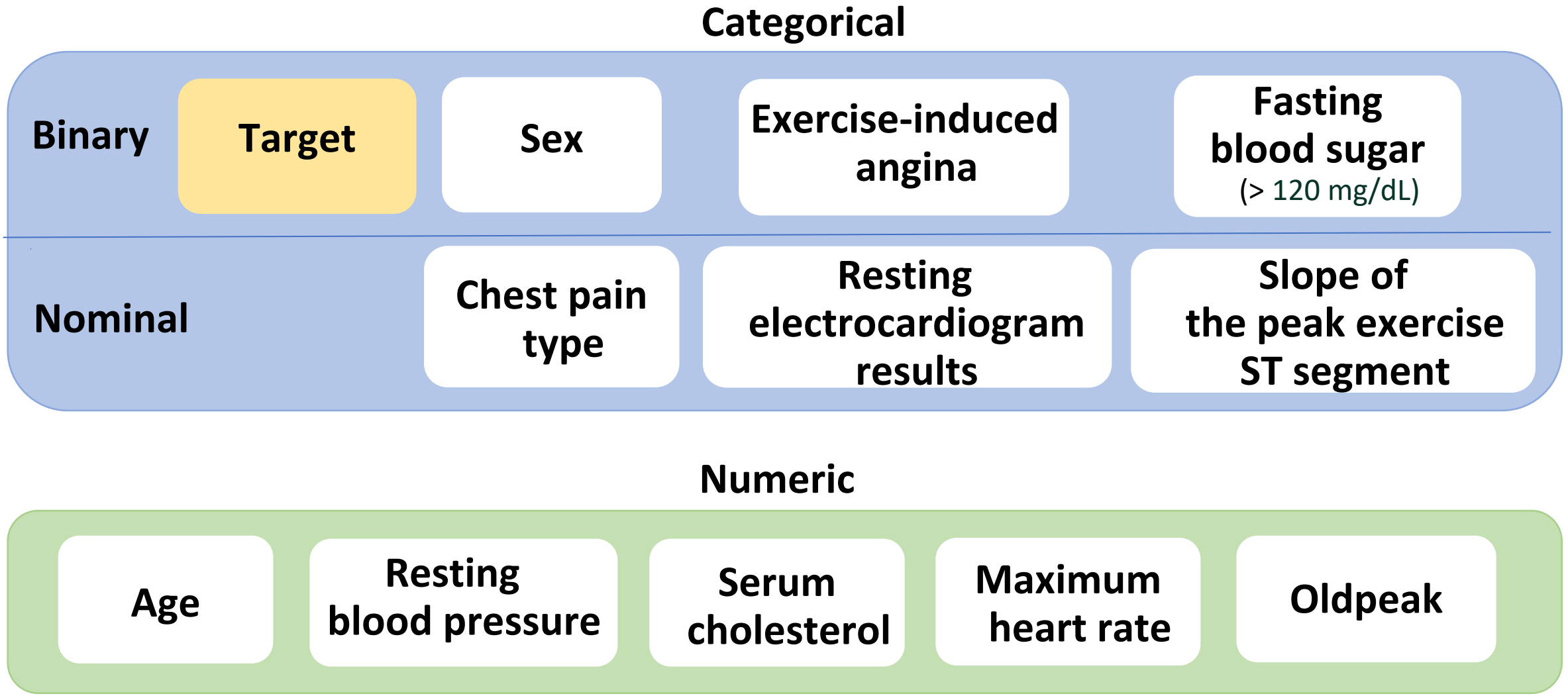
- investigate the association between different variables and the presence of heart disease.

Introduction of Dataset

Source of dataset:

- The dataset utilized in this study consists of five distinct datasets, namely the Cleveland, Hungarian, Switzerland, Long Beach VA, and Statlog datasets.
- These datasets are widely recognized and extensively employed in the medical field for heart disease research.

Original Dataset



Dataset

Categorical

Binary

Target

0: Normal
1: Heart Disease

Exercise-induced angina

0: No
1:
Yes

Sex

0: Female
1: Male

Fasting blood sugar
(> 120 mg/dL)

0: False
1: True

Nominal

Chest pain type

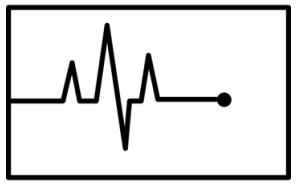
1: Typical angina
2: Atypical angina
3: Non-anginal pain
4: Asymptomatic

Resting electrocardiogram results

0: Normal
1: Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
2: Showing probable or definite left ventricular hypertrophy by Estes' criteria

Slope of the peak exercise ST segment

1: Upsloping
2: Flat
3: Downsloping



Dataset

Numeric

	Age	Resting blood pressure	Serum cholesterol	Maximum heart rate	Oldpeak
Range in dataset	28 ~ 77	92 ~ 200 mg Hg	85 ~ 603 in mg/dL	69 ~ 202	-0.1 ~ 6.2
Normal Range		< 120 mg Hg	< 200 mg/dL	(220 - Age) bpm ±15 bpm	0 ~ 1.5

Data Cleansing

- Remove the unnecessary column 'Unnamed: 0' (#1)
- Check null value (#2)

#1

	Unnamed: 0	age	sex	chest pain type	resting bps
0	0	40	1	2	140
1	1	49	0	3	160
2	2	37	1	2	130
3	3	48	0	4	138
4	4	54	1	3	150

#2

```
# Determine whether there are missing values
df_1.isnull().sum()
```

```
age                   0
sex                   0
chest pain type       0
resting bps           0
cholesterol           0
fasting blood sugar   0
resting ecg           0
max heart rate        0
exercise angina       0
oldpeak               0
ST slope              0
target                0
dtype: int64
```

Data Cleansing

- Check duplicated value (#3)
- Check Dataset size (number of rows, number of columns) (#4)

#3

```
# Determine whether there are duplicated records
df_1.duplicated().sum()

# Drop duplicate records if any
# df_1 = df_1.drop_duplicates(subset=None, keep='first', inplace=True)
```

```
0
```

#4

```
# Explore the data
df_1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048 entries, 0 to 1047
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   age                  1048 non-null   int64  
 1   sex                  1048 non-null   int64  
 2   chest pain type      1048 non-null   int64  
 3   resting bps          1048 non-null   int64  
 4   cholesterol          1048 non-null   float64
 5   fasting blood sugar  1048 non-null   int64  
 6   resting ecg          1048 non-null   int64  
 7   max heart rate       1048 non-null   int64  
 8   exercise angina      1048 non-null   int64  
 9   oldpeak              1048 non-null   float64
 10  ST slope             1048 non-null   int64  
 11  target               1048 non-null   int64  
dtypes: float64(2), int64(10)
memory usage: 98.4 KB
```

Data Cleansing - create new columns

- Add two columns (#5)

1. age group
   1: 21 – 30 years old
   2: 31 – 40 years old
   3: 41 – 50 years old
   4: 51 – 60 years old
   5: 61 – 70 years old
   6: 71 – 80 years old

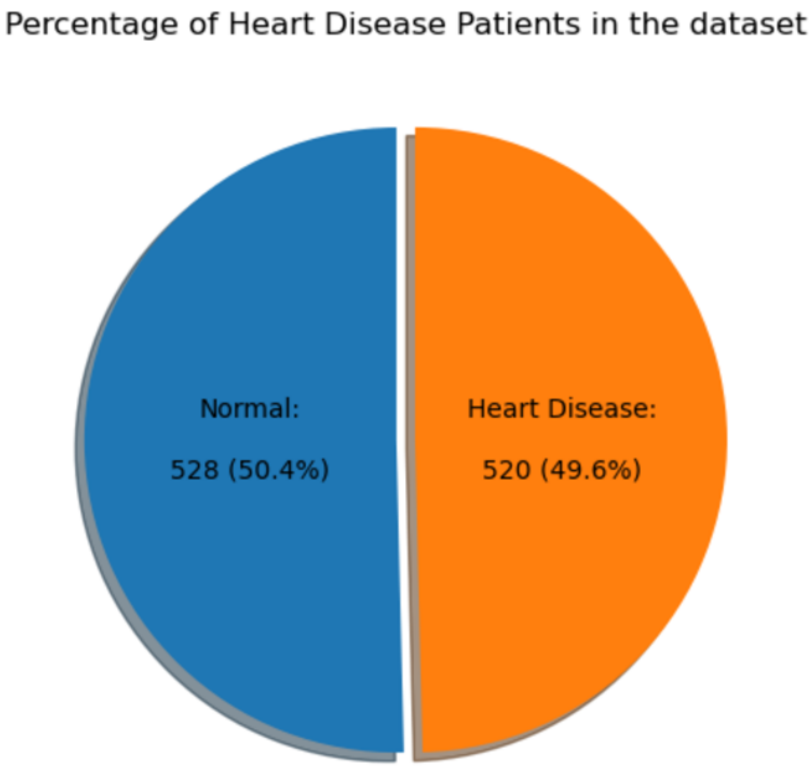
2. max heart rate status
   - Normal: within (220 - Age) $\pm$ 15 bpm
   - Abnormal: outside this reference range

Dataset characteristics

1. Mean and median (50%) of all columns, except oldpeak, are almost the same, meaning that the data is symmetrically distributed.

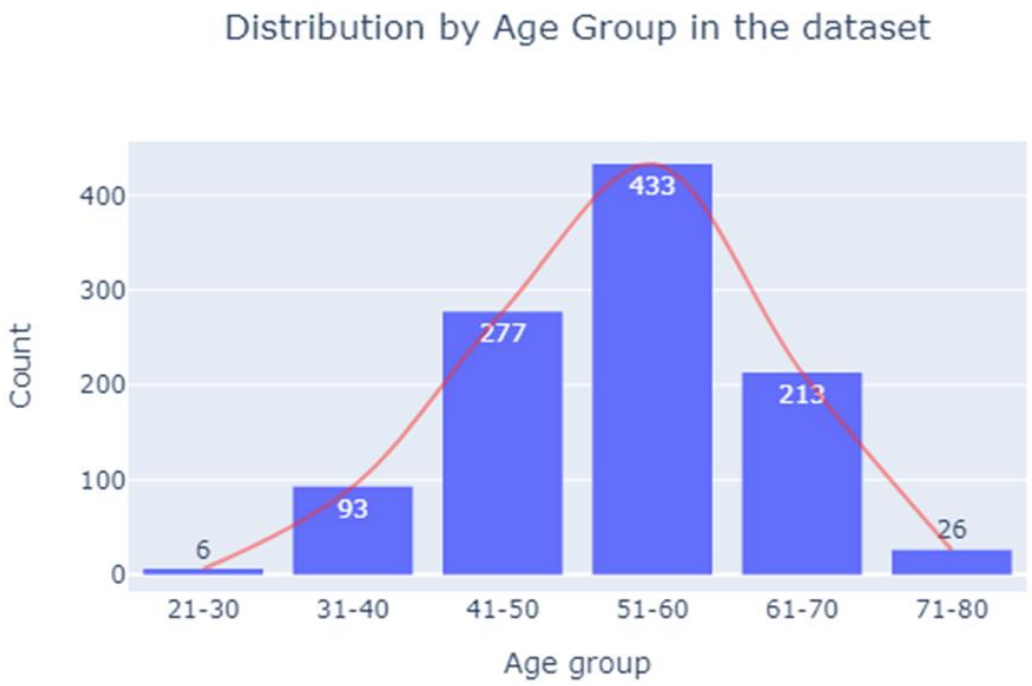
2. The number of normal patients and patients with heart disease are nearly the same, indicating that this is a balanced dataset.

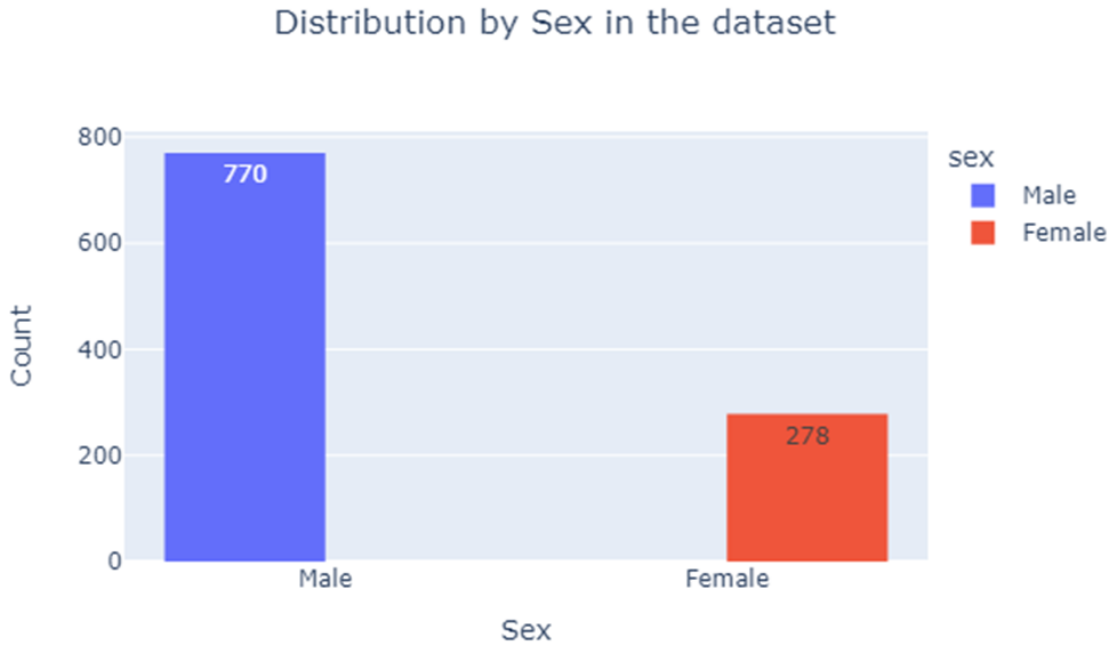
	age	resting bps	cholesterol	max heart rate	oldpeak
count	1048.000000	1048.000000	1048.000000	1048.000000	1048.000000
mean	53.325382	132.613550	245.172710	142.918893	0.942366
std	9.397822	17.367605	57.101359	24.427115	1.100429
min	28.000000	92.000000	85.000000	69.000000	-0.100000
25%	46.000000	120.000000	208.000000	125.000000	0.000000
50%	54.000000	130.000000	239.000000	144.000000	0.600000
75%	60.000000	140.000000	275.000000	162.000000	1.600000
max	77.000000	200.000000	603.000000	202.000000	6.200000



Dataset characteristics

3. The distribution of age follows a normal distribution pattern.





Methodology

- **Investigate association between independent variable(s) and dependent variable (Target, i.e. having heart disease or not)**

	Univariate Analysis	Bivariate Analysis	Multivariate Analysis
No. of independent variable(s)	1	2	3 or above
For Categorical independent variable(s)	**1. Histogram** - distribution by category **2. Heart Disease Prevalence Ratio** **3. Chi-squared test (p-values)** - statistical significance of the association **4.** [if p-value < 0.05] **Cramer's V** - strength & direction of association	**1. Heatmap** - corr(method='kendall') **2. Violin Plot** - 2-dimension comparison **3. Stacked bar chart**	**Binary Logistic Regression** **1. p-value (Scatter Plot)** - statistical significance of the association **2. Coefficient (Bar chart)** - strength & direction of association
For Numeric independent variable(s)	**1. Boxplot** **2. Point-Biserial Correlation Coefficient** - strength & direction of association		

Univariate Analysis

A. Target VS Categorical Variable:

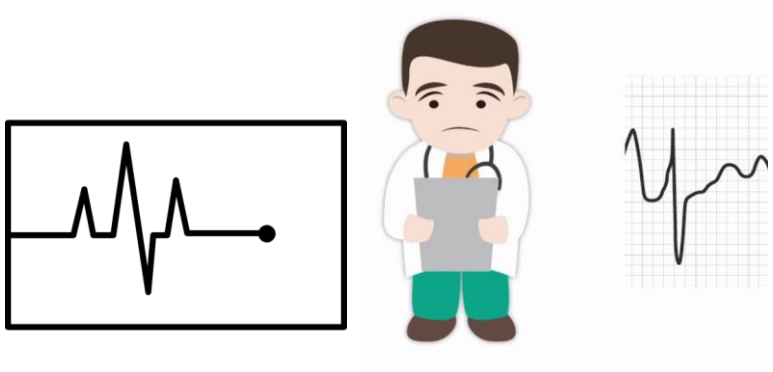
Degrees of freedom	Small	Medium	Large
1	0.10	0.30	0.50
2	0.07	0.21	0.35
3	0.06	0.17	0.29
4	0.05	0.15	0.25
5	0.04	0.13	0.22

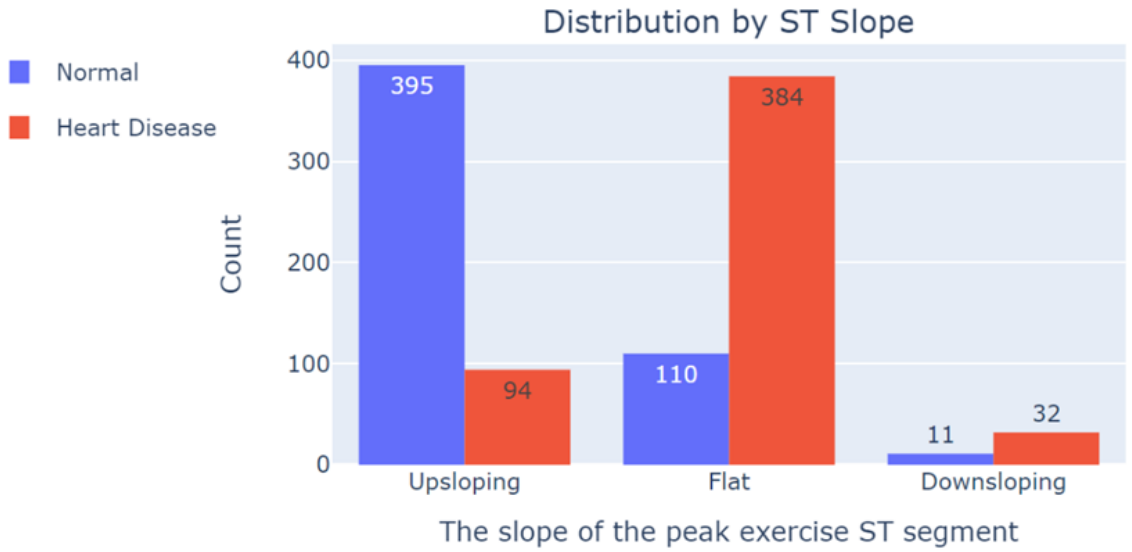
No.	Categorical Variable	Target	Degrees of freedom	Cramer's V	Strength
1	ST slope	Target	2	0.582	large
2	chest pain type	Target	3	0.394	large
3	age_group	Target	5	0.155	medium
4	exercise angina	Target	1	0.271	small ~ medium (close to medium)
5	resting ecg	Target	2	0.153	small
6	sex	Target	1	0.112	small
7	fasting blood sugar	Target	1	0.107	small
8	Maximum Heart Rate Status	Target	1	0.07	small

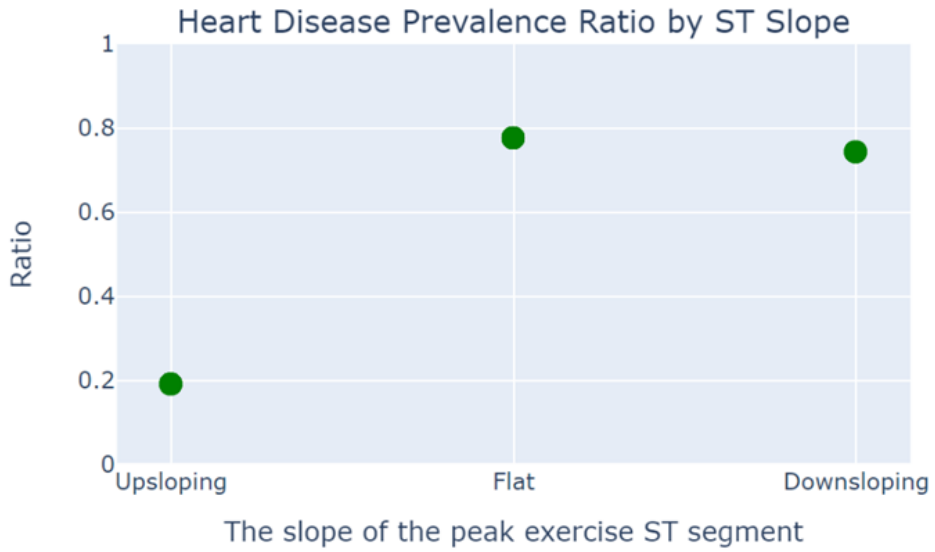
Univariate Analysis

1. ST slope

- Less common with upsloping ST slope
- Most heart disease cases show flat ST segment
- Flat ST segment causes complicate diagnosis



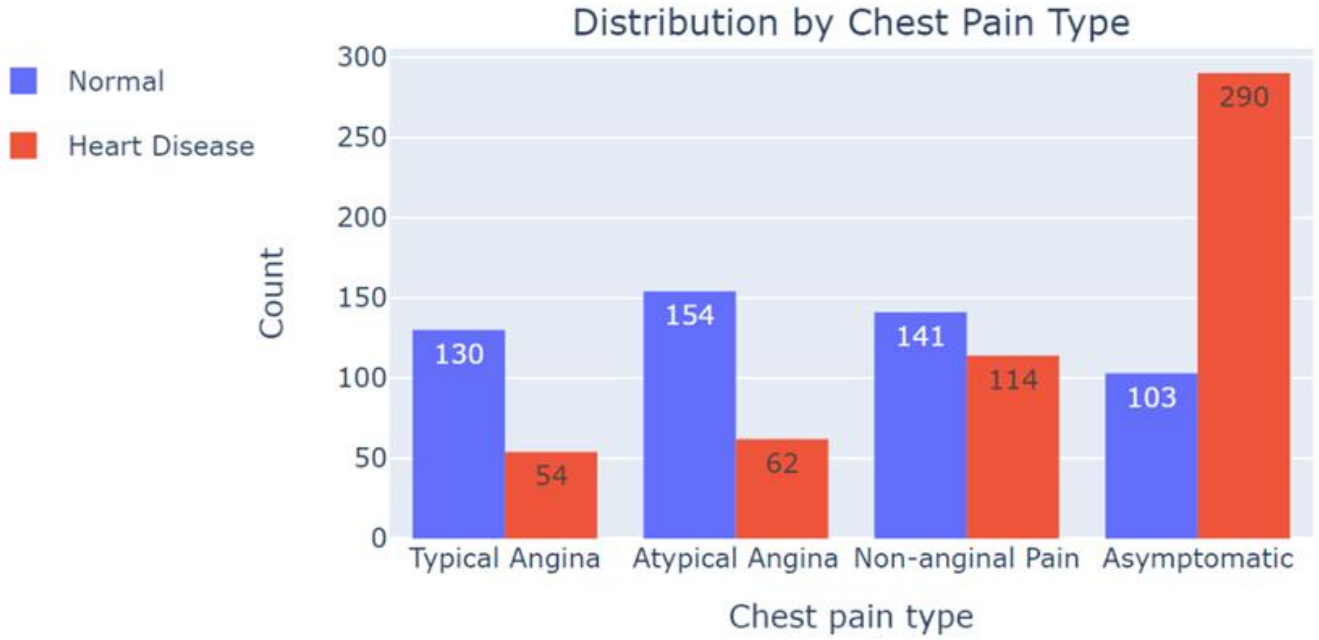


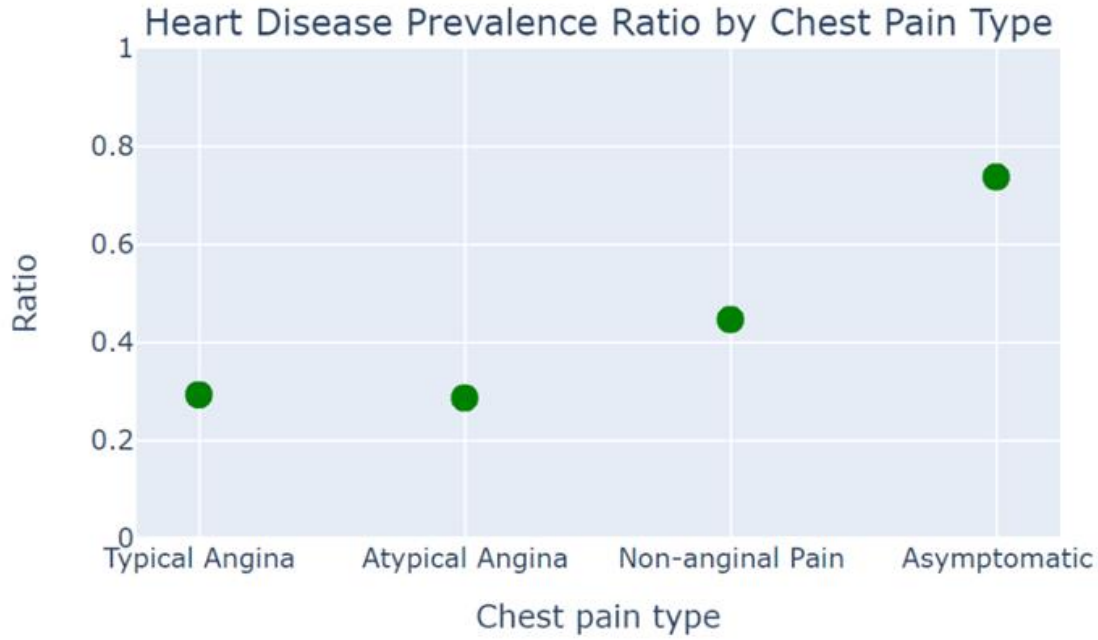


Univariate Analysis

2. Chest pain type

- Highest in number:
  Asymptomatic heart disease patients

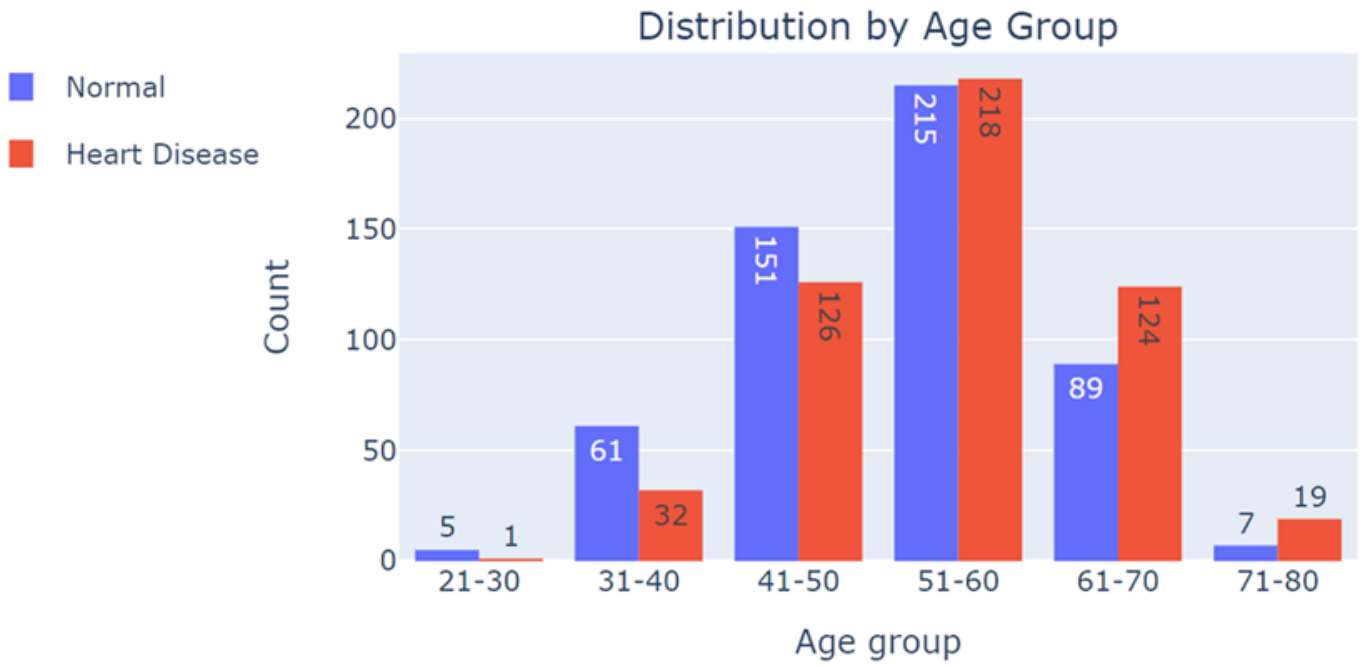


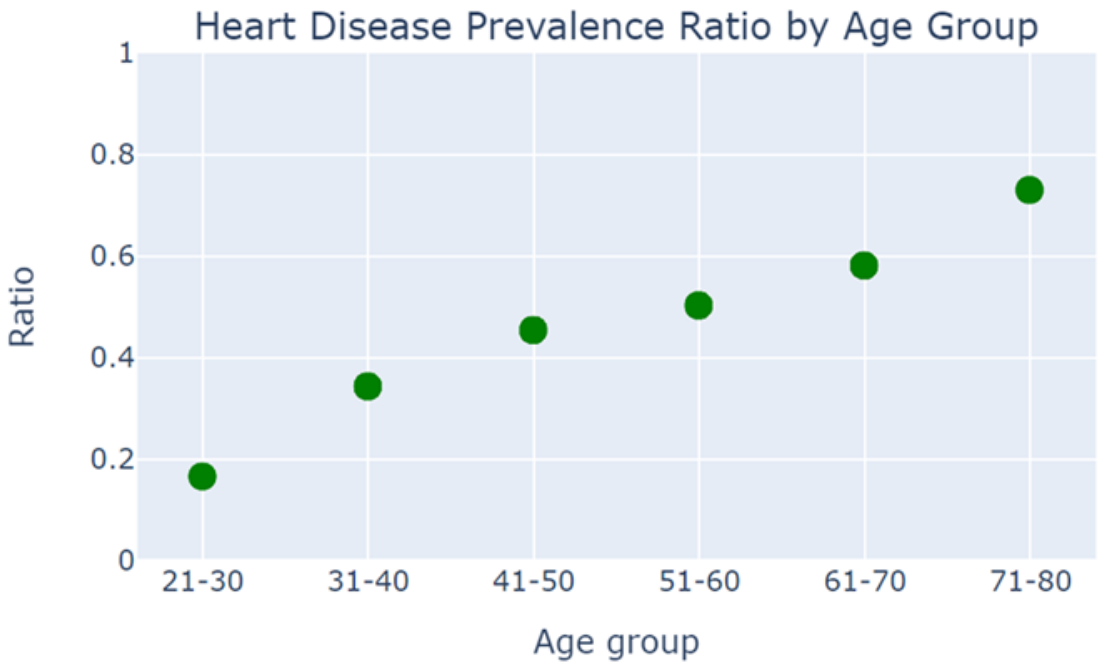


Univariate Analysis

3. Age group

- Positive association

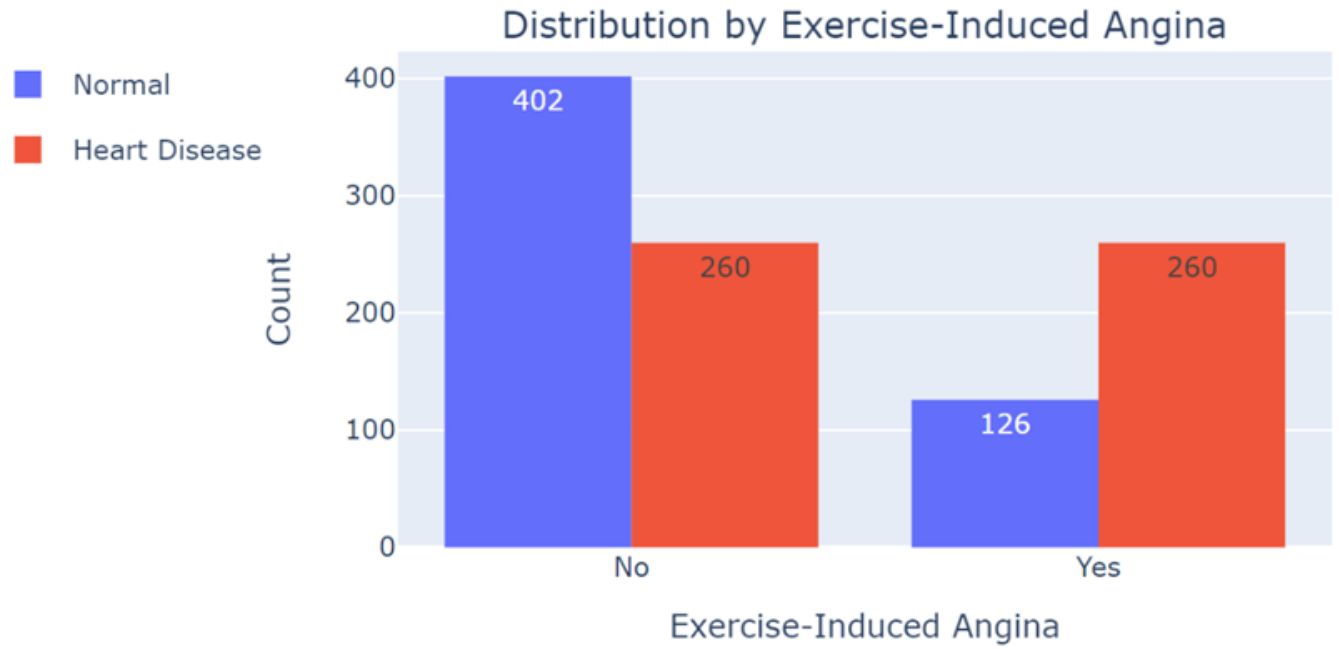


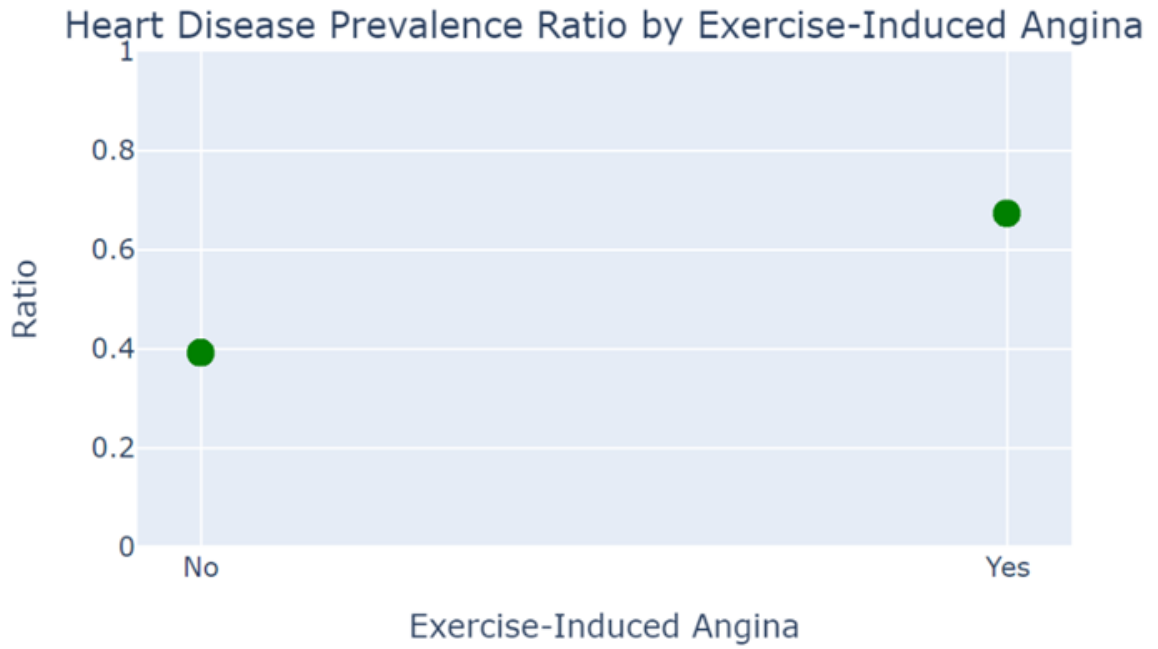


Univariate Analysis

4. Exercise-induced angina

- Heart disease patients: 50% with this condition



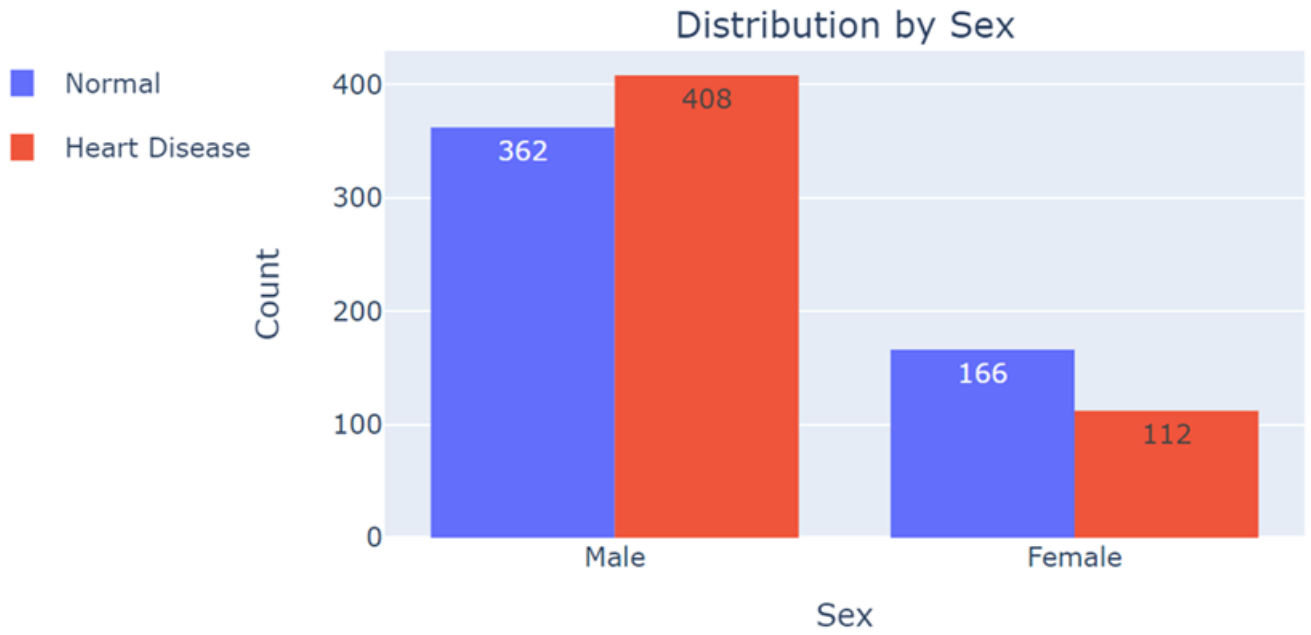


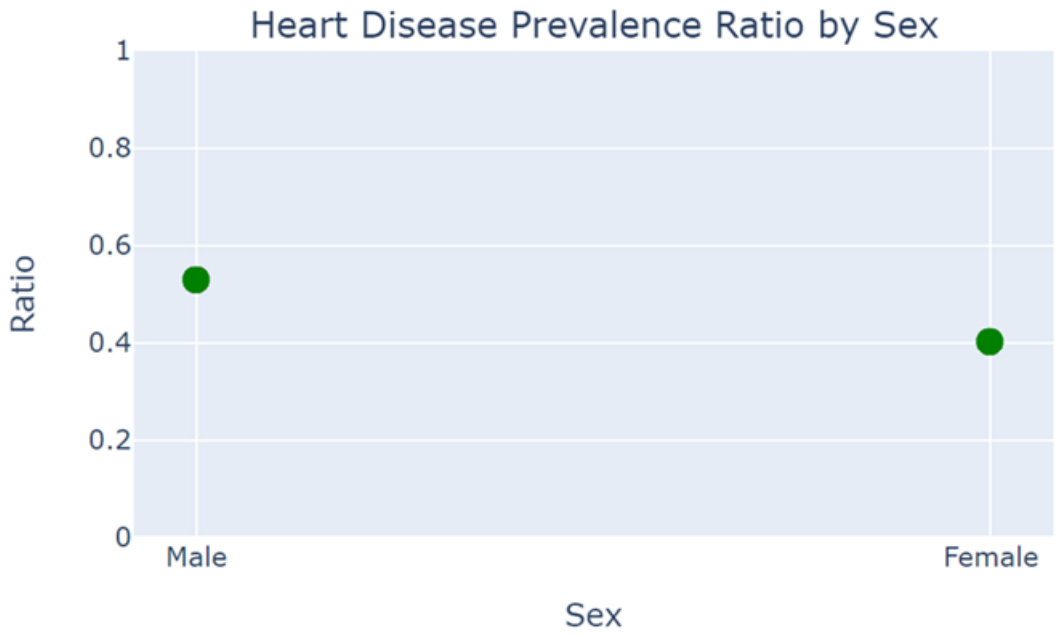
Univariate Analysis

5. Sex

Ratio of having heart disease by Sex: **Male Ratio > Female Ratio**

- Male Ratio: 408 / (408+362) = 0.53
- Female Ratio: 112 / (112+166) = 0.40





Univariate Analysis

B. Target VS Numeric Variable:

Correlation Coefficient	Correlation Interpretation
0.0 – 0.19	Weak
0.2 – 0.29	Normal
0.3 – 0.39	Strong
0.4 – 1.00	Very Strong

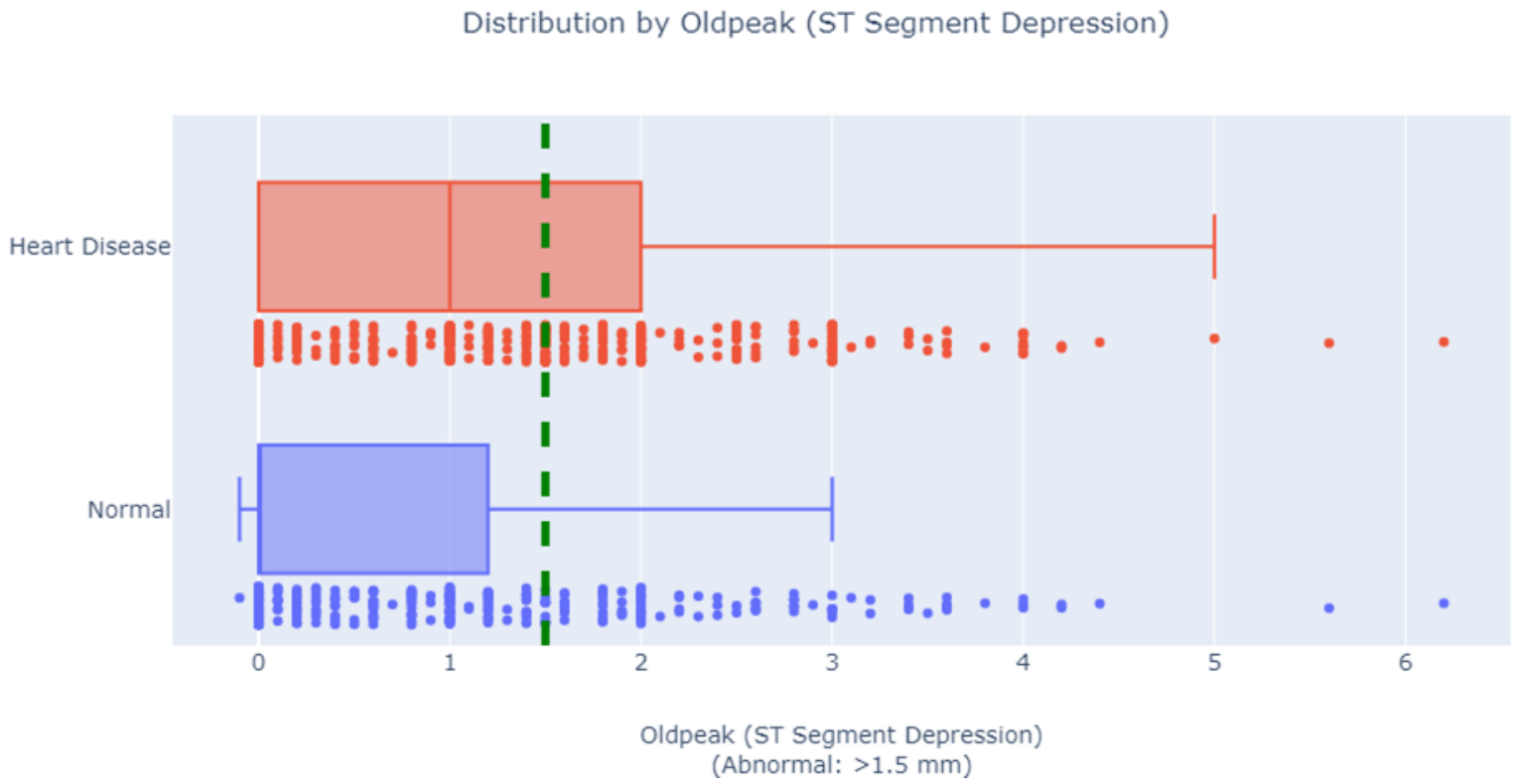
No.	Numeric	Target	p value	Point-Biserial Correlation Coefficient	
1	oldpeak	Target	0	0.2171	Normal, positive
2	resting bps	Target	0.0116	0.0779	weak, positive
3	cholesterol	Target	0.0690	p value > 0.05, cannot conclude	

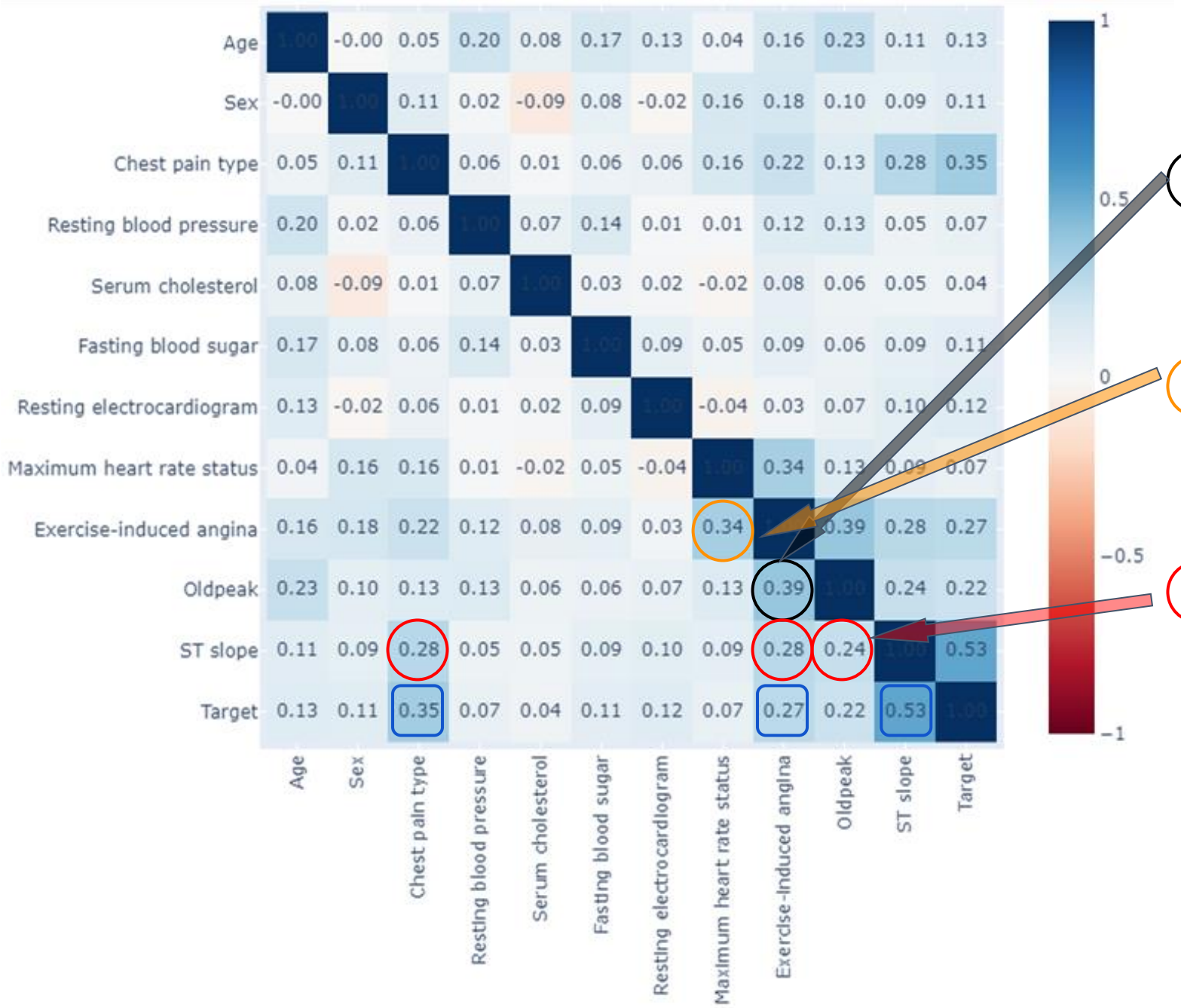
- Oldpeak has a normal correlation with heart disease.

Univariate Analysis

Oldpeak (ST Segment Depression)

- Heart disease patients tend to have higher oldpeak





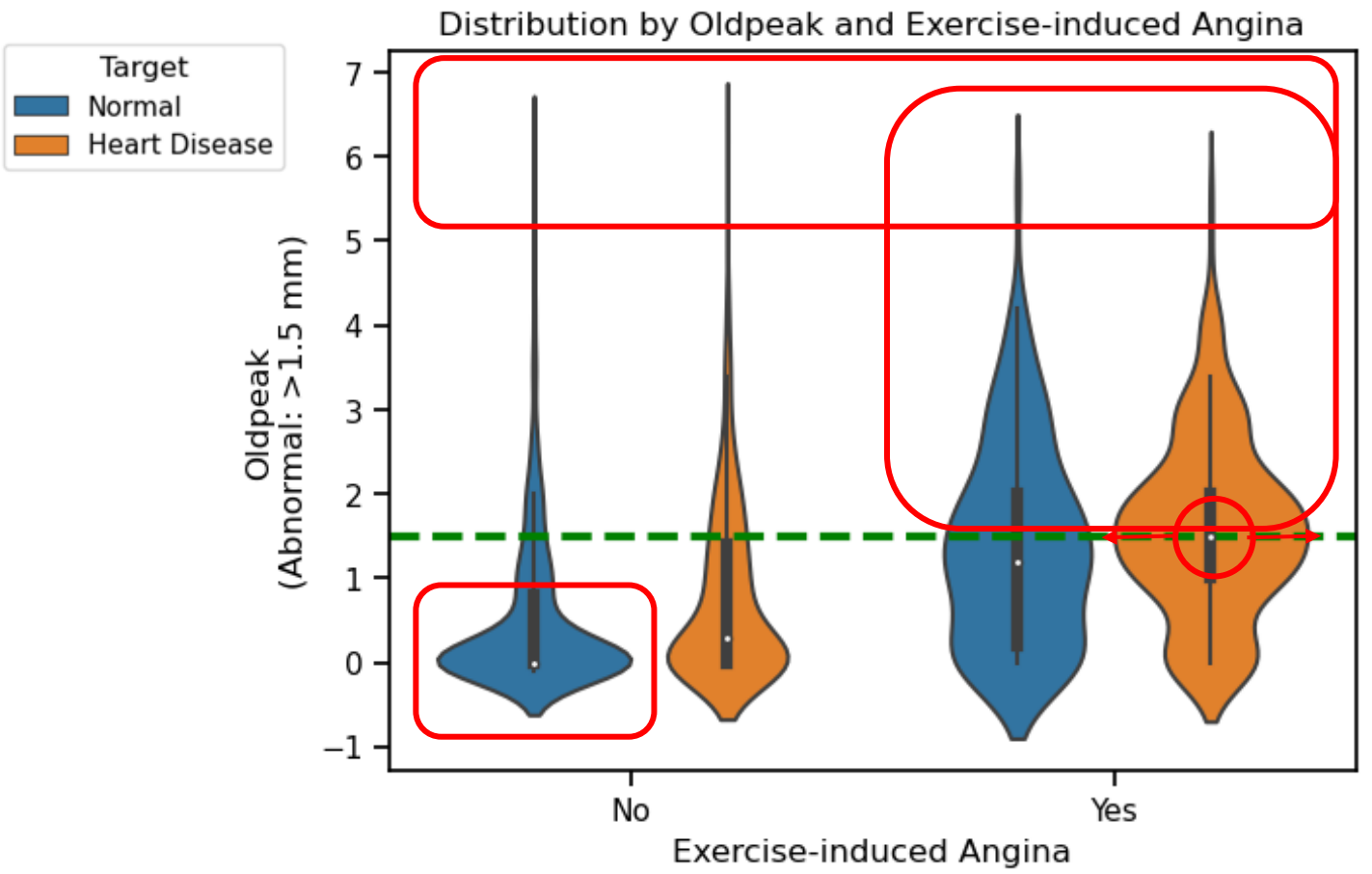
Heatmap

1. Oldpeak
   Exercise-induced angina

2. Maximum heart rate status
   Exercise-induced angina

3. ST slope
   Chest pain type
   Exercise-induced angina
   Oldpeak

Bivariate Analysis & Multivariate Analysis



Observation:

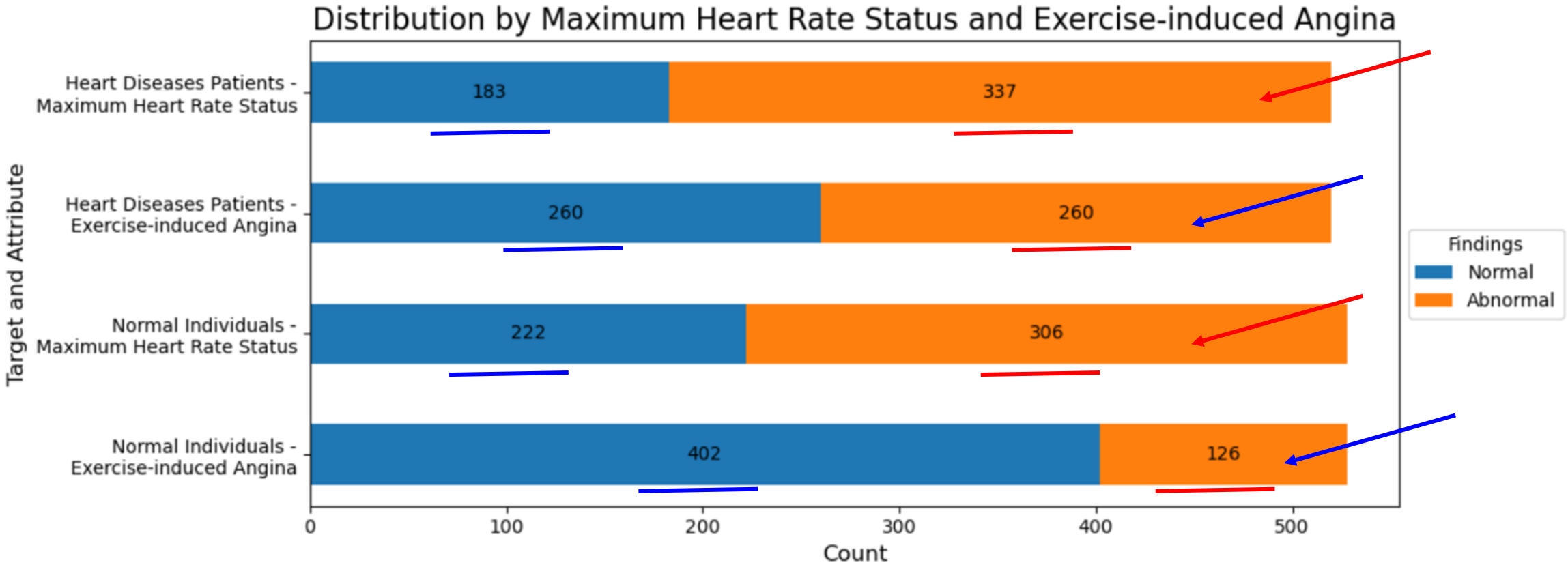
1. Exercise-induced angina:
   ↑ Oldpeak
   (esp. heart disease patients)

2. Heart disease patients with exercise-induced angina:
   - Highest median
   - Denser at higher oldpeak

3. Normal individuals without exercise-induced angina:
   - Denser at lower oldpeak

4. Extremely high oldpeak:
   - Outliers

Bivariate Analysis & Multivariate Analysis



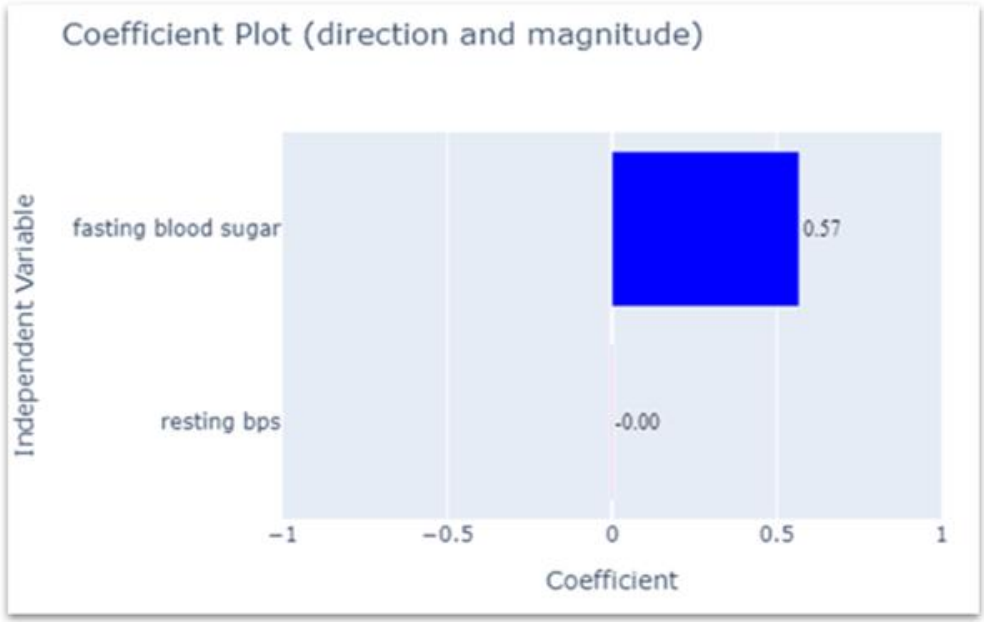
Observation:

1. Both attributes: Heart disease patients > Normal individuals
2. Both patient groups: Abnormal > Normal in Maximum heart rate
3. For exercise-induced angina:
   - Normal individuals: Normal > Abnormal
   - Heart disease patients: Normal = Abnormal

Bivariate Analysis & Multivariate Analysis

- **Logistics Regression**

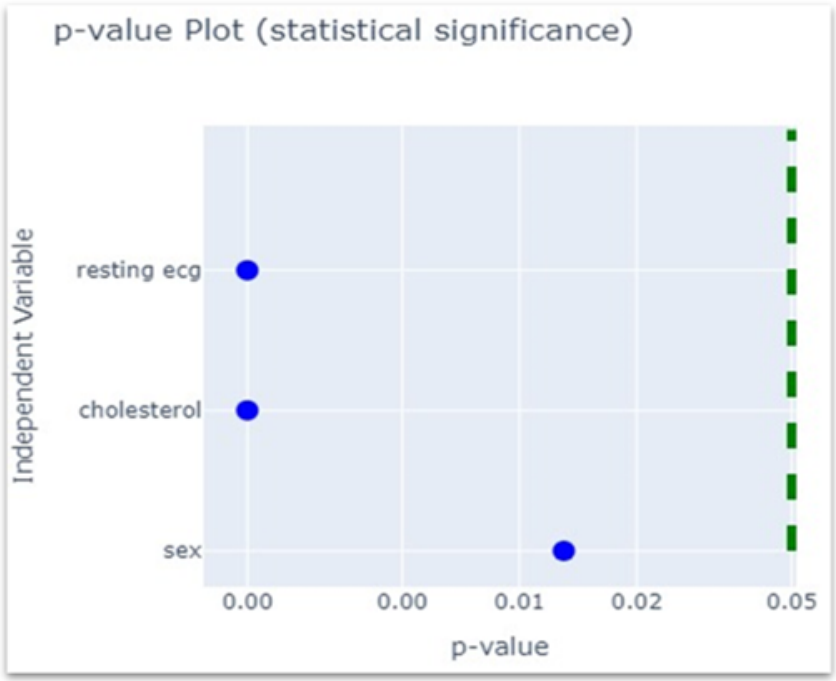
Dependent Variable	**Target**
Independent Variable	**Fasting Blood Sugar** (> 120 mg/dL) · **Resting Blood Pressure** · **Maximum Heart Rate Label**

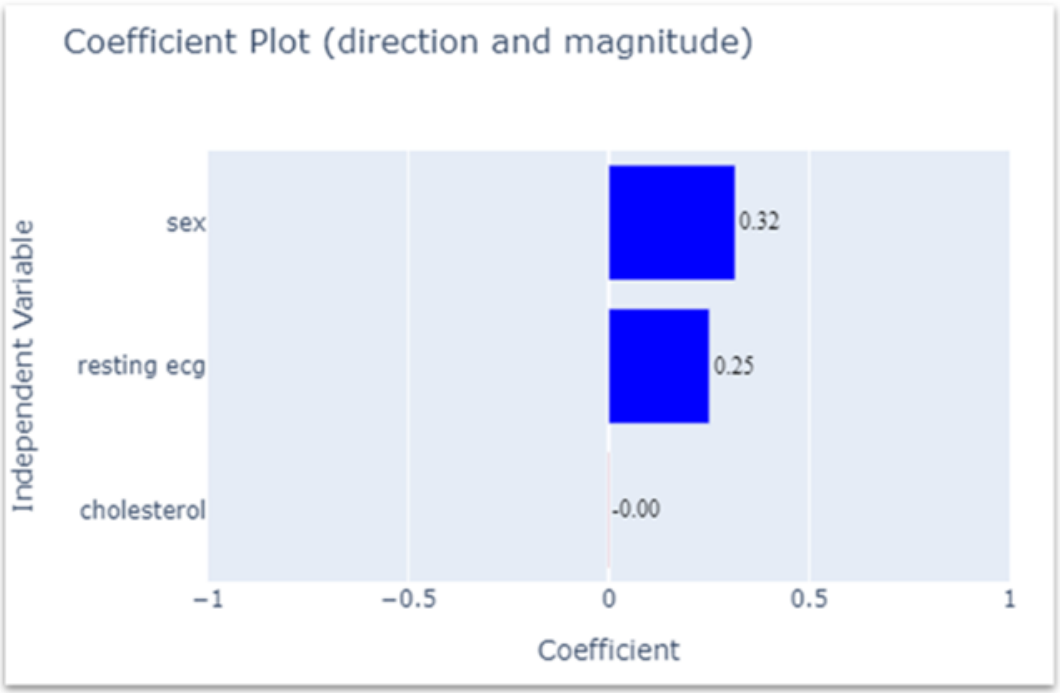


Bivariate Analysis & Multivariate Analysis

- **Logistics Regression**

Dependent Variable	**Target**
Independent Variable	**Sex**, **Resting electrocardiogram**, **Serum Cholesterol**





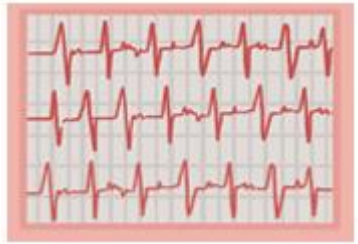
Bivariate Analysis & Multivariate Analysis

- **Logistics Regression**

p-value > 0.05

Dependent Variable	*Independent Variable*	Case 1	Case 2	Case 3	Case 4	Case 5
Target	Oldpeak	X	0.12	X	X	
	ST Slope	0.34	0.39	X		0.34
	Chest Pain Type	- 0.16	- 0.14		X	- 0.16
	Exercise-Induced Angina	0.60		0.55	0.66	0.62

Conclusion



- **variables most strongly related to heart disease**
  - ST Slope, Oldpeak, Chest Pain Type, Exercise-induced Angina
- **Key characteristics**

Require electrocardiogram assessment

Linked to exercise

- Variables from simple tests (e.g blood pressure test and blood test)
  - Weak association with Target

- Unchangeable variables (age, sex)
  - only age has a relatively stronger association with Target (but not as strong as electrocardiogram variables)

Since most heart diseases do not cause chest pain,
it is necessary to incorporate exercise ECG assessments into routine check-ups for early detection and management of cardiac conditions.

Reference

Dataset:

Heart Disease Dataset (kaggle.com)

Heart Disease Dataset (Comprehensive) | IEEE DataPort (ieee-dataport.org)

Chi-squared test & Cramer's V:

Contingency Tables, Chi-Squared and Cramer's V | by Jeffrey Hanif Watson | Towards Data Science

How to Interpret Cramer's V (With Examples) (statology.org)

Point-Biserial Correlation Coefficient:

Conduct and Interpret a Point-Biserial Correlation - Statistics Solutions

Point-Biserial Correlation in R. Point-biserial correlation is used to… | by Rahardito Dio Prastowo | Medium

Binary Logistic Regression:

Binary Logistic Regression – An introduction (datascienceinstitute.net)