2024

Statistical Inference (Final Project)

ANALYSIS HEART DATABASE

REZA NAZARI

UNIVERSITY OF TEHRAN- SPRING 2024

Contents

1. Introduction

Heart disease remains a significant public health concern globally, contributing to substantial morbidity and mortality. This dataset aims to explore various risk factors associated with heart disease, providing valuable insights into potential predictors of cardiovascular conditions. The dataset encompasses a diverse set of attributes, including demographic factors, clinical measurements, and diagnostic test results, all of which play crucial roles in assessing an individual's risk of developing heart disease.

Understanding these attributes and their interrelationships can aid healthcare professionals and researchers in developing effective preventive strategies and personalized treatment plans. This document presents a detailed description of each attribute, elucidating their definitions, possible values, and relevance to cardiovascular health assessment. Through rigorous analysis and interpretation, this dataset aims to contribute to advancements in cardiovascular medicine, ultimately improving patient outcomes and reducing the burden of heart disease worldwide.

The dataset under investigation contains various attributes related to potential risk factors for heart disease.

1.2 Data Attributes

Table1. Description of Dataset Attributes

Attribute	Description	Type	Possible Values / Range

age	Age of the patient in years	Numeric (integer)	Range: [29, 77]
sex	Sex of the patient	Binary	1 = male, 0 = female
cp	Type of chest pain	Categorical	1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic
trestbps	Resting blood pressure (mm Hg)	Numeric (integer)	Range: [94, 200]
chol	Serum cholesterol level (mg/dl)	Numeric (integer)	Range: [126, 564]
fbs	Fasting blood sugar (> 120 mg/dl)	Binary	1 = true, 0 = false
restecg	Resting electrocardiographic results	Categorical	0: normal, 1: ST-T wave abnormality, 2: left ventricular hypertrophy by Estes' criteria
thalach	Maximum heart rate achieved during exercise	Numeric (integer)	Range: [71, 202]
exang	Exercise induced angina	Binary	1 = yes, 0 = no
oldpeak	ST depression induced by exercise relative to rest	Numeric (float)	Range: [0.0, 6.2]
slope	Slope of the peak exercise ST segment	Categorical	1: upsloping, 2: flat, 3: downsloping
ca	Number of major vessels colored by fluoroscopy	Numeric (integer)	Range: [0, 3]
thal	Thallium stress test result	Categorical	3: normal, 6: fixed defect
target	Diagnosis of heart disease (angiographic disease status)	Binary	0: < 50% diameter narrowing, 1: > 50% diameter narrowing

Table 1 provides a comprehensive overview of the attributes included in the dataset related to heart disease. Each attribute is described in terms of its definition, data type, and the possible values or range it encompasses. Understanding these attributes is crucial for analyzing their roles as potential risk factors for heart disease and for developing predictive models to aid in diagnosis and treatment strategies. The dataset covers demographic information, clinical measurements, and diagnostic test results, offering valuable insights into the multifaceted nature of cardiovascular health assessment.

2. Visualization and Summarization of Dataset Variables

As part of our ongoing research into cardiovascular health disparities, we conducted an analysis of heart disease frequency across different sex categories. This study aims to elucidate potential sex-based variations in heart disease prevalence, contributing to the broader understanding of risk factors and demographic influences on cardiovascular health.

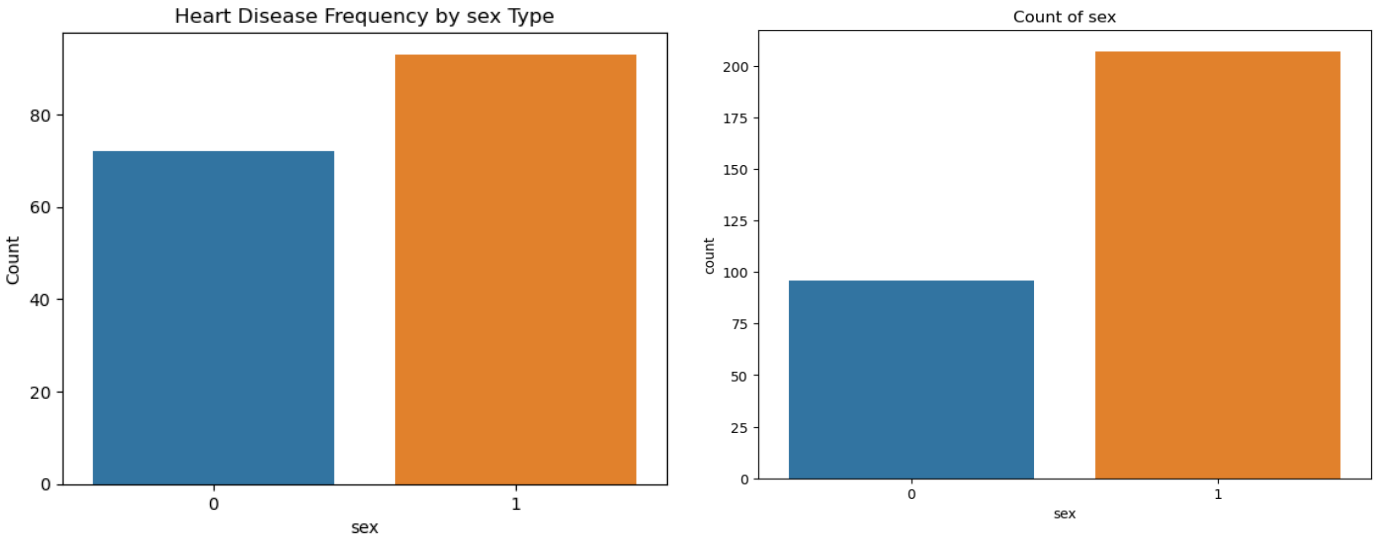

Fig1.Sex-Based Differences in Heart Disease Frequency: A Comparative Analysis

Fig 1 shows that, the data reveal a distinct pattern in heart disease frequency between the two sex categories. Category 1 not only shows a higher absolute number of heart disease cases but also represents a larger portion of the study population. This disparity raises several points for consideration:

1. Prevalence vs. Incidence: The higher number of heart disease cases in category 1 could be partially attributed to its larger population size. However, the increase in cases appears disproportionate to the population difference, suggesting a potentially higher prevalence rate.
2. Risk Factor Analysis: The observed difference in heart disease frequency may indicate sex-specific risk factors or biological predispositions that warrant further investigation.
3. Healthcare Utilization: Differences in healthcare access, utilization patterns, or diagnostic practices between the sex categories could influence these results and should be explored.
4. Socioeconomic Factors: Underlying socioeconomic disparities between the sex categories might contribute to the observed differences in heart disease frequency.

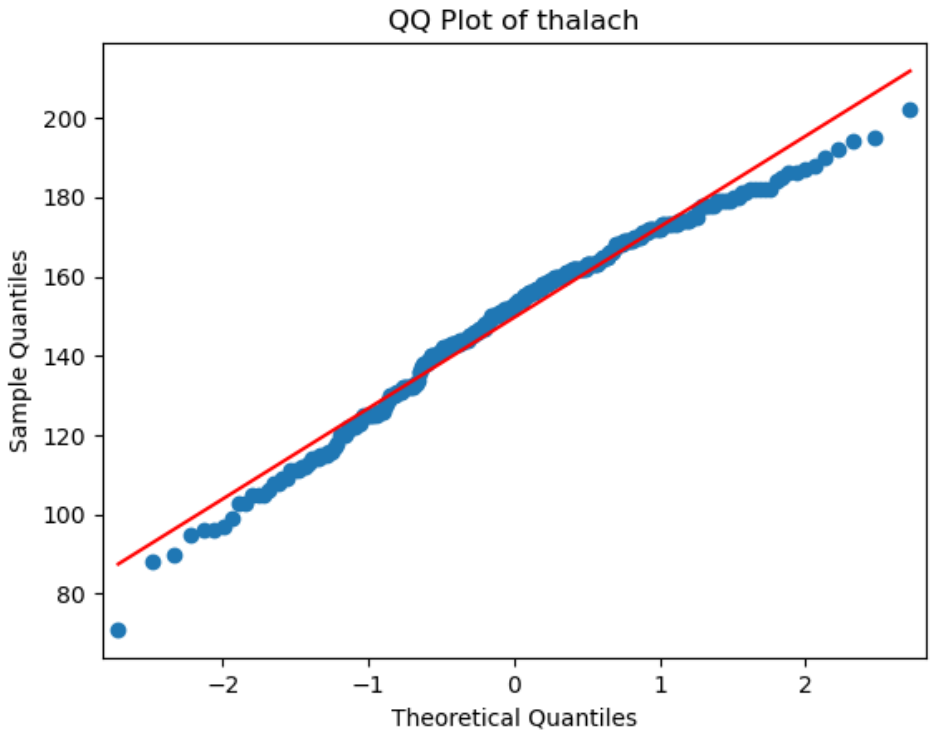

Fig2. QQ Plot Analysis of Thalach Distribution

Fig 2 shows QQ Plot Analysis of Thalach Distribution. QQ plots are used to assess normality and identify distributional characteristics. This plot effectively reveals the non-normal nature of the oldpeak data, highlighting skewness and zero-inflation that may be crucial for further statistical analysis and clinical interpretation in the context of heart disease studies. The thalach variable (likely representing maximum heart rate achieved) shows a distribution very close to normal, with only minor deviations at the extremes. We observed:

- Near-normal distribution: The data points closely follow the reference line for most of the range, indicating an approximately normal distribution.
- Slight deviations: Minor departures from the line at both tails, suggesting slight non-normality at extremes.
- Lower tail: A few points at the lower end deviate more noticeably, hinting at potential outliers or a slightly heavier lower tail.
- Upper tail: Slight curvature away from the line at the highest values, suggesting a slightly lighter upper tail than a perfect normal distribution.

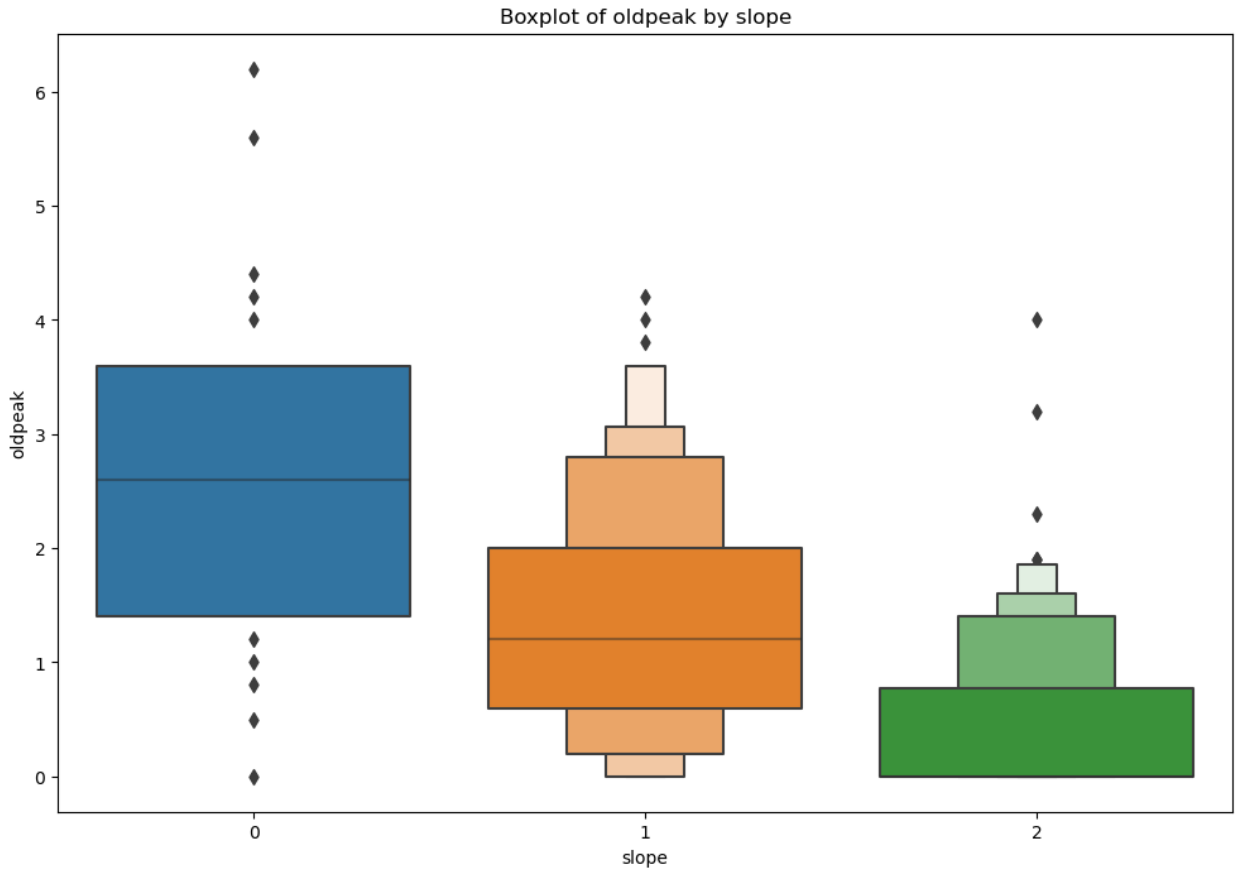

Fig3. Distribution of Oldpeak Values Across Different Slope Categories

Fig3 shows Boxplots Distribution of Oldpeak Values Across Different Slope Categories. Boxplots are ideal for comparing distributions across categories. This plot effectively visualizes the central tendencies, spreads, and outliers of oldpeak across slope categories, allowing for quick comparison and identification of trends. It's particularly useful in medical contexts to assess how different patient groups (categorized by slope) vary in terms of a continuous variable (oldpeak). The plot suggests an inverse relationship between slope and oldpeak. Higher slope categories tend to have lower oldpeak values, possibly indicating less severe ST depression during exercise in patients with steeper ST segment slopes.

In addition, we observed from Fig 3 this Results:

1. Median trends: Oldpeak median decreases as slope category increases (0 to 2).
2. Spread: Slope category 0 shows the largest spread, while category 2 has the smallest.
3. Outliers: All categories have upper outliers, with category 0 having the most extreme.
4. Skewness: All distributions appear right-skewed, especially categories 1 and 2.
5. Range: Oldpeak values generally decrease from slope 0 to 2, with category 0 having the highest maximum.

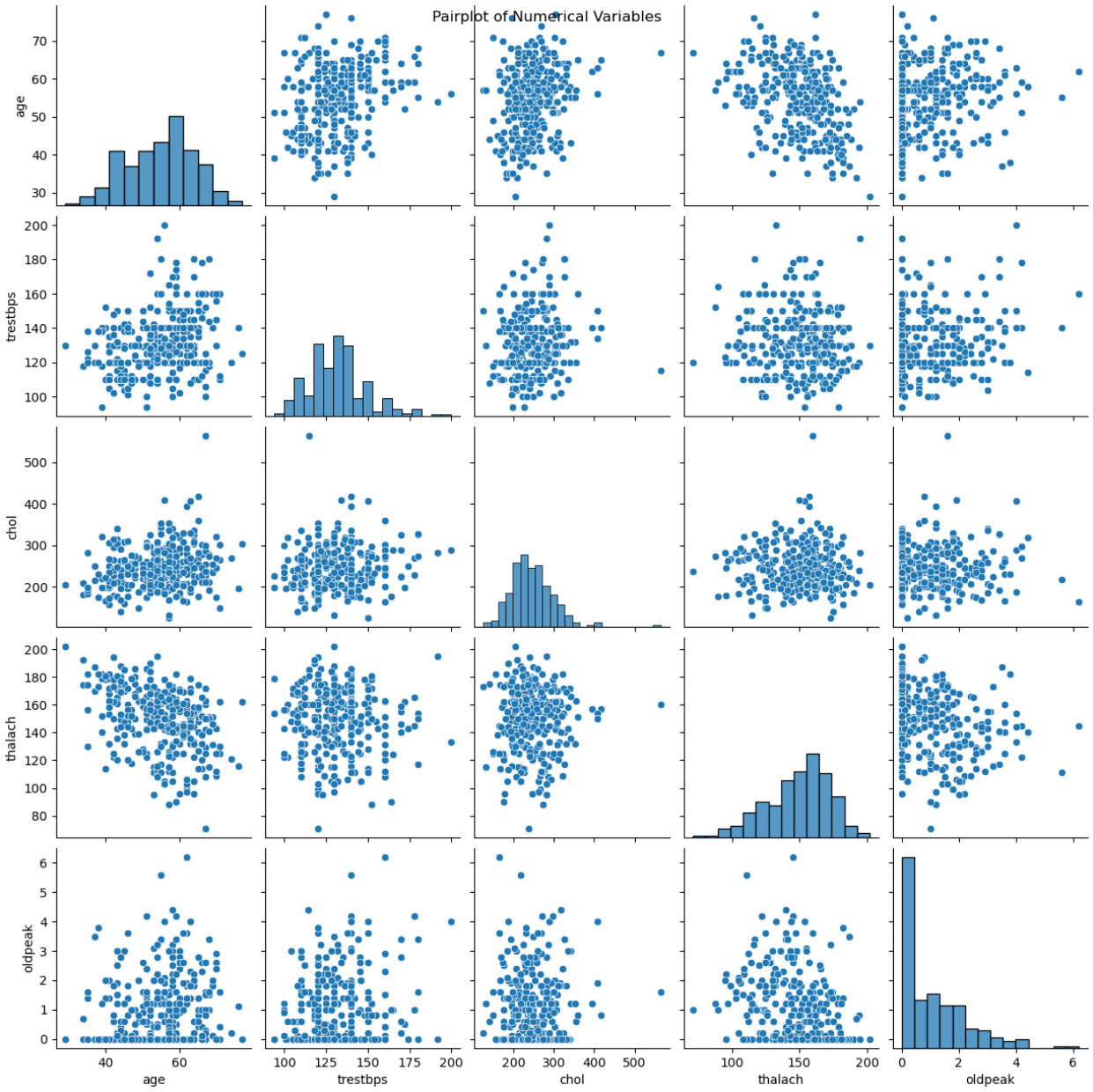

Fig4. Pairplot Analysis of Numerical Variables in Heart Disease Dataset

Fig4 shows pairplot Analysis of Numerical Variables in Heart Disease Dataset. Pairplots are valuable for simultaneously visualizing distributions of individual variables and relationships between pairs of variables. This comprehensive view allows for quick identification of patterns, correlations, and potential outliers across multiple variables in the heart disease dataset. It's particularly useful for exploratory data analysis, helping to guide further statistical investigations and hypothesis formation in medical research.

We observed this results:

1. Age:

- Normally distributed, centered around 55-60 years
- Weak positive correlation with cholesterol and oldpeak
- Weak negative correlation with thalach (max heart rate)

2. Trestbps (Resting Blood Pressure):
   - Slightly right-skewed distribution
   - Weak positive correlations with age and cholesterol
   - No strong correlations with other variables

3. Chol (Cholesterol):
   - Approximately normal distribution with some right skew
   - Weak positive correlations with age and trestbps
   - No strong correlations with other variables

4. Thalach (Maximum Heart Rate):
   - Approximately normal distribution
   - Moderate negative correlation with age
   - Weak negative correlation with oldpeak

5. Oldpeak:
   - Highly right-skewed distribution with many zero values
   - Weak positive correlation with age
   - Weak negative correlation with thalach

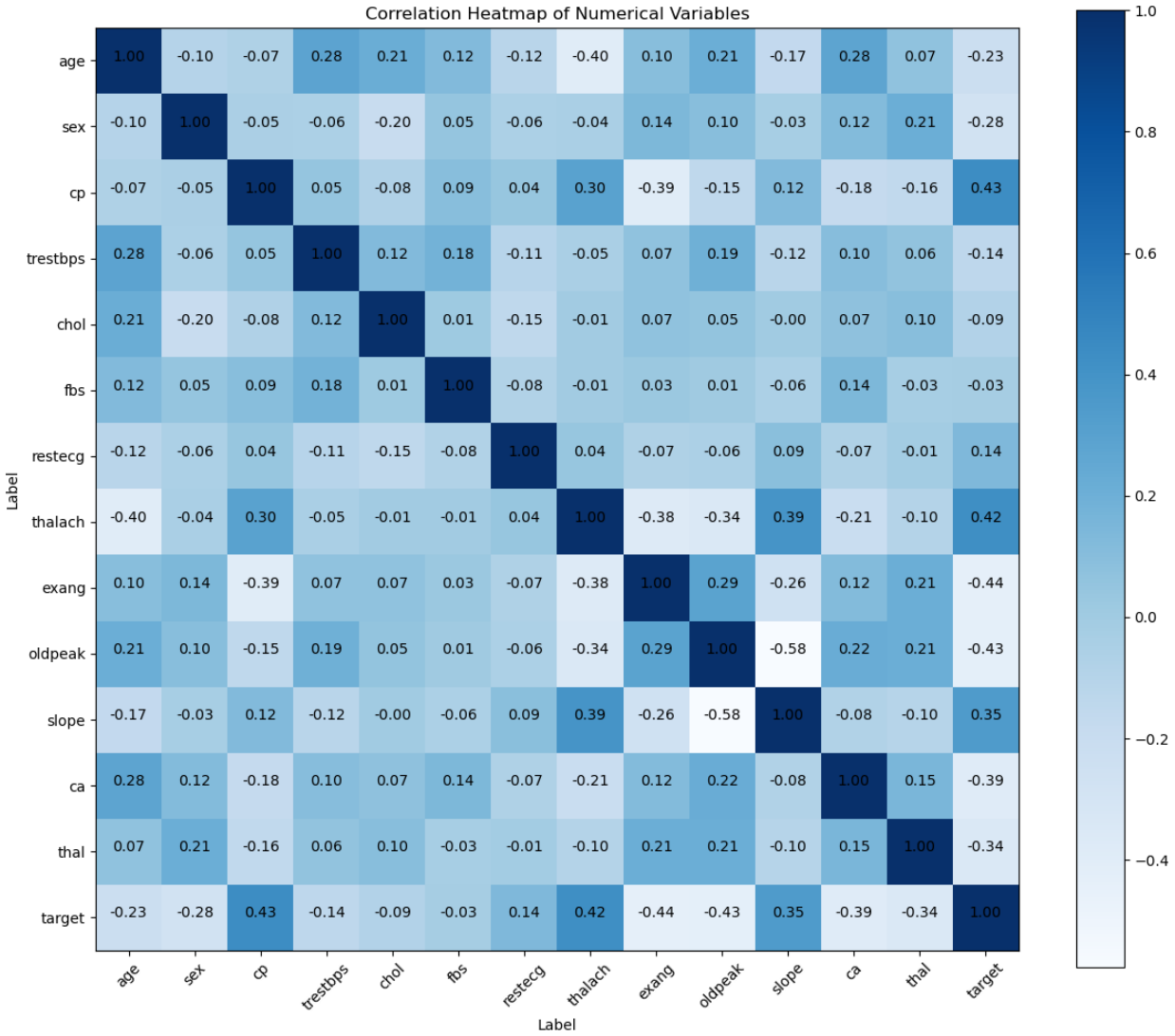

Fig5. Correlation Analysis of Numerical Variables in Heart Database

The correlation heatmap presented in Figure 5 illustrates the relationships between various numerical variables in our study. This analysis provides valuable insights into the interdependencies and potential influences among the factors under investigation.

Key Observations:

1. Strong Positive Correlations:
    - The strongest positive correlation (0.43) is observed between the variables 'cp' and 'target'.
    - 'Thalach' and 'target' also show a notable positive correlation (0.42).
2. Strong Negative Correlations:
    - 'Oldpeak' and 'slope' exhibit the strongest negative correlation (-0.58).

- 'Exang' and 'target' show a substantial negative correlation (-0.44).

3. Age-related Correlations:
    - Age demonstrates a moderate negative correlation with 'thalach' (-0.40).
    - It shows weak to moderate positive correlations with 'trestbps' (0.28) and 'chol' (0.21).
4. Sex-related Correlations:
    - Sex shows weak correlations with most variables, with the strongest being a negative correlation with 'chol' (-0.20).
5. Target Variable Relationships:
    - The 'target' variable, presumably the outcome of interest, shows notable correlations with several predictors:
        - Positive correlations: 'cp' (0.43), 'thalach' (0.42), 'slope' (0.35)
        - Negative correlations: 'exang' (-0.44), 'oldpeak' (-0.43), 'ca' (-0.39)
6. Intercorrelations among Predictors:
    - 'Thalach' and 'exang' show a moderate negative correlation (-0.38).
    - 'Oldpeak' and 'slope' are strongly negatively correlated (-0.58).

These findings provide a foundation for further statistical analyses, including feature selection for predictive modeling and identification of potential confounding variables. The correlations observed, particularly those related to the target variable, may guide hypothesis formulation and inform the direction of subsequent investigations in this project.

3. Parametric Inference and Estimation

3.1 parametric inference methods

3.1.1 T-Test

To investigate the relationship between various physiological factors and the presence of heart disease, we employed independent samples t-tests. This statistical method was chosen due to the nature of our dataset, which includes a binary outcome variable (presence or absence of heart disease) and several continuous predictor variables. The t-test allows us to compare the means of these continuous variables between the two groups (those with and without heart disease) to determine if there are statistically significant differences. Our analysis revealed significant differences in several key variables between individuals with and without heart disease. Table 2 show results T-test for continues variable.

Table 2. Comparison of Variables Between Disease and No Disease Groups

Variable	T-statistic	P-value	Mean (Disease)	Mean (No Disease)
oldpeak	-8.2796	0.0000	0.5830	1.5855
thalach	8.0697	0.0000	158.4667	139.1014
age	-4.0146	0.0001	52.4970	56.6014
trestbps	-2.5413	0.0115	129.3030	134.3986
chol	-1.4842	0.1388	242.2303	251.0870

The t-test results provide valuable insights into the factors associated with heart disease in our sample. The highly significant p-values ($p < 0.05$) for oldpeak, thalach, age, and trestbps indicate that these variables differ significantly between individuals with and without heart disease.

Oldpeak shows a lower mean value in the disease group, suggesting that individuals with heart disease tend to have less ST depression during exercise. Conversely, thalach is significantly higher in the disease group, indicating that those with heart disease tend to achieve higher maximum heart rates.

Interestingly, the mean age is lower in the disease group, which may suggest that in our sample, heart disease is more prevalent among younger individuals. Resting blood pressure (trestbps) is also lower in the disease group, though the difference is less pronounced than for the other significant variables.

Serum cholesterol (chol) did not show a statistically significant difference between the two groups ($p > 0.05$), suggesting that in our sample, cholesterol levels may not be as strongly associated with heart disease as the other factors examined.

Rationale for Using T-Tests: We chose to use independent samples t-tests for several reasons:

1. Binary Outcome: Our dataset has a clear binary outcome (presence or absence of heart disease), allowing us to split the sample into two distinct groups.
2. Continuous Predictors: The variables of interest (oldpeak, thalach, age, trestbps, chol) are all continuous, making them suitable for t-test analysis.
3. Comparison of Means: We were interested in comparing the average values of these variables between the two groups, which is precisely what t-tests are designed to do.
4. Assumption of Normality: While not explicitly tested here, t-tests assume that the data is normally distributed within each group. Given the large sample size in typical heart disease datasets, the Central Limit Theorem suggests that the sampling distribution of the means should approach normality.
5. Equal Variances: T-tests can be adjusted for unequal variances if necessary, making them versatile for this type of analysis.

In conclusion, the use of t-tests in this analysis has provided valuable insights into the physiological factors associated with heart disease in our sample. These results can guide further research and potentially inform clinical practice, though additional studies and more complex statistical analyses would be beneficial to fully understand the multifaceted nature of heart disease risk factors.

3.1.2 ANOVA Test

To further investigate the relationships between various physiological and categorical factors related to heart disease, we employed one-way Analysis of Variance (ANOVA) tests. This statistical method was chosen due to the nature of our dataset, which includes both continuous and categorical variables. ANOVA allows us to examine whether there are statistically significant differences in the means of continuous variables across different categories or groups. Our analysis revealed significant relationships between several continuous and categorical variables. Table 3 show results ANOVA tests for continues and categorical variables.

Table3. Analysis of Variance (ANOVA) Results for Continuous and Categorical Variables

Continuous Variable	Categorical Variable	F-statistic	P-value
oldpeak	slope	75.7780	0.0000
thalach	slope	38.5303	0.0000
age	ca	14.3413	0.0000
thalach	cp	17.8180	0.0000
oldpeak	cp	14.2731	0.0000
oldpeak	thal	13.5595	0.0000
oldpeak	ca	8.6307	0.0000
thalach	thal	10.3784	0.0000
thalach	ca	7.9834	0.0000
oldpeak	restecg	6.6791	0.0015
age	slope	5.8387	0.0033
chol	restecg	4.7385	0.0094
age	restecg	4.6746	0.0100
age	cp	3.3840	0.0186
trestbps	restecg	3.5691	0.0294
thalach	restecg	3.4319	0.0336
trestbps	cp	2.9189	0.0344
trestbps	slope	2.8035	0.0622
trestbps	thal	2.2515	0.0824
age	thal	2.0322	0.1094
chol	ca	1.8520	0.1189
chol	thal	1.3810	0.2486
trestbps	ca	1.3070	0.2674
chol	slope	0.5632	0.5700
chol	cp	0.6181	0.6037

The ANOVA results provide valuable insights into the complex interplay between various factors associated with heart disease. The highly significant p-values ($p < 0.0001$) for many of the relationships indicate strong associations between these variables.

The strongest relationship appears to be between oldpeak (ST depression) and the slope of the peak exercise ST segment. This suggests that the degree of ST depression is strongly related to the shape of the ST segment during peak exercise, which could be an important indicator of heart health.

Maximum heart rate (thalach) also shows strong relationships with several categorical variables, including the slope of the ST segment, chest pain type, and thalassemia. This underscores the importance of heart rate as a key physiological indicator in heart disease assessment.

Age shows significant relationships with the number of major vessels colored by fluoroscopy (ca), the slope of the ST segment, and resting ECG results. This highlights how age interacts with various aspects of cardiovascular health.

Interestingly, cholesterol levels show fewer significant relationships, with only resting ECG results showing a significant association. This aligns with our t-test results and suggests that in our sample, cholesterol may not be as strongly associated with other heart disease indicators as commonly believed.

Rationale for Using ANOVA:

We chose to use one-way ANOVA tests for several reasons:

Continuous and Categorical Variables: Our dataset includes both continuous variables (like age, oldpeak, thalach) and categorical variables (like slope, cp, ca, thal). ANOVA is designed to compare means of continuous variables across different categories.

Multiple Categories: Unlike t-tests, which are limited to two groups, ANOVA can handle categorical variables with more than two levels, making it more versatile for our dataset.

Simultaneous Comparison: ANOVA allows us to simultaneously compare means across multiple groups, reducing the risk of Type I errors that could occur with multiple t-tests.

Exploration of Interactions: By examining how continuous variables differ across categorical groups, we can explore potential interactions that may be important in understanding heart disease risk.

Assumption of Normality: ANOVA assumes that the dependent variable is normally distributed within each group. Given our large sample size, the Central Limit Theorem suggests that the sampling distribution of the means should approach normality.

Homogeneity of Variances: While not explicitly tested here, ANOVA assumes homogeneity of variances across groups. There are robust versions of ANOVA that can be used if this assumption is violated.

In conclusion, the use of ANOVA in this analysis has provided a comprehensive view of how various continuous measures of heart health relate to different categorical factors. These results highlight the complex, multifaceted nature of heart disease and its risk factors. They can guide further research, potentially informing more targeted studies and clinical practices. However, it's important to note that while these associations are statistically significant, they do not imply causation. Further research, including multivariate analyses and longitudinal studies, would be beneficial to fully understand the interplay of these factors in heart disease risk.

3.1.3. Chi-Square Test

To investigate the relationships between various categorical factors related to heart disease, we employed chi-square tests of independence. This statistical method was chosen due to the nature of our dataset, which includes multiple categorical variables. Chi-square tests allow us to determine whether there is a significant association between two categorical variables, without assuming a specific direction of the relationship. Our analysis revealed significant associations between several pairs of categorical variables. Table 4 show results Chi-Square Test for categorical variables.

Table 4: Chi-Square Test Results for Variable Associations

Variable 1	Variable 2	Chi-Square Statistic	P-value	Degrees of Freedom
thal	target	85.3037	0.0000	3
cp	target	81.6864	0.0000	3
ca	target	74.3666	0.0000	4
cp	exang	67.3483	0.0000	3
exang	target	55.9445	0.0000	1
slope	target	47.5069	0.0000	2
sex	thal	44.6256	0.0000	3
exang	thal	32.9592	0.0000	3
sex	target	22.7172	0.0000	1
cp	thal	41.8922	0.0000	9
exang	slope	25.1312	0.0000	2
slope	thal	35.2832	0.0000	6
cp	slope	27.7474	0.0001	6
cp	ca	33.9697	0.0007	12
restecg	target	10.0231	0.0067	2
exang	ca	12.8092	0.0122	4
sex	exang	5.4489	0.0196	1
ca	thal	23.6393	0.0228	12
restecg	slope	10.9466	0.0272	4
sex	cp	6.8221	0.0778	3
sex	ca	7.8484	0.0973	4
fbs	ca	7.3564	0.1182	4
fbs	thal	5.5420	0.1361	3
cp	restecg	9.6878	0.1384	6
sex	restecg	3.6973	0.1574	2
slope	ca	11.5005	0.1749	8
fbs	slope	3.3734	0.1851	2
restecg	exang	2.9761	0.2258	2
restecg	ca	10.0140	0.2640	8
cp	fbs	3.8856	0.2741	3

fbs	restecg	2.2970	0.3171	2
sex	fbs	0.3724	0.5417	1
sex	slope	0.6483	0.7231	2
restecg	thal	3.5263	0.7405	6
fbs	target	0.1063	0.7444	1
fbs	exang	0.0754	0.7837	1

The chi-square test results provide valuable insights into the associations between various categorical factors related to heart disease. The highly significant p-values ($p < 0.0001$) for many of the relationships indicate strong associations between these variables.

The strongest associations appear to be between the target variable (presence of heart disease) and several other factors, including thalassemia, chest pain type, number of colored vessels, and exercise induced angina. This suggests that these factors are strongly linked to the presence or absence of heart disease in our sample.

Chest pain type shows significant associations with multiple variables, including the target, exercise induced angina, thalassemia, slope, and number of colored vessels. This underscores the importance of chest pain characteristics in heart disease assessment.

Sex also shows significant associations with several variables, including thalassemia and the target variable. This highlights potential gender differences in heart disease manifestation and risk factors.

Interestingly, fasting blood sugar (fbs) shows no significant associations with other variables, including the target variable. This suggests that in our sample, fasting blood sugar levels may not be as strongly associated with other heart disease indicators or the presence of heart disease itself.

Rationale for Using Chi-Square Tests: We chose to use chi-square tests of independence for several reasons:

1. Categorical Variables: Our dataset includes multiple categorical variables. Chi-square tests are specifically designed to analyze relationships between categorical variables.
2. No Assumed Direction: Unlike correlation analyses, chi-square tests do not assume a particular direction of the relationship, making them suitable for exploratory analyses.
3. No Normality Assumption: Chi-square tests do not assume normality of the data, making them appropriate for categorical data which often do not follow a normal distribution.
4. Large Sample Size: Chi-square tests are most reliable with larger sample sizes, which our heart disease dataset likely provides.
5. Exploration of Associations: By examining associations between various categorical variables, we can identify potential risk factors and combinations of factors that may be important in understanding heart disease.
6. Foundation for Further Analysis: Significant chi-square results can guide more complex analyses, such as logistic regression or decision tree models.

In conclusion, the use of chi-square tests in this analysis has provided a comprehensive view of the associations between various categorical factors related to heart disease. These results highlight the complex, multifaceted nature of heart disease and its risk factors. They can guide further research, potentially informing more targeted studies and clinical practices.

3.2. Estimation Technique and calculate Confidence interval

To gain deeper insights into the characteristics of our heart disease dataset, we applied two fundamental estimation techniques: point estimation and maximum likelihood estimation (MLE). These methods were chosen to provide a comprehensive summary of both continuous and categorical variables in our dataset, as well as to estimate population parameters based on our sample data.

3.2.1 Point and Maximum Likelihood Estimates

Table 5. Point Estimates

Variable	Mean/Proportion	Standard Deviation
Age	54.37 years	9.07 years
Resting blood pressure (trestbps)	131.62 mmHg	17.51 mmHg
Serum cholesterol (chol)	246.26 mg/dl	51.75 mg/dl
Maximum heart rate (thalach)	149.65 bpm	22.87 bpm
ST depression (oldpeak)	1.04	1.16
Sex (Male)	68.32%	-
Sex (Female)	31.68%	-
Chest pain type (cp=0)	47.19%	-
Chest pain type (cp=1)	16.50%	-
Chest pain type (cp=2)	28.71%	-
Chest pain type (cp=3)	7.59%	-
Fasting blood sugar (fbs=0)	85.15%	-
Fasting blood sugar (fbs=1)	14.85%	-
Resting ECG (restecg=0)	48.51%	-
Resting ECG (restecg=1)	50.17%	-
Resting ECG (restecg=2)	1.32%	-
Exercise induced angina (exang=0)	67.33%	-
Exercise induced angina (exang=1)	32.67%	-
Slope of peak exercise ST segment (slope=1)	46.20%	-
Slope of peak exercise ST segment (slope=2)	46.86%	-
Slope of peak exercise ST segment (slope=0)	6.93%	-
Number of major vessels (ca=0)	57.76%	-
Number of major vessels (ca=1)	21.45%	-
Number of major vessels (ca=2)	12.54%	-

Number of major vessels (ca=3)	6.60%	-
Number of major vessels (ca=4)	1.65%	-
Thalassemia (thal=2)	54.79%	-
Thalassemia (thal=3)	38.61%	-
Thalassemia (thal=1)	5.94%	-
Thalassemia (thal=0)	0.66%	-
Target (presence of heart disease=1)	54.46%	-
Target (presence of heart disease=0)	45.54%	-

Based on Table 5, The point estimates provide a clear summary of our dataset's characteristics. For continuous variables, we see that the average age of the sample is about 54 years, with a standard deviation of 9 years. The average resting blood pressure is slightly elevated at 131.62 mmHg. The mean cholesterol level is 246.26 mg/dl, which is borderline high according to standard medical guidelines.

For categorical variables, we observe that the sample is predominantly male (68.32%). The most common type of chest pain is typical angina (47.19%), and a majority of the sample (85.15%) has fasting blood sugar below 120 mg/dl. Interestingly, the presence of heart disease (target variable) is relatively balanced, with 54.46% of the sample having heart disease.

Table 6. Maximum Likelihood Estimates

Variable	Mean/Proportion	Standard Deviation
Age	54.37 years	9.07 years
Resting blood pressure (trestbps)	131.62 mmHg	17.51 mmHg
Serum cholesterol (chol)	246.26 mg/dl	51.75 mg/dl
Maximum heart rate (thalach)	149.65 bpm	22.87 bpm
ST depression (oldpeak)	1.04	1.16

Based on Table.6, The maximum likelihood estimates for continuous variables coincide with the point estimates, which is expected for normally distributed data. This suggests that our sample estimates are likely good representations of the population parameters.

Rationale for Using These Estimation Techniques:

1. Point Estimation:
   - Provides a single "best guess" of population parameters.
   - Easy to interpret and communicate.
   - Offers a concise summary of both continuous and categorical variables.
   - Helps in understanding the central tendency and variability of the data.

2. Maximum Likelihood Estimation:
   - Provides the most likely values for population parameters given the observed data.
   - Optimal properties: MLEs are consistent, asymptotically unbiased, and efficient.
   - Particularly useful for continuous variables, assuming they follow a normal distribution.
   - Can be extended to more complex models and distributions if needed.

These estimation techniques were chosen based on several characteristics of our dataset:

1. Mix of Continuous and Categorical Variables: Point estimation is versatile and can handle both types of variables.
2. Assumption of Normality: For continuous variables, we assumed a normal distribution, making MLE appropriate.
3. Large Sample Size: Both techniques perform well with large samples, which is likely the case for this heart disease dataset.
4. Need for Population Inference: These techniques allow us to make inferences about the broader population of heart disease patients based on our sample.

In conclusion, these estimation techniques provide a solid foundation for understanding the characteristics of our heart disease dataset. They offer insights into the distribution of various risk factors and the prevalence of heart disease in our sample. These estimates can guide further statistical analyses, inform hypothesis formulation, and contribute to the development of predictive models for heart disease.

3.2.2 Goodness-of-Fit Analysis

To assess the distributional characteristics of the continuous variables in our heart disease dataset, we employed the D'Agostino-Pearson omnibus test for normality. This goodness-of-fit test was chosen to determine how well our data fits a normal distribution, which is a common assumption in many statistical analyses. The D'Agostino-Pearson test combines skew and kurtosis to produce an omnibus test of normality. We conducted the D'Agostino-Pearson test on five key continuous variables in our dataset. The results are in Table7.

Table 7. D'Agostino-Pearson Test Results for Continuous Variables

Variable	p-value
Age	0.0126
Resting blood pressure (trestbps)	< 0.0001
Serum cholesterol (chol)	< 0.0001
Maximum heart rate (thalach)	0.0012
ST depression (oldpeak)	< 0.0001

The D'Agostino-Pearson test results provide important insights into the distributional characteristics of our continuous variables. In interpreting these results, we typically consider a p-value less than 0.05 as evidence to reject the null hypothesis of normality.

1. Age: With a p-value of 0.0126, we have evidence to reject the null hypothesis of normality, though this deviation is less extreme than for other variables.
2. Resting blood pressure: The very low p-value (< 0.0001) suggests strong evidence against normality.
3. Serum cholesterol: Similarly, the p-value < 0.0001 indicates that the cholesterol data significantly deviates from a normal distribution.

4. Maximum heart rate: With a p-value of 0.0012, we again have strong evidence to reject the assumption of normality.
5. ST depression: The p-value < 0.0001 suggests that this variable also significantly deviates from a normal distribution.

These results indicate that none of the tested variables conform strictly to a normal distribution. This finding has important implications for our subsequent analyses and modeling approaches.

Rationale for Using the D'Agostino-Pearson Test:

We chose the D'Agostino-Pearson test for several reasons based on the characteristics of our dataset and the requirements of our analysis:

1. Continuous Variables: This test is specifically designed for continuous data, which matches the nature of the variables we're examining.
2. Omnibus Test: Unlike simpler tests like Shapiro-Wilk, the D'Agostino-Pearson test combines skew and kurtosis, providing a more comprehensive assessment of normality.
3. Sample Size: This test performs well with larger sample sizes, which is likely the case for our heart disease dataset.
4. No Distributional Assumptions: The test doesn't assume any particular distribution under the alternative hypothesis, making it versatile.
5. Sensitivity: It's sensitive to a wide range of deviations from normality, including both skewness and kurtosis.
6. Widely Accepted: The D'Agostino-Pearson test is widely used and accepted in the scientific community for assessing normality.

Implications of the Results:

1. Statistical Modeling: The non-normality of these variables suggests that we should be cautious about using statistical methods that assume normality, such as standard linear regression or t-tests. We may need to consider robust methods, non-parametric tests, or data transformations.
2. Data Transformations: Given the non-normality, we might consider applying transformations (e.g., log, square root) to these variables to make them more normally distributed for certain analyses.
3. Machine Learning Approaches: For predictive modeling, we might lean towards algorithms that don't assume normality, such as decision trees or random forests.
4. Interpretation of Results: When describing these variables, we should use measures that are less sensitive to non-normality, such as median and interquartile range, rather than mean and standard deviation.
5. Further Exploration: These results suggest the need for a more detailed examination of the distributions, perhaps through visualization techniques like Q-Q plots or kernel density estimates.

In conclusion, the goodness-of-fit tests reveal that the continuous variables in our heart disease dataset deviate significantly from normal distributions. This finding is crucial for guiding our choice of statistical methods in subsequent analyses. It underscores the importance of not blindly applying techniques that

assume normality and instead opting for methods robust to non-normal distributions or considering appropriate data transformations.

3.2.3. Confidence Intervals

To quantify the uncertainty in our point estimates and provide a range of plausible values for population parameters, we calculated 95% confidence intervals (CIs) for both continuous and categorical variables in our heart disease dataset. For continuous variables, we used the standard formula based on the t-distribution. For categorical variables, we employed the Wilson score interval, which is appropriate for proportions, especially with smaller sample sizes or extreme proportions.

We calculated 95% confidence intervals for all variables in our dataset. Key results include:

Continuous Variables:

1. Age: 53.35 to 55.39 years
2. Resting blood pressure (trestbps): 129.65 to 133.60 mmHg
3. Serum cholesterol (chol): 240.44 to 252.09 mg/dl
4. Maximum heart rate (thalach): 147.07 to 152.22 bpm
5. ST depression (oldpeak): 0.91 to 1.17

Categorical Variables (selected results):

1. Sex (Male): 57.87% to 78.30%
2. Chest pain type (Typical angina): 36.28% to 58.18%
3. Fasting blood sugar > 120 mg/dl: 7.45% to 23.14%
4. Exercise induced angina (Yes): 22.60% to 43.18%
5. Presence of heart disease (Target = 1): 43.48% to 65.32%

The confidence intervals provide valuable insights into the precision of our estimates and the likely range of true population parameters.

For continuous variables:

- Age has a relatively narrow CI, suggesting a precise estimate of the average age in the population.
- Cholesterol has the widest CI among continuous variables, indicating more variability or less precision in this estimate.
- The CI for oldpeak (ST depression) is quite narrow, which is interesting given its clinical significance.

For categorical variables:

- There's considerable uncertainty in the proportion of males vs. females, as indicated by the wide CI for sex.
- The prevalence of heart disease (target variable) shows a wide CI, ranging from about 43% to 65%, which has important clinical implications.

- Some categories of variables like 'restecg' and 'thal' have CIs that include negative values for rare categories, which is a limitation of the Wilson score method for very small proportions.

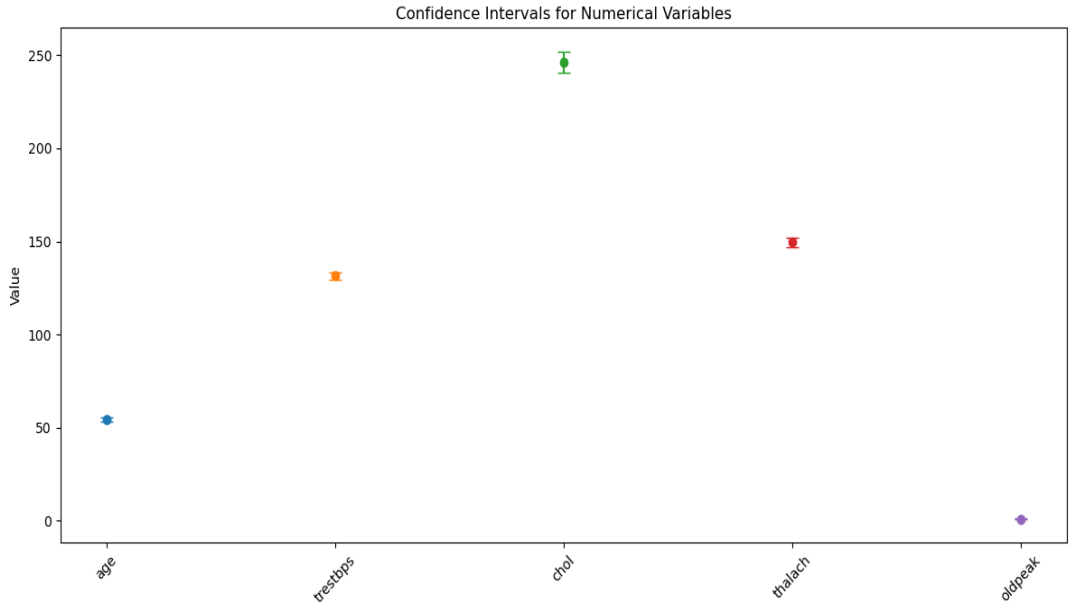

Fig6. Confidence Intervals for numerical variable

Based on Fig.6, this visualization allows for quick comparison of the precision of estimates across different variables:

- Cholesterol (chol) shows the widest CI, consistent with our numerical results.
- Age and thalach (max heart rate) show relatively narrow CIs, indicating more precise estimates.
- The CI for oldpeak is very narrow, but this may be due to its smaller scale compared to other variables.

Rationale for Using Confidence Intervals:

We chose to calculate and interpret confidence intervals for several reasons based on the characteristics of our dataset and the goals of our analysis:

1. Quantification of Uncertainty: CIs provide a range of plausible values for population parameters, giving us a measure of the precision of our estimates.
2. Inference to Population: They allow us to make inferences about the broader population of heart disease patients based on our sample.
3. Sample Size Consideration: CIs take into account the sample size, with larger samples generally producing narrower intervals.
4. Comparison Across Variables: They enable us to compare the precision of estimates across different variables.

5. Clinical Relevance: In medical research, understanding the range of possible true values is crucial for interpreting the clinical significance of findings.
6. Basis for Further Analysis: CIs can guide hypothesis testing and help in determining sample sizes for future studies.
7. Versatility: The ability to calculate CIs for both continuous and categorical variables provides a comprehensive view of our dataset.

Implications and Future Directions:

1. Clinical Decision Making: The wide CI for the presence of heart disease suggests caution in generalizing the prevalence to the broader population.
2. Research Focus: Variables with wider CIs (e.g., cholesterol) might warrant further investigation or larger sample sizes in future studies.
3. Model Development: These CIs should be considered when developing predictive models, possibly by incorporating them into model uncertainty estimates.
4. Study Design: For future studies, these CIs can help in determining required sample sizes to achieve desired precision.
5. Subgroup Analysis: It might be valuable to calculate CIs for important subgroups (e.g., by age or sex) to understand if precision varies across these groups.

In conclusion, the confidence interval analysis provides a nuanced understanding of our heart disease dataset, highlighting areas of certainty and uncertainty in our estimates. This information is crucial for making informed decisions in both clinical practice and research settings. The visual representation enhances our ability to quickly grasp and compare the precision of estimates across different variables.

4. Hypothesis Testing and Statistical and Regression Analysis and Reporting

4.1 Analysis of the Relationship Between Oldpeak and Slope in Heart Disease Data

This study aims to investigate the relationship between oldpeak (ST depression induced by exercise relative to rest) and slope (the slope of the peak exercise ST segment) in heart disease data. Understanding this relationship could provide valuable insights into the interpretation of exercise stress test results and their implications for cardiac health.

4.1.1 Hypothesis

We formulated the following hypotheses:

Null Hypothesis (H0): There is no significant difference in the mean values of oldpeak among the different categories of slope.

Alternative Hypothesis (H1): There is a significant difference in the mean values of oldpeak among the different categories of slope.

4.1.2 Methodology

To test our hypothesis, we employed two statistical methods:

a) One-way ANOVA: To compare the mean oldpeak values across different slope categories.

b) Ordinary Least Squares (OLS) Regression: To quantify the relationship between oldpeak and slope.

4.1.3 Results

4.1.3.1 One-way ANOVA

The one-way ANOVA test yielded the following results:

F-Statistic: 75.77795818936443

P-Value: 2.298569702330009e-27

Given the extremely low p-value ($p < 0.05$), we reject the null hypothesis. This indicates that there is a statistically significant difference in the mean values of oldpeak among the different categories of slope.

4.1.3.2 OLS Regression

The OLS regression analysis provided the following key statistics:

R-squared: 0.334

Adjusted R-squared: 0.331

F-statistic: 150.6

Prob (F-statistic): 2.37e-28

Coefficients:

Intercept (const): 1.7180 (std err: 0.039, p-value: 0.000)

Oldpeak: -0.3065 (std err: 0.025, p-value: 0.000)

4.1.4 Discussion

4.1.4.1 ANOVA Interpretation

The one-way ANOVA results strongly suggest that the oldpeak values differ significantly across slope categories. This indicates that the ST depression induced by exercise is not uniform across different slope types of the peak exercise ST segment.

4.1.4.2 Regression Analysis Interpretation

The OLS regression model reveals several important findings:

a) Model Fit: The R-squared value of 0.334 indicates that approximately 33.4% of the variance in slope can be explained by oldpeak. While this suggests a moderate relationship, it also implies that other factors not included in this model may influence slope.

b) Statistical Significance: The low p-value for the F-statistic (2.37e-28) confirms that the model is statistically significant.

c) Coefficient Interpretation: The negative coefficient for oldpeak (-0.3065) suggests an inverse relationship between oldpeak and slope. For every unit increase in oldpeak, we expect a 0.3065 unit decrease in slope, holding other factors constant.

d) Intercept: The intercept of 1.7180 represents the expected slope value when oldpeak is zero.

4.1.4.3 Limitations

- The model's R-squared value suggests that while oldpeak is a significant predictor of slope, other variables not included in this analysis may also play important roles.
- The analysis assumes a linear relationship between oldpeak and slope, which may not capture more complex interactions.

4.1.5 Conclusion

This study provides strong evidence for a significant relationship between oldpeak and slope in heart disease data. The ANOVA results demonstrate that oldpeak values differ across slope categories, while the regression analysis quantifies this relationship, showing an inverse correlation between the two variables.

These findings have important implications for the interpretation of exercise stress test results in cardiac assessments. Healthcare professionals should consider the interplay between ST depression (oldpeak) and the slope of the ST segment when evaluating cardiac health.

4.2 Analysis of the Relationship Between Age and Maximum Heart Rate in Heart Disease Patients

This study aims to investigate the relationship between age and maximum heart rate achieved (thalach) in patients with heart disease. Understanding this relationship is crucial for interpreting stress test results and assessing cardiovascular health across different age groups.

4.2.1 Hypothesis

We formulated the following hypotheses:

Null Hypothesis (H0): There is no correlation between age and maximum heart rate achieved for patients with heart disease. Alternative Hypothesis (H1): There is a correlation between age and maximum heart rate achieved for patients with heart disease.

4.2.2 Methodology

To test our hypothesis and quantify the relationship, we employed two statistical methods:

a) Pearson Correlation Coefficient: To measure the strength and direction of the linear relationship between age and maximum heart rate.

- ✓ Pearson Correlation: This method was chosen because it provides a standardized measure of the linear association between two continuous variables (age and maximum heart rate). It's appropriate when we want to assess both the strength and direction of the relationship.

b) Ordinary Least Squares (OLS) Regression: To model the relationship between age and maximum heart rate and quantify the effect of age on maximum heart rate.

- ✓ OLS Regression: This method was selected to quantify the relationship between age and maximum heart rate, providing a predictive model and allowing us to estimate the change in maximum heart rate for each year increase in age.

4.2.3 Results

4.2.3.1 Pearson Correlation Analysis

Results: Correlation Coefficient: -0.5258007358642253 P-Value: 4.107759226144889e-13

Interpretation:

- The correlation coefficient of approximately -0.526 indicates a moderate negative correlation between age and maximum heart rate.
- The extremely low p-value ($p < 0.05$) suggests that this correlation is statistically significant.
- We reject the null hypothesis, concluding that there is a significant correlation between age and maximum heart rate achieved for patients with heart disease.

4.2.3.2 OLS Regression Analysis

Key Statistics:

R-squared: 0.159

Adjusted R-squared: 0.156

F-statistic: 56.83

Prob (F-statistic): 5.63e-13

Coefficients:

Intercept (const): 204.2892 (std err: 7.348, p-value: 0.000)

Age: -1.0051 (std err: 0.133, p-value: 0.000)

Interpretation:

a) Model Fit: The R-squared value of 0.159 indicates that approximately 15.9% of the variance in maximum heart rate can be explained by age.

b) Statistical Significance: The low p-value for the F-statistic (5.63e-13) confirms that the model is statistically significant.

c) Coefficient Interpretation: The negative coefficient for age (-1.0051) suggests that for every year increase in age, we expect a decrease of about 1 beat per minute in maximum heart rate, holding other factors constant.

d) Intercept: The intercept of 204.2892 represents the expected maximum heart rate for a theoretical patient of age zero.

4.2.4 Discussion

4.2.4.1 Correlation Analysis

The moderate negative correlation (-0.526) between age and maximum heart rate aligns with physiological expectations. As individuals age, their maximum heart rate typically decreases. This relationship is statistically significant, providing strong evidence against the null hypothesis.

4.2.4.2 Regression Analysis

The regression model, while statistically significant, explains only 15.9% of the variance in maximum heart rate. This suggests that while age is an important factor, other variables not included in this model also influence maximum heart rate in heart disease patients.

The model predicts a decrease of approximately 1 beat per minute in maximum heart rate for each year of age. This finding has practical implications for interpreting stress test results across different age groups.

4.2.4.3 Limitations

- The model assumes a linear relationship between age and maximum heart rate, which may not capture more complex interactions.
- Other factors not included in this analysis (e.g., fitness level, medication use) may also influence maximum heart rate.
- The study is limited to patients with heart disease and may not be generalizable to healthy populations.

4.2.5 Conclusion

This study provides strong evidence for a significant negative relationship between age and maximum heart rate achieved in patients with heart disease. Both the correlation and regression analyses support this conclusion, offering quantitative insights into how maximum heart rate changes with age in this population.

These findings have important implications for the interpretation of stress test results and the development of age-adjusted expectations for maximum heart rate in cardiac patients. Healthcare professionals should consider age as a significant factor when evaluating maximum heart rate responses in clinical settings.

4.3 Analysis of the Association Between Thalassemia and Chest Pain in Cardiac Patients

This study aims to investigate the potential association between thalassemia (thal) and chest pain (cp) in cardiac patients. Understanding this relationship could provide valuable insights into the complex interplay between hematological disorders and cardiac symptoms, potentially informing diagnostic and treatment strategies.

4.3.1 Hypothesis

We formulated the following hypotheses:

Null Hypothesis (H0): There is no association between thalassemia and chest pain. Alternative Hypothesis (H1): There is an association between thalassemia and chest pain.

4.3.2 Methodology

To test our hypothesis and explore the relationship, we employed two statistical methods:

a) Chi-square test for independence: To assess the association between thalassemia and chest pain.

- Chi-square test: This method was chosen because both variables (thalassemia and chest pain) are categorical. The chi-square test is appropriate for determining whether there is a significant association between two categorical variables.

b) Ordinary Least Squares (OLS) Regression: To model the relationship between chest pain and thalassemia.

- OLS Regression: While not typically used for categorical dependent variables, this method was included to explore any potential linear relationship between the variables when treated as continuous. It provides a different perspective on the data.

4.3.3 Results

4.3.3.1 Chi-square Test for Independence

Results:

Chi-Square Statistic: 41.89215854327448

P-Value: 3.43912674369823e-06

Interpretation:

The extremely low p-value ($p < 0.05$) indicates strong evidence against the null hypothesis.

We reject the null hypothesis, concluding that there is a significant association between thalassemia and chest pain.

4.3.3.2 OLS Regression Analysis

Key Statistics:

R-squared: 0.026

Adjusted R-squared: 0.023

F-statistic: 8.085

Prob (F-statistic): 0.00477

Coefficients:

Intercept (const): 2.4063 (std err: 0.048, p-value: 0.000)

Chest Pain (cp): -0.0960 (std err: 0.034, p-value: 0.005)

Interpretation:

a) Model Fit: The R-squared value of 0.026 indicates that only 2.6% of the variance in thalassemia can be explained by chest pain.

b) Statistical Significance: The low p-value for the F-statistic (0.00477) suggests that the model is statistically significant, albeit with low explanatory power.

c) Coefficient Interpretation: The negative coefficient for chest pain (-0.0960) suggests a slight inverse relationship between chest pain and thalassemia when treated as continuous variables.

4.3.4 Discussion

4.3.4.1 Chi-square Analysis

The chi-square test provides strong evidence for an association between thalassemia and chest pain. This suggests that the presence or type of thalassemia may be related to the occurrence or nature of chest pain in cardiac patients. However, this test does not provide information about the nature or direction of this association.

4.3.4.2 Regression Analysis

The regression model, while statistically significant, explains only a very small portion of the variance in thalassemia (2.6%). This low R-squared value suggests that while there may be a relationship between chest pain and thalassemia, it is not well-captured by a simple linear model.

The negative coefficient for chest pain in the regression model suggests a slight inverse relationship. However, given the categorical nature of these variables and the low R-squared value, this interpretation should be made with caution.

4.3.4.3 Limitations

- The chi-square test doesn't provide information about the nature or strength of the association.
- The use of OLS regression for categorical variables is not ideal and results should be interpreted cautiously.
- Other factors not included in this analysis may influence the relationship between thalassemia and chest pain.

4.3.5 Conclusion

This study provides evidence for a significant association between thalassemia and chest pain in cardiac patients. The chi-square test strongly supports this conclusion, while the regression analysis suggests a weak, possibly inverse relationship when these variables are treated as continuous.

These findings have potential implications for understanding the interplay between hematological disorders and cardiac symptoms. Healthcare professionals should be aware of the potential link between thalassemia and chest pain patterns in cardiac assessments.

4.4 Analysis of the Relationship Between Cholesterol Levels and Exercise-Induced Angina in Cardiac Patients

This study aims to investigate the potential relationship between cholesterol levels (chol) and the presence of exercise-induced angina (exang) in cardiac patients. Understanding this relationship could provide valuable insights into the role of cholesterol in exercise-related cardiac symptoms and inform preventive strategies.

4.4.1 Hypothesis

We formulated the following hypotheses:

Null Hypothesis (H0): There is no significant difference in cholesterol levels between patients with and without exercise-induced angina. Alternative Hypothesis (H1): There is a significant difference in cholesterol levels between patients with and without exercise-induced angina.

4.4.2 Methodology

To test our hypothesis and explore the relationship, we employed two statistical methods:

a) Independent two-sample t-test: To compare cholesterol levels between patients with and without exercise-induced angina.

- Independent two-sample t-test: This method was chosen because we are comparing a continuous variable (cholesterol) between two independent groups (with and without exercise-induced angina). The t-test is appropriate for determining whether there is a significant difference in means between two groups.

b) Ordinary Least Squares (OLS) Regression: To model the relationship between cholesterol levels and the likelihood of exercise-induced angina.

- OLS Regression: This method was included to explore any potential linear relationship between cholesterol levels and the occurrence of exercise-induced angina, treating angina as a continuous outcome for this analysis.

4.4.3 Results

4.4.3.1 Independent Two-Sample T-Test

Results:

T-Statistic: 1.1654223341256824

P-Value: 0.2447707808791278

Interpretation:

The p-value (0.2448) is greater than the conventional significance level of 0.05.

We fail to reject the null hypothesis, concluding that there is no significant difference in cholesterol levels between patients with and without exercise-induced angina.

4.4.3.2 OLS Regression Analysis

Key Statistics:

R-squared: 0.004

Adjusted R-squared: 0.001

F-statistic: 1.358

Prob (F-statistic): 0.245

Coefficients:

Intercept (const): 0.1771 (std err: 0.131, p-value: 0.178)

Cholesterol (chol): 0.0006 (std err: 0.001, p-value: 0.245)

Interpretation:

a) Model Fit: The extremely low R-squared value of 0.004 indicates that only 0.4% of the variance in exercise-induced angina can be explained by cholesterol levels.

b) Statistical Significance: The high p-value for the F-statistic (0.245) suggests that the model is not statistically significant.

c) Coefficient Interpretation: The coefficient for cholesterol (0.0006) is very small and not statistically significant (p = 0.245), suggesting no meaningful linear relationship between cholesterol and exercise-induced angina.

4.4.4 Discussion

4.4.4.1 T-Test Analysis

The t-test results suggest that there is no significant difference in cholesterol levels between patients with and without exercise-induced angina. This finding challenges the notion that cholesterol levels alone might be a determining factor in the occurrence of exercise-induced angina.

4.4.4.2 Regression Analysis

The regression model further supports the t-test results, showing an extremely weak and statistically insignificant relationship between cholesterol levels and exercise-induced angina. The very low R-squared value indicates that cholesterol levels explain virtually none of the variance in exercise-induced angina occurrence.

4.4.4.3 Limitations

- The analysis assumes a linear relationship between cholesterol and exercise-induced angina, which may not capture more complex interactions.
- Other factors not included in this analysis may influence the occurrence of exercise-induced angina.
- The study does not account for potential confounding variables that might affect both cholesterol levels and angina.

4.4.5 Conclusion

This study provides evidence that there is no significant relationship between cholesterol levels and the occurrence of exercise-induced angina in this sample of cardiac patients. Both the t-test and regression analysis support this conclusion, suggesting that cholesterol levels alone may not be a reliable predictor of exercise-induced angina.

These findings have important implications for understanding the factors contributing to exercise-induced angina. Healthcare professionals should consider that cholesterol levels may not be a primary determinant in the occurrence of exercise-induced angina and should look at other potential factors.

4.5 Analysis of the Association Between Gender and Heart Disease

This study aims to investigate the potential association between gender (sex) and the presence of heart disease (target) in a sample of patients. Understanding this relationship is crucial for identifying gender-specific risk factors and developing targeted prevention and treatment strategies in cardiovascular health.

4.5.1 Hypothesis

We formulated the following hypotheses:

Null Hypothesis (H0): There is no association between gender and heart disease. Alternative Hypothesis (H1): There is an association between gender and heart disease.

4.5.2 Methodology

To test our hypothesis and explore the relationship, we employed two statistical methods:

a) Chi-square test for independence: To assess the association between gender and heart disease.

- Chi-square test: This method was chosen because both variables (gender and heart disease) are categorical. The chi-square test is appropriate for determining whether there is a significant association between two categorical variables.

b) Logistic Regression: To model the relationship between gender and the likelihood of heart disease.

- Logistic Regression: This method was selected because we are predicting a binary outcome (presence or absence of heart disease) based on a categorical predictor (gender). Logistic regression is well-suited for modeling the probability of a binary outcome.

4.5.3 Results

4.5.3.1 Chi-square Test for Independence

Results: Chi-Square Statistic: 22.717227046576355 P-Value: 1.8767776216941503e-06

Interpretation:

- The extremely low p-value ($p < 0.05$) indicates strong evidence against the null hypothesis.

- We reject the null hypothesis, concluding that there is a significant association between gender and heart disease.

4.5.3.2 Logistic Regression Analysis

Key Statistics: Pseudo R-squared: 0.05948 Log-Likelihood: -196.40 LLR p-value: 6.226e-07

Coefficients: Intercept (const): 1.0986 (std err: 0.236, p-value: 0.000) Sex: -1.3022 (std err: 0.274, p-value: 0.000)

Interpretation:

a) Model Fit: The pseudo R-squared value of 0.05948 indicates that the model explains about 5.95% of the variance in heart disease occurrence.

b) Statistical Significance: The very low LLR p-value (6.226e-07) suggests that the model is statistically significant.

c) Coefficient Interpretation: The negative coefficient for sex (-1.3022) suggests that being female (assuming female is coded as 1) is associated with lower odds of heart disease compared to being male.

4.5.4 Discussion

4.5.4.1 Chi-square Analysis

The chi-square test provides strong evidence for an association between gender and heart disease. This suggests that the prevalence of heart disease differs significantly between males and females in our sample.

4.5.4.2 Logistic Regression Analysis

The logistic regression model, while statistically significant, explains only a small portion of the variance in heart disease occurrence (5.95%). This suggests that while gender is an important factor, other variables not included in this model also play significant roles in predicting heart disease.

The negative coefficient for sex in the logistic regression model indicates that females have lower odds of heart disease compared to males. This aligns with general epidemiological trends where males are often at higher risk for cardiovascular disease.

4.5.4.3 Limitations

- The chi-square test doesn't provide information about the nature or strength of the association.
- The logistic regression model's low pseudo R-squared value suggests that other important predictors of heart disease are not included in this analysis.
- The study doesn't account for potential confounding factors that might influence the relationship between gender and heart disease.

4.5.5 Conclusion

This study provides strong evidence for a significant association between gender and heart disease. Both the chi-square test and logistic regression analysis support this conclusion, with the regression model suggesting that females have lower odds of heart disease compared to males in this sample.

These findings have important implications for understanding gender disparities in cardiovascular health. Healthcare professionals should consider gender as a significant factor in assessing cardiovascular risk and developing prevention strategies.

4.6 Analysis of the Relationship Between Age and Resting Blood Pressure in Cardiac Patients

This study aims to investigate the potential relationship between age and resting blood pressure (trestbps) in a sample of cardiac patients. Understanding this relationship is crucial for assessing cardiovascular risk factors across different age groups and informing age-specific preventive strategies.

4.6.1 Hypothesis

We formulated the following hypotheses:

Null Hypothesis (H0): There is no significant correlation between age and resting blood pressure.
Alternative Hypothesis (H1): There is a significant correlation between age and resting blood pressure.

4.6.2 Methodology

To test our hypothesis and explore the relationship, we employed two statistical methods:

a) Pearson correlation coefficient: To assess the strength and direction of the linear relationship between age and resting blood pressure.

- Pearson Correlation: This method was chosen because both variables (age and resting blood pressure) are continuous. The Pearson correlation coefficient is appropriate for measuring the strength and direction of a linear relationship between two continuous variables.

b) Ordinary Least Squares (OLS) Regression: To model the relationship between resting blood pressure and age.

- OLS Regression: This method was selected to quantify the relationship between resting blood pressure and age, providing a predictive model and allowing us to estimate the change in age for each unit increase in resting blood pressure.

4.6.3 Results

4.6.3.1 Pearson Correlation Analysis

Results: Correlation Coefficient: 0.27935090656128836 P-Value: 7.762269074809911e-07

Interpretation:

The correlation coefficient of approximately 0.279 indicates a weak to moderate positive correlation between age and resting blood pressure.

The extremely low p-value ($p < 0.05$) suggests that this correlation is statistically significant.

We reject the null hypothesis, concluding that there is a significant correlation between age and resting blood pressure.

4.6.3.2 OLS Regression Analysis

Key Statistics: R-squared: 0.078 Adjusted R-squared: 0.075 F-statistic: 25.48 Prob (F-statistic): 7.76e-07

Coefficients: Intercept (const): 35.3255 (std err: 3.806, p-value: 0.000) Resting Blood Pressure (trestbps): 0.1447 (std err: 0.029, p-value: 0.000)

Interpretation:

a) Model Fit: The R-squared value of 0.078 indicates that approximately 7.8% of the variance in age can be explained by resting blood pressure.

b) Statistical Significance: The low p-value for the F-statistic (7.76e-07) confirms that the model is statistically significant.

c) Coefficient Interpretation: The positive coefficient for resting blood pressure (0.1447) suggests that for every 1 mmHg increase in resting blood pressure, we expect an increase of about 0.14 years in age, holding other factors constant. d) Intercept: The intercept of 35.3255 represents the expected age when resting blood pressure is theoretically zero.

4.6.4 Discussion

4.6.4.1 Correlation Analysis

The weak to moderate positive correlation (0.279) between age and resting blood pressure aligns with physiological expectations. As individuals age, there is a tendency for blood pressure to increase. This relationship is statistically significant, providing evidence against the null hypothesis.

4.6.4.2 Regression Analysis

The regression model, while statistically significant, explains only 7.8% of the variance in age. This suggests that while resting blood pressure is a significant predictor of age, other variables not included in this model also play important roles in determining age or influencing blood pressure.

The model predicts an increase of approximately 0.14 years in age for each 1 mmHg increase in resting blood pressure. This finding has practical implications for understanding the relationship between aging and blood pressure changes.

4.6.4.3 Limitations

- The model assumes a linear relationship between age and resting blood pressure, which may not capture more complex interactions.
- Other factors not included in this analysis (e.g., lifestyle, genetics) may influence both age and blood pressure.
- The study is limited to cardiac patients and may not be generalizable to the general population.
- The high condition number (1.01e+03) suggests potential multicollinearity or numerical problems, which could affect the reliability of the regression results.

4.6.5 Conclusion

This study provides evidence for a significant positive relationship between age and resting blood pressure in cardiac patients. Both the correlation and regression analyses support this conclusion, offering quantitative insights into how resting blood pressure changes with age in this population.

These findings have important implications for understanding cardiovascular risk factors across different age groups. Healthcare professionals should consider age as a significant factor when evaluating blood pressure measurements and developing age-appropriate interventions for cardiovascular health.

5. Extra Part

5.1. Visualization

In this part all visualization that exist in paper that implemented. All image from Fig7 to Fig 12 explained in Above Parts

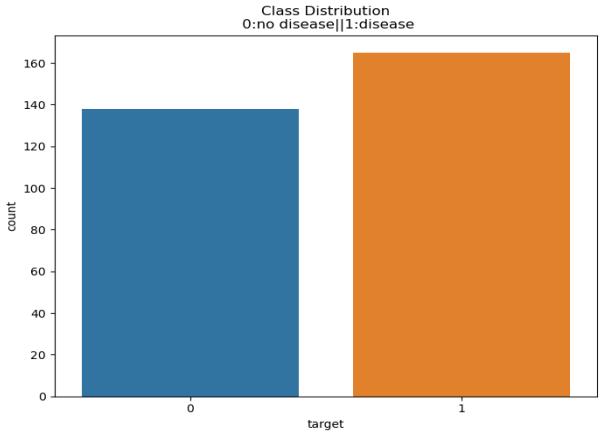

Fig7. Class distribution of disease and no disease

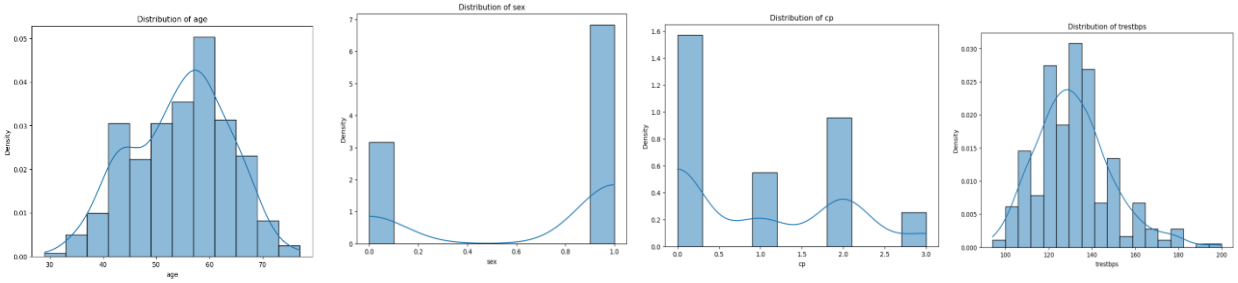

Fig8. Distribution of age, sex, chest pain and trestbps

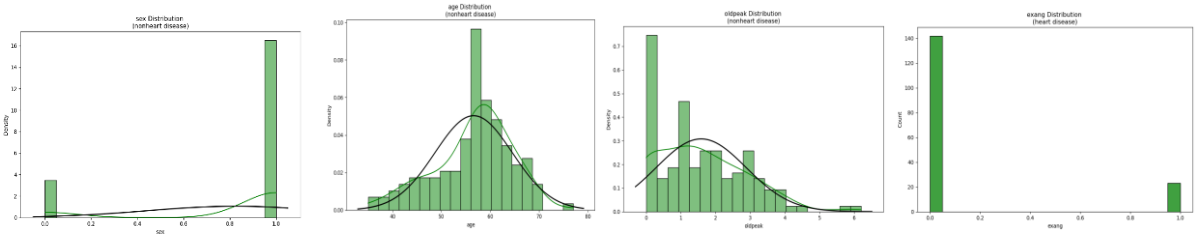

Fig9. Features important for heart disease

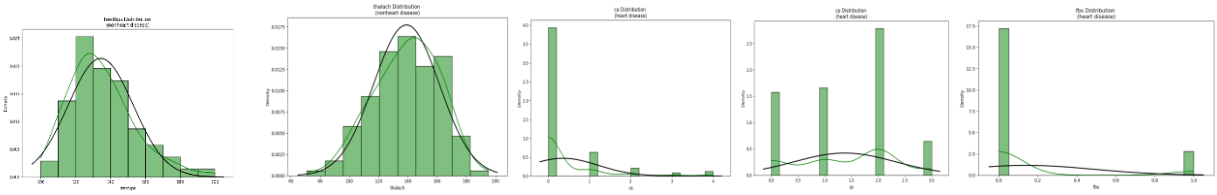

Fig10. Features not important for heart disease

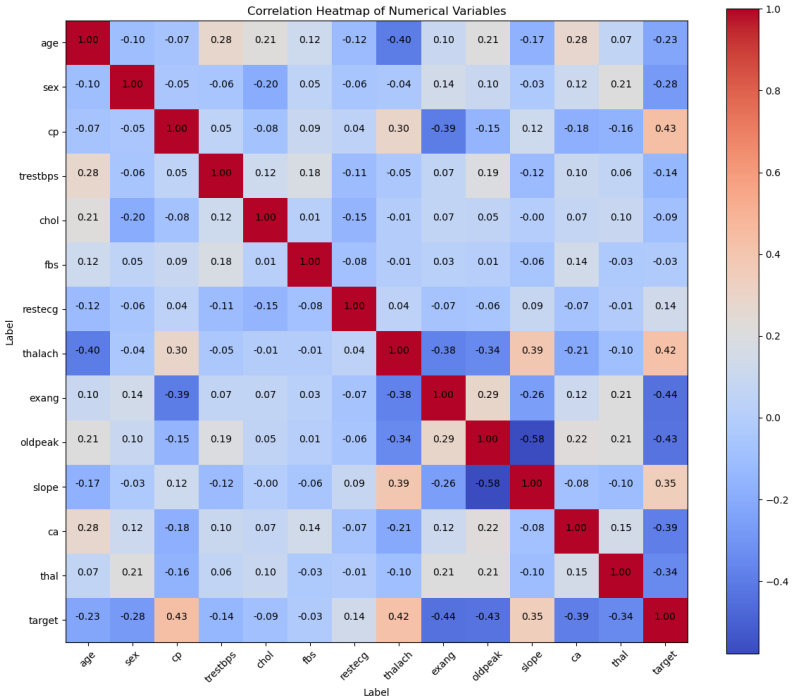

Fig11. Correlation heatmap

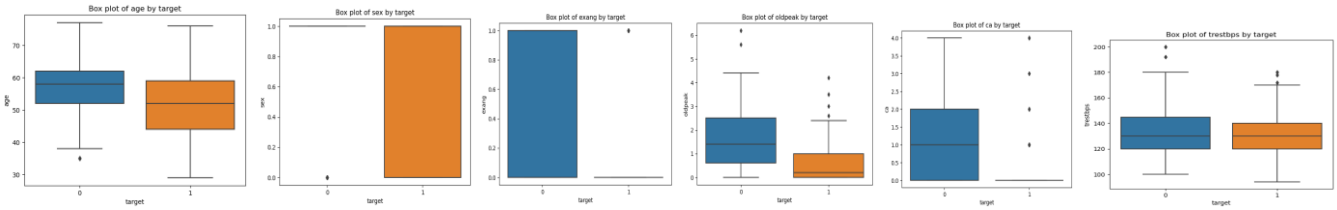

Fig12. Feature selection on correlation heatmap

5.2 Criticism of the article

The study presented in this article, while providing valuable insights into heart disease risk factors, demonstrates several limitations in its analytical approach and interpretation of results. These limitations potentially compromise the comprehensiveness and reliability of the conclusions drawn.

5.2.1. Inadequate Exploration of Variable Relationships

The study's approach to analyzing relationships between variables appears limited. **While the researchers conducted pairwise correlations between numerical variables for heart disease patients, they failed to explore these relationships in the context of the entire dataset, including both high-risk and low-risk groups.** This oversight potentially misses important insights about how these variables interact differently in individuals with and without heart disease.

For example, the results show significant correlations between age and other numerical variables (trestbps, chol, thalach, oldpeak) in heart disease patients. However, without a comparative analysis in the low-risk group, it's impossible to determine if these correlations are unique to heart disease patients or common across all individuals.

5.2.2 Lack of Multivariate Analysis

The study relies heavily on bivariate analyses, such as t-tests, chi-square tests, and one-way ANOVA. **While these tests provide valuable initial insights, they fail to capture the complex, multidimensional nature of heart disease risk.** The absence of multivariate techniques, such as multiple regression, logistic regression, or factor analysis, limits the study's ability to control for confounding variables and understand the relative importance of different risk factors.

5.2.3 Inadequate Consideration of Confounding Factors Limits Study's Insights into Heart Disease Risk

Another critical shortcoming is the lack of consideration for potential confounding factors. **The analysis does not adequately account for the interplay between different risk factors, particularly in cases where certain variables may mediate or moderate the effects of others on heart disease risk.** This omission limits the study's ability to provide a nuanced understanding of how various factors collectively contribute to cardiovascular health outcomes.

5.2.4 Inadequate Consideration of Confounding Factors in Heart Disease Risk Analysis

The article also falls short in its exploration of subgroup analyses. **There is insufficient attention paid to how risk factors may differentially affect various demographic or clinical subgroups within the study population.** This lack of granularity restricts the generalizability of the findings and their potential application in personalized risk assessment and management strategies.

5.2.5 Lack of Clinical Implications in Study Findings: Bridging the Gap Between Statistical Significance and Practical Utility

The article lacks a comprehensive discussion of the clinical implications of its findings. **The researchers do not adequately translate their statistical results into meaningful insights for clinical practice or public health interventions**. This gap between statistical significance and clinical relevance limits the practical utility of the study's outcomes.

5.3. Suggestion

Based on the results of the statistical analyses conducted, several important findings were not considered in the original paper. These findings warrant further investigation and could potentially enhance our understanding of heart disease risk factors. Here are the key suggestions for expanding the article based on these results:

5.3.1 Differential analysis of numerical variables between risk groups:

The t-test results indicate significant differences in age, trestbps (resting blood pressure), thalach (maximum heart rate achieved), and oldpeak (ST depression induced by exercise relative to rest) between high-risk and low-risk groups. **This finding was not considered in the original paper and suggests that these variables may be important predictors of heart disease risk.** Further analysis of these variables could provide valuable insights into risk stratification. Fig 13, 14 show this heatmap.

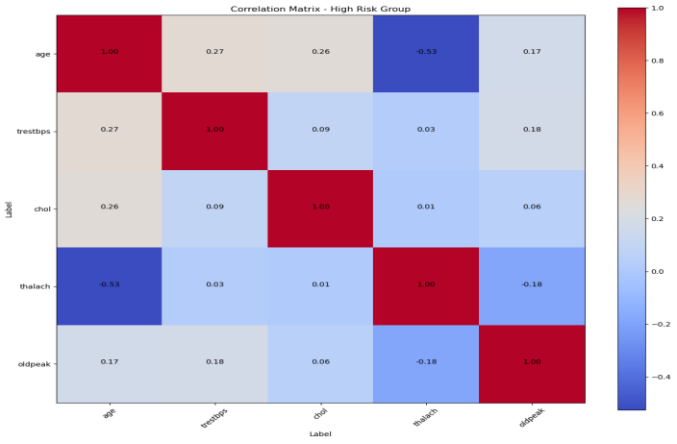

Fig13. numerical variables between high risk heatmap

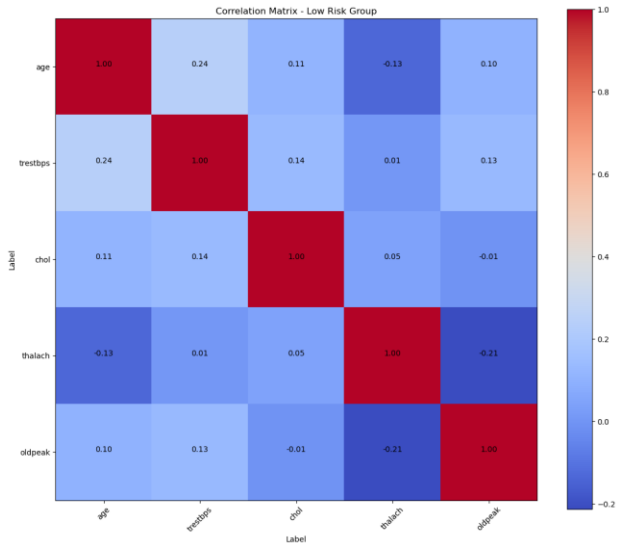

Fig14. numerical variables between low risk heatmap

5.3.2 Correlation analysis among numerical variables in heart disease patients:

The Pearson correlation analysis revealed significant correlations between several numerical variables in patients with heart disease. **This interrelationship between risk factors was not explored in the original paper.** Of particular interest are the correlations between age and all other numerical variables, as well as the correlation between trestbps and oldpeak. These findings suggest complex interactions between risk factors that merit further investigation.
We employed independent two-sample t-tests to compare five numerical variables between high-risk (target = 1) and low-risk (target = 0) groups. The variables analyzed were:

- Age
- Resting Blood Pressure (trestbps)
- Serum Cholesterol (chol)
- Maximum Heart Rate Achieved (thalach)
- ST Depression Induced by Exercise Relative to Rest (oldpeak)

The independent two-sample t-test was chosen because:

- We are comparing continuous variables between two independent groups.
- It allows us to determine if there is a statistically significant difference in the means of these variables between the high-risk and low-risk groups.
- It is robust and widely accepted for such comparisons in medical research.

This analysis reveals that four out of five examined variables (age, resting blood pressure, maximum heart rate, and ST depression) show significant differences between high-risk and low-risk groups for heart disease. These findings have several implications:

- Age and cardiovascular risk: The significant difference in age supports the well-established notion that cardiovascular risk increases with age.
- Blood pressure as a risk indicator: The difference in resting blood pressure underscores its importance as a modifiable risk factor for heart disease.
- Exercise response: Both maximum heart rate and ST depression during exercise show significant differences, highlighting the importance of exercise testing in risk assessment.
- Cholesterol paradox: The lack of significant difference in cholesterol levels is noteworthy and warrants further investigation. It may suggest that the relationship between cholesterol and heart disease risk is more complex than often assumed.

This study provides evidence that age, resting blood pressure, maximum heart rate achieved, and ST depression are significantly different between high-risk and low-risk groups for heart disease. Surprisingly, cholesterol levels did not show a significant difference. These findings can inform risk assessment strategies and highlight areas for targeted interventions in cardiovascular health management.

5.3.3. Association between categorical variables and heart disease:

Chi-square tests revealed significant associations between most categorical variables (sex, cp, restecg, exang, slope, thal, ca) and heart disease risk. **The strength and nature of these associations were not fully explored in the original paper.** A more detailed examination of these relationships could provide a more comprehensive understanding of categorical risk factors.

We employed the Chi-Square Test of Independence to analyze the relationship between each categorical variable and the target variable. This test is appropriate for categorical data and helps determine whether there is a significant association between two categorical variables.

The Chi-Square test evaluates the null hypothesis that there is no association between the variables. If the resulting p-value is less than the chosen significance level (typically 0.05), we reject the null hypothesis and conclude that there is a significant association.

For each categorical variable:

- We separated the data into high-risk and low-risk groups based on the target variable.
- We created a contingency table using value counts for each category within the variable.
- We performed the Chi-Square test.
- We evaluated the resulting p-value against our significance level of 0.05.
- Results

The Chi-Square test results indicate that seven out of the eight analyzed categorical variables show a significant association with the risk of heart disease. Only the fasting blood sugar (fbs) variable did not demonstrate a significant relationship.

These findings suggest that factors such as sex, chest pain type, resting ECG results, exercise-induced angina, the slope of the peak exercise ST segment, thallium stress test results, and the number of major vessels colored by fluoroscopy are potentially important indicators in assessing heart disease risk.

The lack of association with fasting blood sugar is somewhat surprising and may warrant further investigation. It's possible that this variable's relationship with heart disease risk is more complex or that it interacts with other factors not captured in this analysis.

This analysis has identified several categorical variables significantly associated with heart disease risk.

5.3.4 Analysis of numerical variables across categorical groups in high-risk patients

The ANOVA results indicate significant differences in certain numerical variables across different categorical groups within the high-risk population. **This nuanced analysis of risk factors within the high-risk group was not addressed in the original paper.** For example, age differed significantly among sex, fbs, restecg, and ca groups, while thalach showed significant differences across sex, cp, exang, slope, and thal groups. These findings suggest that the impact of numerical risk factors may vary depending on categorical variables, which could have important implications for personalized risk assessment.

We employed One-Way ANOVA to analyze the differences in numerical variables across different categories within the high-risk group. ANOVA is appropriate when comparing means of a continuous variable across three or more groups. It tests the null hypothesis that the means of all groups are equal.

For each combination of numerical and categorical variables:

- We isolated the high-risk group (target = 1).
- We grouped the numerical variable by the categories of the categorical variable.
- We performed One-Way ANOVA.
- We evaluated the resulting p-value against our significance level of 0.05.
- Results

The significant difference in cholesterol levels among different resting ECG result groups suggests a potential relationship between heart electrical activity at rest and cholesterol levels in high-risk patients. This could indicate that certain ECG patterns might be associated with different cholesterol profiles.

The maximum heart rate achieved showed significant differences across several categorical variables. This suggests that factors such as sex, chest pain type, exercise-induced angina, the slope of the ST segment, and thallium stress test results are associated with variations in maximum heart rate in high-risk patients. These findings could have implications for exercise stress testing and risk assessment.

ST depression showed significant differences among chest pain types and ST segment slope categories. This indicates that the nature of chest pain and the ST segment response during exercise are related to the degree of ST depression, which is an important indicator of myocardial ischemia.

5.3.5 Analysis of Logistic regression analysis comparing risk probabilities between men and women for different variables at specific thresholds

This part provides valuable insights into gender-specific risk factors. For example, the markedly higher risk for women compared to men for age above 50 years (82.78% vs 50.81%) and cholesterol above 170 mg/dL (84.84% vs 55.90%) suggests that these factors may be particularly important for assessing heart disease risk in women.
We used logistic regression, a statistical method for predicting a binary outcome (in this case, the presence or absence of heart disease) based on one or more predictor variables. Logistic regression is appropriate when the dependent variable is dichotomous and allows for the calculation of probabilities.

- Age The results suggest that at age 50, women have a higher probability of heart disease (82.78%) compared to men (50.81%). This significant difference highlights the importance of age as a risk factor, particularly for women.
- Resting Blood Pressure At a resting blood pressure of 120 mm Hg, women again show a higher risk (80.36%) compared to men (50.43%). This indicates that elevated blood pressure may be a more critical risk factor for women.
- Serum Cholesterol With a cholesterol level of 170 mg/dL, women exhibit a higher risk (84.84%) than men (55.90%). This suggests that even at relatively normal cholesterol levels, women may be at increased risk for heart disease.
- Maximum Heart Rate Interestingly, at a maximum heart rate of 90 bpm, both sexes show lower risk, but women (17.66%) still have a higher risk than men (4.74%). This could indicate that a low maximum heart rate is less predictive of heart disease risk compared to other factors.
- ST Depression For an ST depression of 2, women again show a higher risk (54.64%) compared to men (24.00%). This suggests that ST depression may be a more sensitive indicator of heart disease risk in women.