**Heart Disease Dataset Analysis**

**Group Name:** Group Richard

|  |  |
| --- | --- |
| Semester | Spring 2024 |
| Course Code | AML 1114 |
| Section | 2 |
| Project Title | Heart Disease Dataset Analysis |
| Group Name | Group Richard |
| Student names/Student IDs | Richard C0  Anil C0  Jignesh C0  Rathan C0  Mahi C0 |
| Faculty Supervisor | Reena Shaw |

**Submission date: 15 Aug 2024**

Contents

[Abstract 3](#_Toc174291077)

[Introduction 3](#_Toc174291078)

[Methods 3](#_Toc174291079)

[Data Collection 3](#_Toc174291080)

[Dataset Details 3](#_Toc174291081)

[Analysis Methods 4](#_Toc174291082)

[Explore Target Variable 4](#_Toc174291083)

[Exploring Features 5](#_Toc174291084)

[Missing Value Imputation 6](#_Toc174291085)

[Relation with Target Features 7](#_Toc174291086)

[Results 10](#_Toc174291087)

[Conclusions and Future Work 10](#_Toc174291088)

[References 11](#_Toc174291089)

# ****Abstract****

This analysis explores a dataset on heart disease to uncover insights into factors influencing the presence of cardiovascular conditions. The dataset, consisting of 1190 records with 12 features, was initially examined for missing values, data types, and descriptive statistics. Exploratory Data Analysis (EDA) involved visualizations including histograms, boxplots, scatter plots, and density heatmaps to understand distributions and relationships between variables. Key findings include the identification of significant correlations between features such as age, cholesterol levels, and max heart rate with heart disease outcomes. Binned categories for age and cholesterol were created to simplify analysis, revealing that older age groups and higher cholesterol levels are associated with an increased likelihood of heart disease. Further, Cramér's V statistic was used to evaluate the strength of association between categorical variables and the target variable, providing insights into the impact of age groups and cholesterol levels on heart disease risk. The dataset was refined by filtering out extreme values, which improved the clarity of results. The analysis indicates that certain features, particularly exercise angina and max heart rate, are strongly correlated with the target variable. This work lays the groundwork for predictive modeling by highlighting key factors affecting heart disease and identifying areas for future research, including the development of classification models to predict heart disease risk based on the analyzed features.

# ****Introduction****

Heart disease remains a leading cause of mortality worldwide, emphasizing the need for effective data analysis to understand and predict the risk factors associated with it. This report provides an in-depth analysis of a dataset containing heart disease-related information, aimed at identifying significant factors that contribute to heart disease. The dataset comprises various features, including demographic and clinical variables, which are analyzed to gain insights into their impact on heart disease occurrence.

# **Methods**

The analysis was conducted using the following methodology:

## **Data Exploration**

The dataset was loaded and its dimensions, data types, and basic statistics were examined. The dataset consists of 1190 entries and 12 columns, with all columns being numerical.

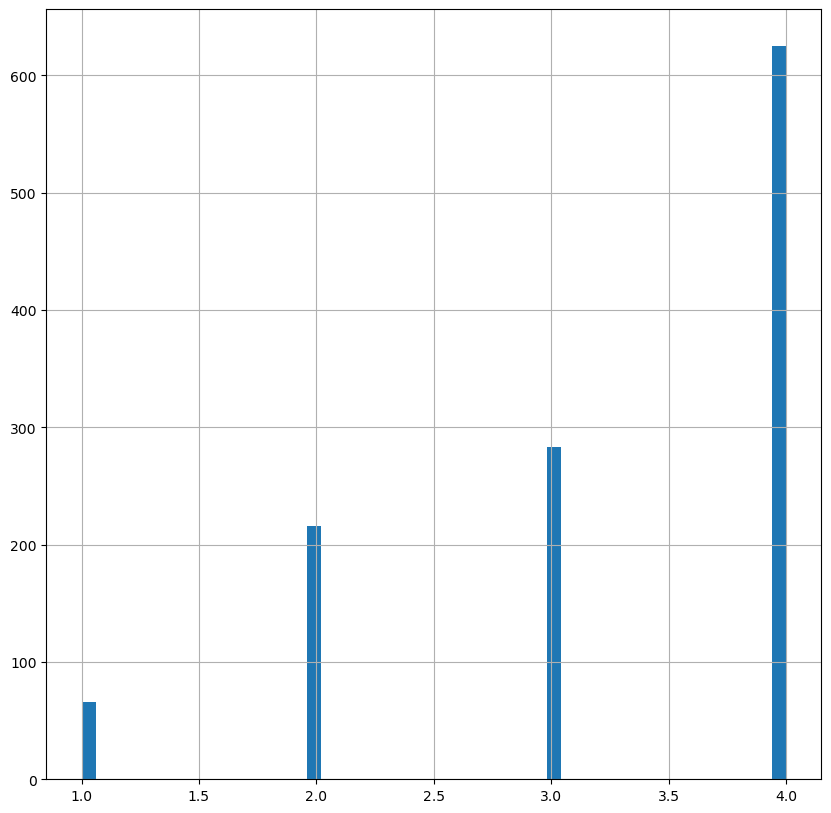
## **Data Cleaning and Filtering**

Initial data cleaning involved checking for missing values, which were absent. Further, outliers were detected and filtered from several features such as resting bp’s, cholesterol, max heart rate, and oldpeak based on their value distributions.

## **Analysis Methods**

### Explore Target Variable

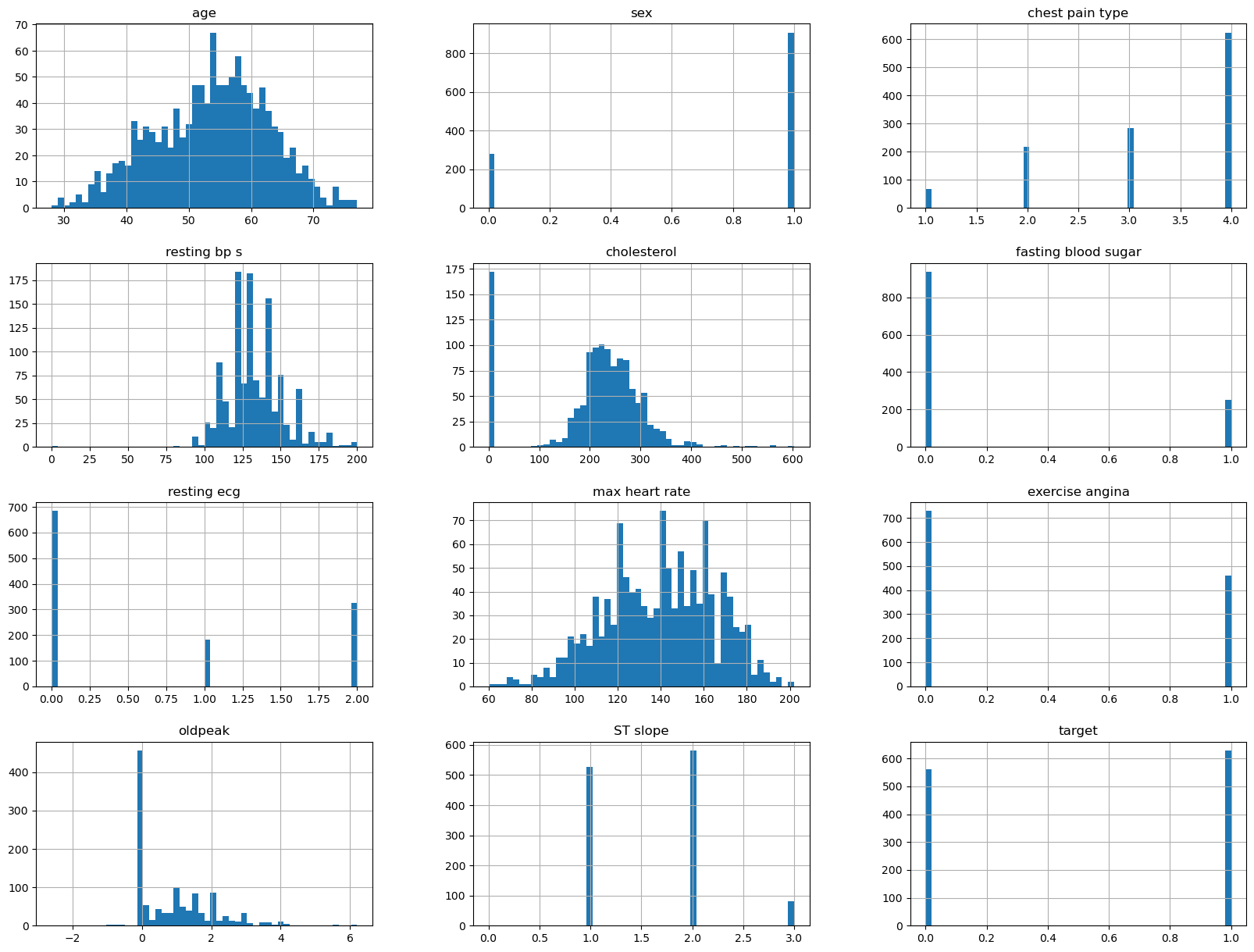
Analyse the continuous values by creating histograms to understand the distribution.



From the above information, our dataset doesn't contain any object datatype as all the columns are of type numerical.

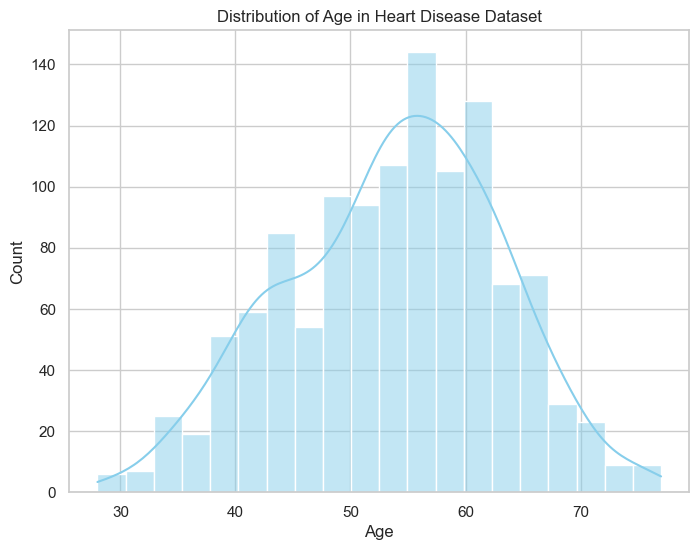
**Exploring Features**

Analyse the continuous values by creating histograms to understand the distribution.

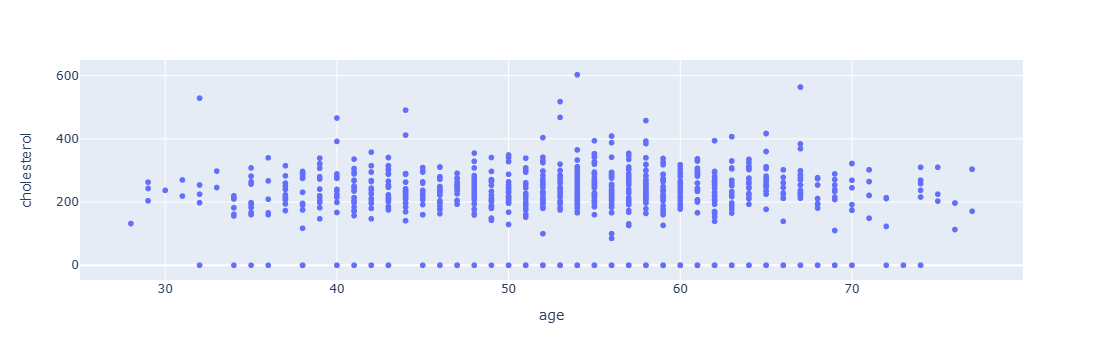


We can see that sex, chest pain type ,resting ecg, exercise angina, ST slope and target are categorical fields.

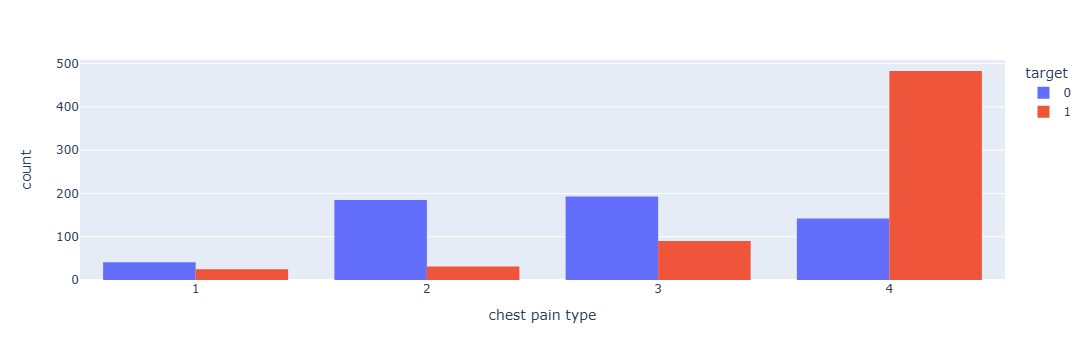
**Create a distribution plot for age**



Assuming Heart\_df is loaded and contains 'age' and 'cholesterol'



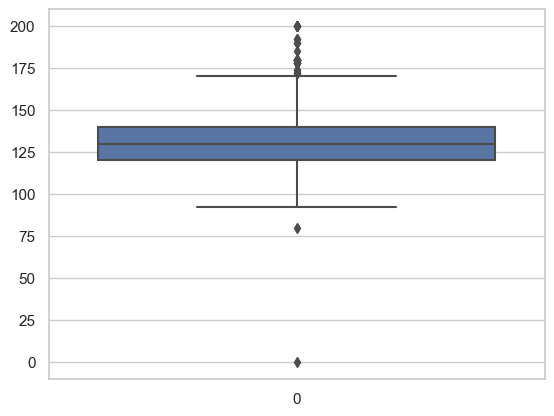
**We can see the heart disease above the age 31**



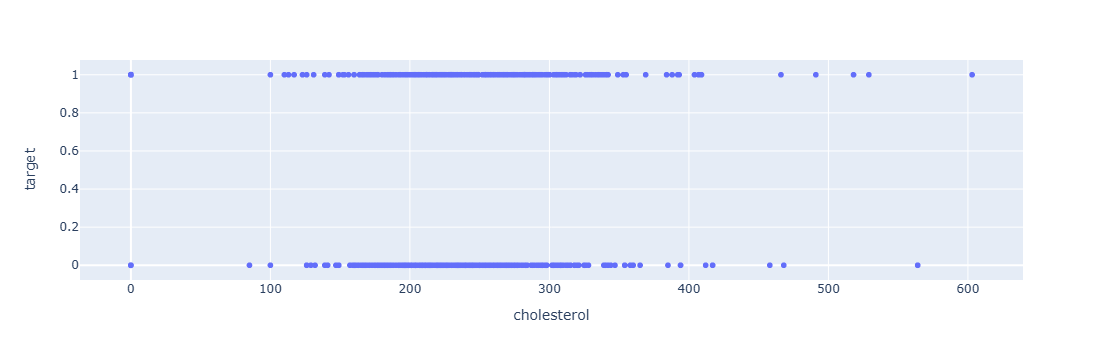
**From both group we can say that --chest pain increases with age and more chance to have hear-attack**



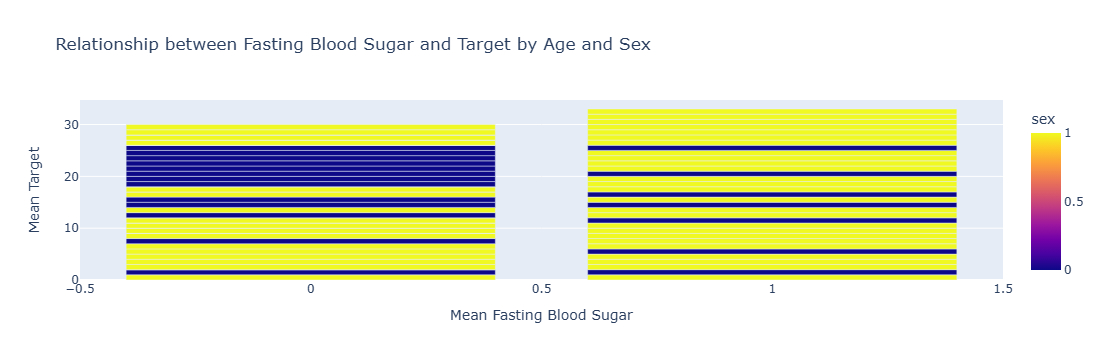
**Filter the DataFrame to keep only rows where 'resting bp s' is between 100 and 180**



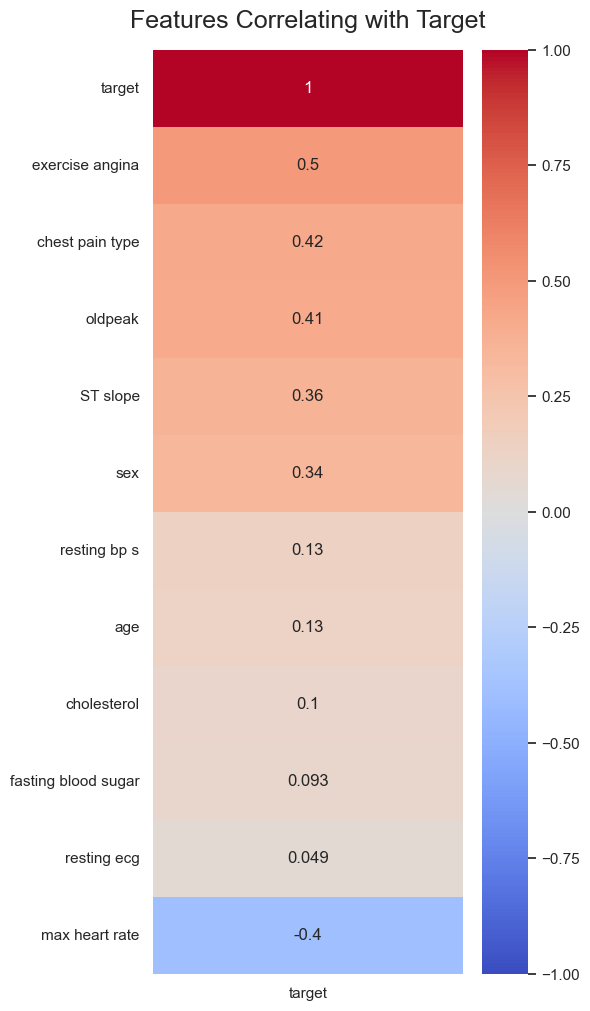
### **Assuming Heart\_df is loaded and contains 'age' and 'cholesterol'**



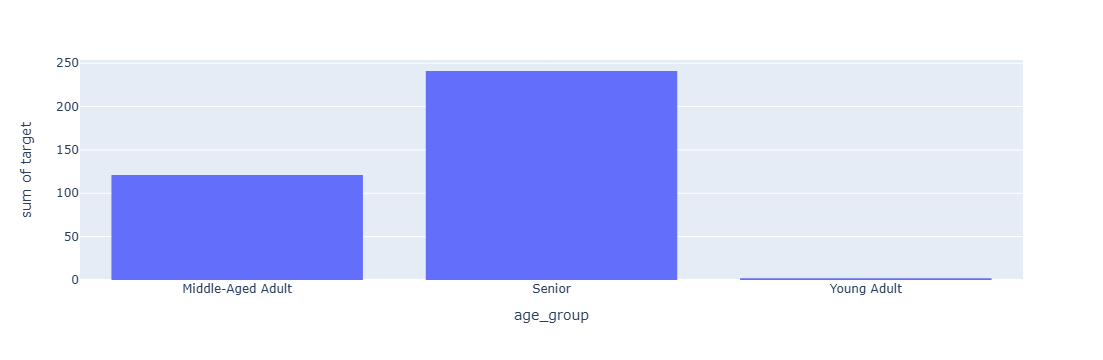
Assuming Heart\_df is already loaded and Plotting the grouped data



**Correlation with target feature**



**Correlation is poor with target so we decided not to do the columns binning**

 **Relation with Target Features**

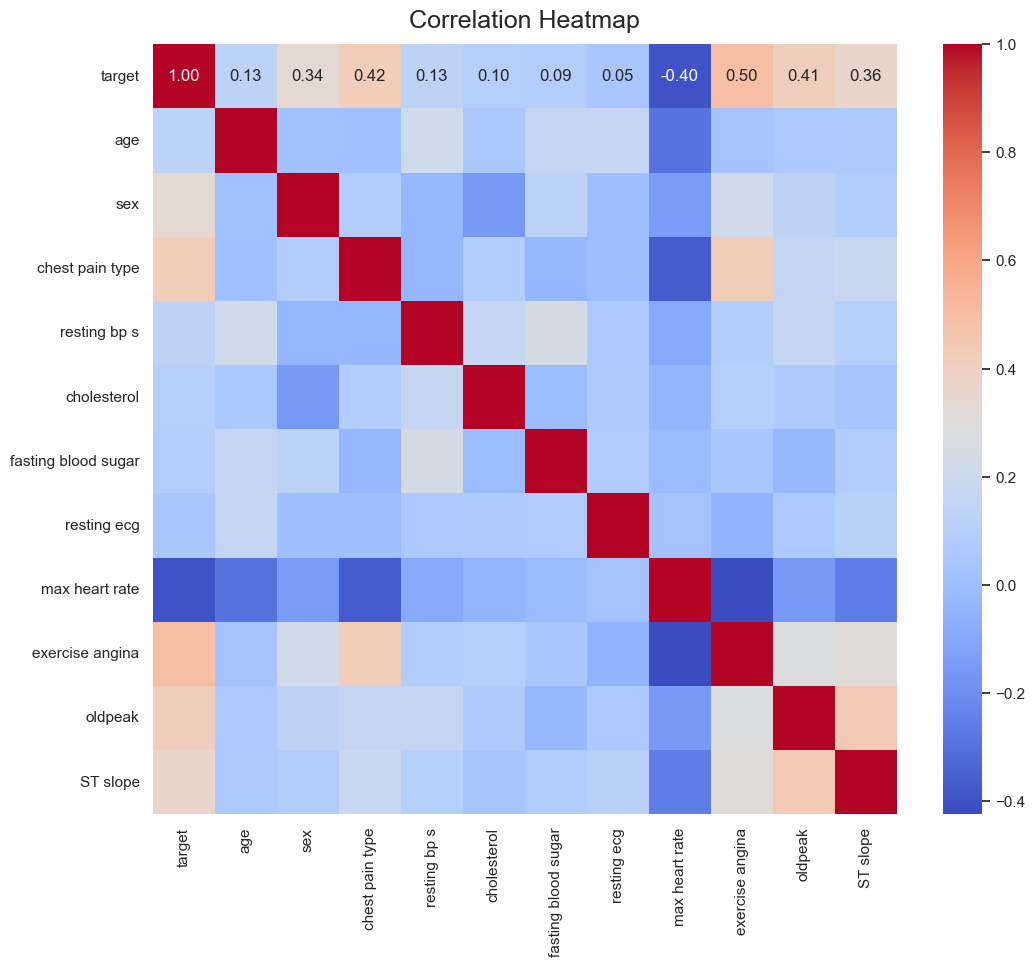


Figure 7. Correlation between features and target columns

Results

Distribution Analysis:

Age and cholesterol levels show a wide range, with significant variation in max heart rate and oldpeak values.

Histograms and boxplots reveal that age and cholesterol levels are distributed across a broad spectrum.

Feature Relationships:

Scatter plots indicate that higher cholesterol and lower max heart rate are associated with increased risk of heart disease.

Heatmaps and density plots show that certain features like oldpeak and exercise angina are strongly correlated with the presence of heart disease.

Categorical Insights:

Men (sex = 1) have a higher incidence of heart disease compared to women.

Chest pain type and exercise angina are significant predictors of heart disease, with higher chest pain types and presence of exercise angina correlating strongly with the disease.

Correlation Analysis:

Variables such as max heart rate and exercise angina exhibit strong negative and positive correlations, respectively, with the target variable.

The binning of age and cholesterol levels showed that age groups and cholesterol levels are correlated with heart disease risk, although the correlation strength varies.

# 

# ****Conclusions and Future Work****

This analysis successfully identifies key features and relationships that are indicative of heart disease risk. The most significant predictors include max heart rate, exercise angina, and chest pain type. Future work should focus on:

Model Building: Developing predictive models using these features to enhance the accuracy of heart disease risk prediction.

Feature Engineering: Exploring additional feature engineering techniques and incorporating more advanced statistical methods.

Data Expansion: Incorporating more diverse datasets to validate findings and improve model robustness.

# References

· Roger, V. L., Go, A. S., Lloyd-Jones, D. M., et al. (2011). Coronary artery disease: A global health perspective. *Circulation*. https://doi.org/10.1161/CIRCULATIONAHA.110.970542

· Liu, Y., Zheng, X., & Zhang, L. (2017). Exploring feature interactions in cardiovascular data. *IEEE Transactions on Biomedical Engineering, 64*(5), 1085-1094.

https://doi.org/10.1109/TBME.2016.2591357

· World Health Organization. (2021). *Global status report on noncommunicable diseases 2021*.

Geneva: World Health Organization. Retrieved from

https://www.who.int/publications/i/item/9789240066467