Introduction

In this report, our group embarked on an exploratory journey into a heart disease dataset, aiming to prepare it for data mining endeavors. We meticulously examined the dataset, assessing variables, data types, and values, while also scrutinizing for missing data, duplications, and outliers. Our efforts focused on ensuring data cleanliness and reliability, crucial for deriving meaningful insights. Through calculated summary statistics and informative visualizations, we sought to understand the dataset's nuances and unearth potential trends. Our findings not only shed light on the dataset's characteristics but also provided a foundation for future data mining endeavors.

EDA

Dataset Introduction

Dataset Overview: The dataset contains 11 columns and 2943 rows. It consists of records of symptoms and whether they turned out to be indicative of heart disease or not. Data is from Kaggle originated from UCI health data repository.

Dependent Variable: Heart disease (0 or 1)
Independent Variables: Age, Sex, Chest Pain, Resting Blood Pressure, Cholesterol, Blood Pressure, Resting Electrocardiogram (ECG), Heart Rate, Angina, ST Depression Induced by Exercise Relative to Rest.

Description of Variables

Variable Name	Variable Description
HeartDisease (Target Variable)	Does patient have heart disease
Age	Age of individual that ranges between 20 to 80
Sex	This is the gender of the individual. It is represented as a binary value where 1 stands for male and 0 stands for female.
ChestPainType	This categorizes the type of chest pain experienced by the individual • Value 1: Typical angina, which is chest pain related to the heart. • Value 2: Atypical angina, which is chest pain not related to the heart. • Value 3: Non-anginal pain, which is typically sharp and non-continuous. • Value 4: Asymptomatic, meaning the individual experiences no symptoms.
RestingBP	This is the individual's resting blood pressure (in mm Hg) when they are at rest.

Cholesterol	This is the individual's cholesterol level, measured in mg/dl.
FastingBS	This indicates whether the individual's fasting blood sugar is greater than 120 mg/dl. It is represented as a binary value where 1 stands for true and 0 stands for false.
RestingECG	
MaxHR	This is the maximum heart rate achieved by the individual.
ExerciseAngina	This indicates whether the individual experiences angina (chest pain) induced by exercise. It is represented as a binary value where 1 stands for yes and 0 stands for no.

Table 1. Variable Description

Data Cleaning

We checked the dataset for null values and duplicates, but the data has 0 null values and no duplicates. Hence, we processed to further analysis of variables.

Exploration

We began our exploratory data analysis (EDA) by examining the distribution of our data in terms of the frequency of patients with heart disease and without heart disease. We found that we have almost similar occurrences, in fact, 55% of the data pertains to patients with heart disease and 45% to their counterparts. We also investigated the data in terms of sex. Surprisingly, there are only 25% females and 75% males in our dataset.

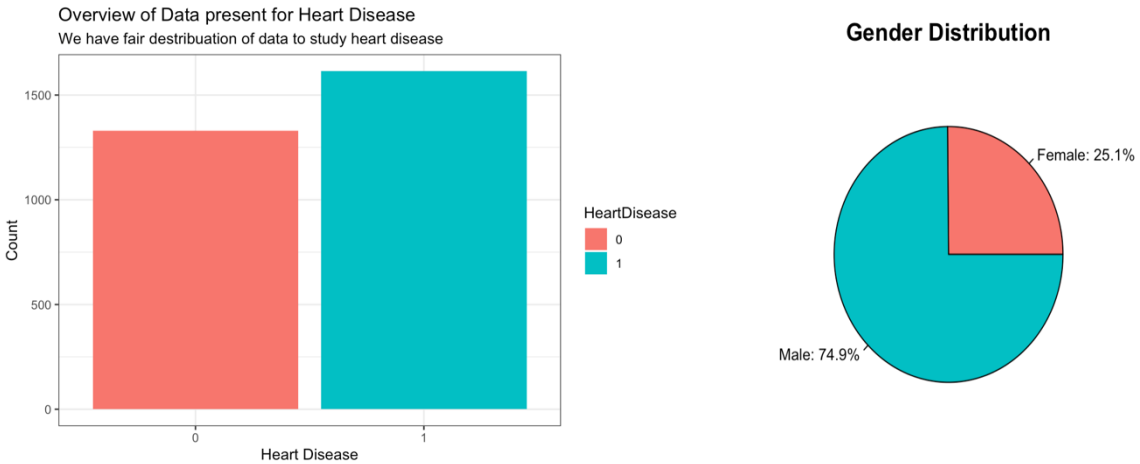

Fig. 1 Bar plot and Pie chart for exploration

Following this, we plotted histograms for age to verify if our data is normally distributed. The results of the graphs showed a normal distribution for both males and females. As previously discovered, the frequency of males is relatively high compared to females, but both genders follow a normal distribution in terms of age.

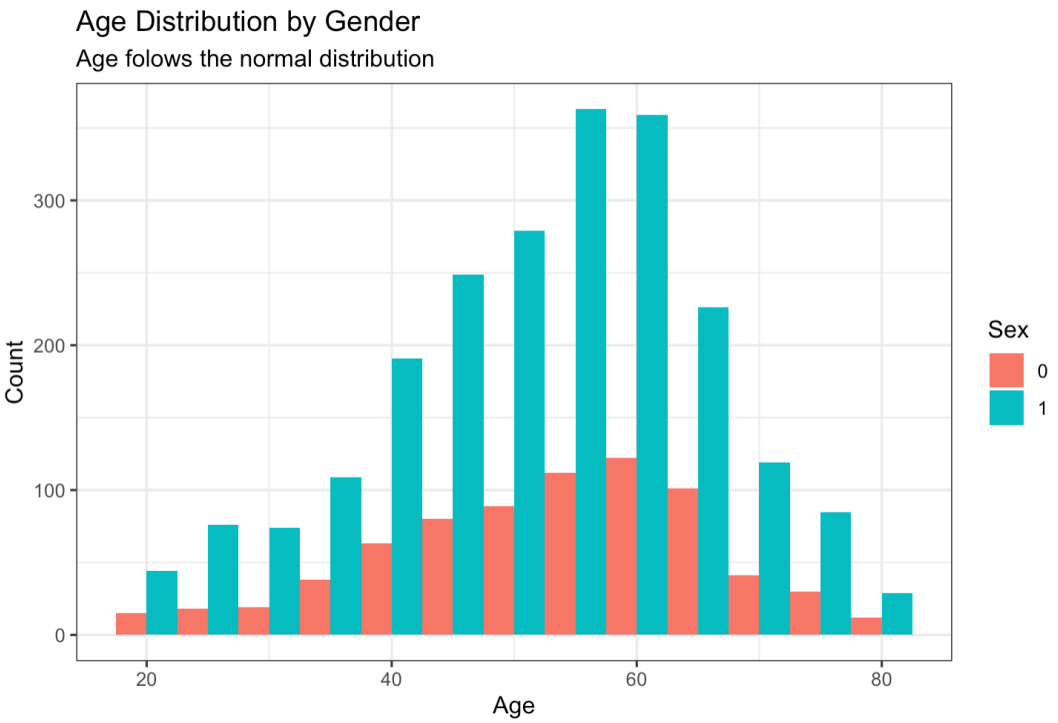

Fig. 2: Histogram of age.

Following this, we investigated age. Firstly, we converted age into different age groups with a span of 10 years, starting from 20 years. We calculated the total instances present in each age group and then found the percentage for each group regarding the occurrence of heart disease. The results confirm that the data is consistently spread across all groups because the percentage value for each group is around 55%, which is the same as the overall dataset distribution for the occurrence of heart disease.

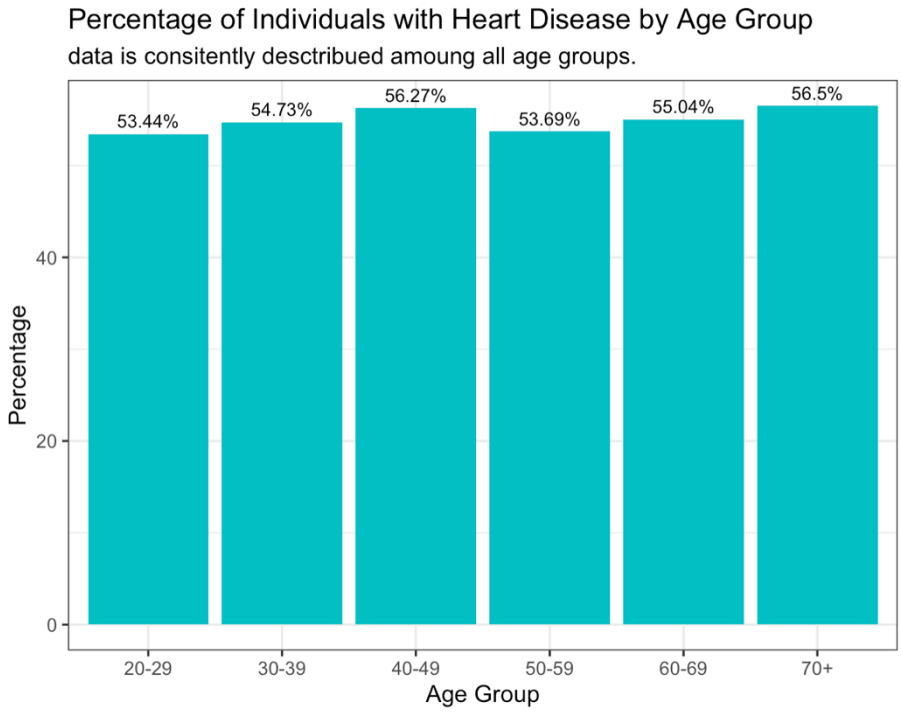

Fig. 3: Distribution of heart disease observation among different age group

In the next plot we draw density plot for cholesterol level for heart disease, which reveals there are two peaks on the graph, one on the left and one on the right. This suggests that there is a correlation between cholesterol levels and heart disease. People with high cholesterol levels (on the right side of the graph) are slightly more likely to have heart disease than people with low cholesterol levels (on the left side of the graph).

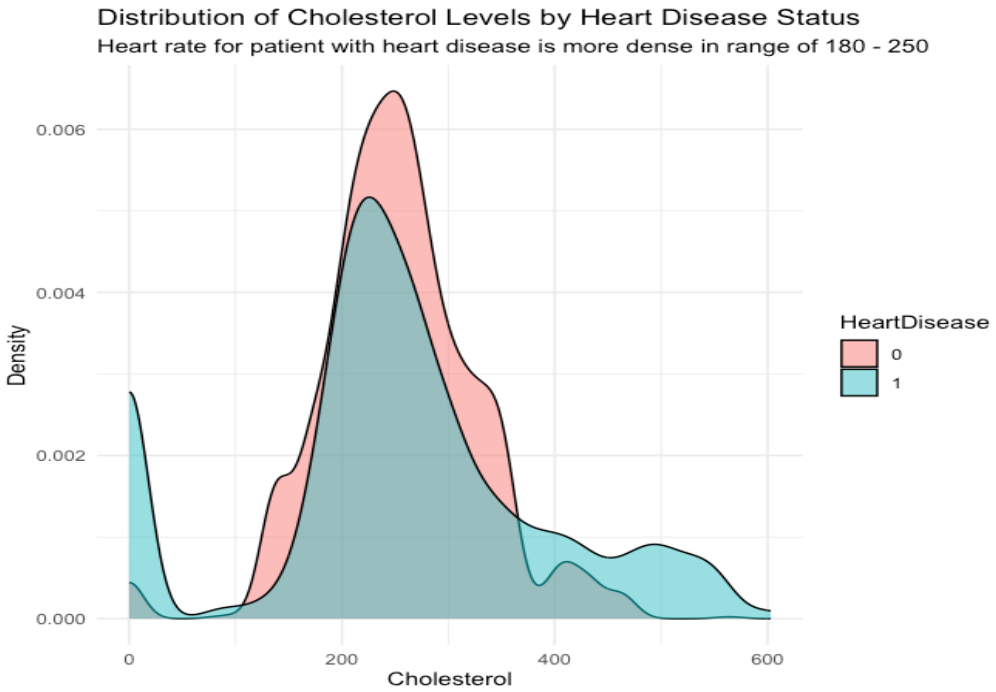

Fig. 4: Density plot of cholesterol level

In the next step, we plotted a correlation chart to identify potential relationships between features. The correlation plot revealed a strong positive relation between Fasting BS, Chest Pain, Maximum heart rate, and resting Blood pressure. This result can provide valuable hints for further analysis and findings. Additionally, we encountered a few interesting relationships, such as high heart rate with angina and age, and the relationship of cholesterol with chest pain and sex. Investigating these relationships further may reveal potentially life-saving findings.

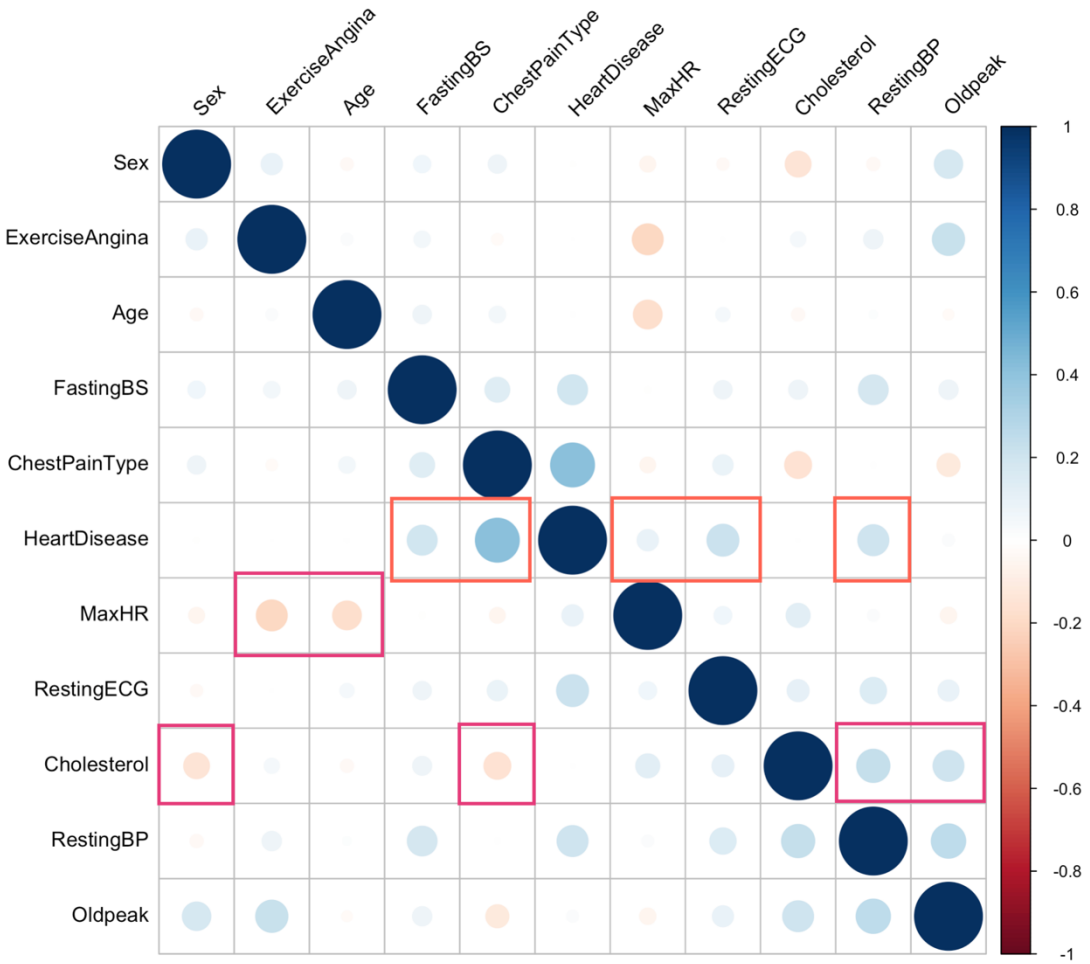

Fig 5: correlation plot

Conclusion

In our report, we thoroughly explored a heart disease dataset to prepare it for subsequent data mining endeavors. We carefully examined the dataset's variables, data types, and values, finding no null values or duplicates. Our exploratory data analysis (EDA) revealed similar frequencies of heart disease occurrences among patients and a notable gender skew towards males. We observed normal age distributions and consistent heart disease occurrences across different age groups. Density plots suggested a correlation between cholesterol levels and heart disease. Through correlation analysis, we identified strong positive relationships between various features, providing valuable insights for future analyses. These findings establish a solid foundation for further exploration, potentially leading to critical insights in heart disease diagnosis and treatment.

Reference:

Heart Disease Data Compiled from UCI. (n.d.). Www.kaggle.com. Retrieved April 25, 2024, from https://www.kaggle.com/datasets/rcratos/heart-disease-data-compiled-from-uci