Exploratory Data Analysis On Heart Diseases Datase

Introduction

In this study, data obtained from this dataset https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data is used to make statistically significant inferences about individuals and their history of heart disease. The dataset contains 17 different attributes, which will be explored and analyzed to gain insights into the factors influencing heart disease.

These attributes describe various features used in diagnosing heart disease. Below is an explanation of each attribute:

- **id**: A unique identifier for each patient. It is distinct and non-repetitive for every patient.
- **age**: The age of the patient, expressed in years.
- **origin**: The location or site of the study. This might refer to the hospital or medical center where the study was conducted.
- **sex**: The gender of the patient, with possible values being "Male" or "Female."
- **cp (chest pain type)**: Describes the type of chest pain the patient experiences, with four possible types:
    - **typical angina**: Chest pain typically related to heart disease, often triggered by physical activity.
    - **atypical angina**: Chest pain that differs from typical angina symptoms.
    - **non-anginal pain**: Chest pain that is likely unrelated to the heart.
    - **asymptomatic**: No symptoms are present.
- **trestbps (resting blood pressure)**: The patient's blood pressure at rest, measured in millimeters of mercury (mm Hg) upon admission to the hospital.
- **chol (serum cholesterol)**: The patient's total cholesterol level in the blood, measured in milligrams per deciliter (mg/dl).
- **fbs (fasting blood sugar)**: Indicates whether the patient's fasting blood sugar is greater than 120 mg/dl. The value is "True" if the fasting blood sugar is above 120 mg/dl, and "False" if it is below.
- **restecg (resting electrocardiographic results)**: Results of the patient's resting electrocardiogram (ECG), with possible values:
    - **normal**: A normal ECG result.
    - **stt abnormality**: ST-T wave abnormalities, which could indicate heart disease or injury.
    - **lv hypertrophy**: Left ventricular hypertrophy, indicating thickening of the heart muscle.
- **thalach**: The maximum heart rate achieved by the patient during exercise, measured in beats per minute (BPM).
- **exang (exercise-induced angina)**: Indicates whether the patient experienced angina (chest pain) induced by exercise. "True" indicates the presence of exercise-induced angina, while "False" indicates its absence.
- **oldpeak**: The ST segment depression induced by exercise, measured as the difference from the resting state. This value is typically in millivolts (mV).
- **slope**: The slope of the peak exercise ST segment. The type of slope can provide information about heart conditions.
- **ca**: The number of major coronary vessels (ranging from 0 to 3) colored by fluoroscopy.
- **thal**: The patient's thalassemia status, with possible values:
    - **normal**: No thalassemia.
    - **fixed defect**: A permanent defect in the heart muscle.
    - **reversible defect**: A temporary or reversible defect in the heart muscle.
- **num**: The predicted attribute, typically indicating the presence or absence of heart disease. It also indicate the severity of the disease.

Analysis stages

Raw data were subjected to exploratory analysis, revealing a total of 920 entries, with varying amounts of missing data in each attribute (Figure 1). Additionally, the results of other descriptive analyses are shown in Table 2. The average age of individuals in the data is 53 (+/- 9), with a blood pressure of 132 (+/- 19), cholesterol level of 199 (+/- 110), maximum heart rate of 137 (+/- 25), and the number of major vessels observed via fluoroscopy ranges from 0 to 3, with an average of 0.67. When examining the raw data overall, the number of diseases most frequently ranges from 0 to 4, with an average value of 0.99

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 920 entries, 0 to 919
Data columns (total 16 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   id        920 non-null    int64
 1   age       920 non-null    int64
 2   sex       920 non-null    object
 3   dataset   920 non-null    object
 4   cp        920 non-null    object
 5   trestbps  861 non-null    float64
 6   chol      890 non-null    float64
 7   fbs       830 non-null    object
 8   restecg   918 non-null    object
 9   thalch    865 non-null    float64
 10  exang     865 non-null    object
 11  oldpeak   858 non-null    float64
 12  slope     611 non-null    object
 13  ca        309 non-null    float64
 14  thal      434 non-null    object
 15  num       920 non-null    int64
dtypes: float64(5), int64(3), object(8)
memory usage: 115.1+ KB
```

Table 1. Descriptive Analytics of Raw numeric data

	id	age	trestbps	chol	thalch	oldpeak	ca	num
count	920.000000	920.000000	861.000000	890.000000	865.000000	858.000000	309.000000	920.000000
mean	460.500000	53.510870	132.132404	199.130337	137.545665	0.878788	0.676375	0.995652
std	265.725422	9.424685	19.066070	110.780810	25.926276	1.091226	0.935653	1.142693
min	1.000000	28.000000	0.000000	0.000000	60.000000	-2.600000	0.000000	0.000000
25%	230.750000	47.000000	120.000000	175.000000	120.000000	0.000000	0.000000	0.000000
50%	460.500000	54.000000	130.000000	223.000000	140.000000	0.500000	0.000000	1.000000
75%	690.250000	60.000000	140.000000	268.000000	157.000000	1.500000	1.000000	2.000000
max	920.000000	77.000000	200.000000	603.000000	202.000000	6.200000	3.000000	4.000000

Figure 1 Raw Descriptive analytics

To detect outliers that could potentially affect statistical calculations, a box plot was used, and potential outliers are shown as circles in Figure 2. As a result, potential outliers for the variables trestbps, chol, thalch, and oldpeak are shown as circles in the box plot

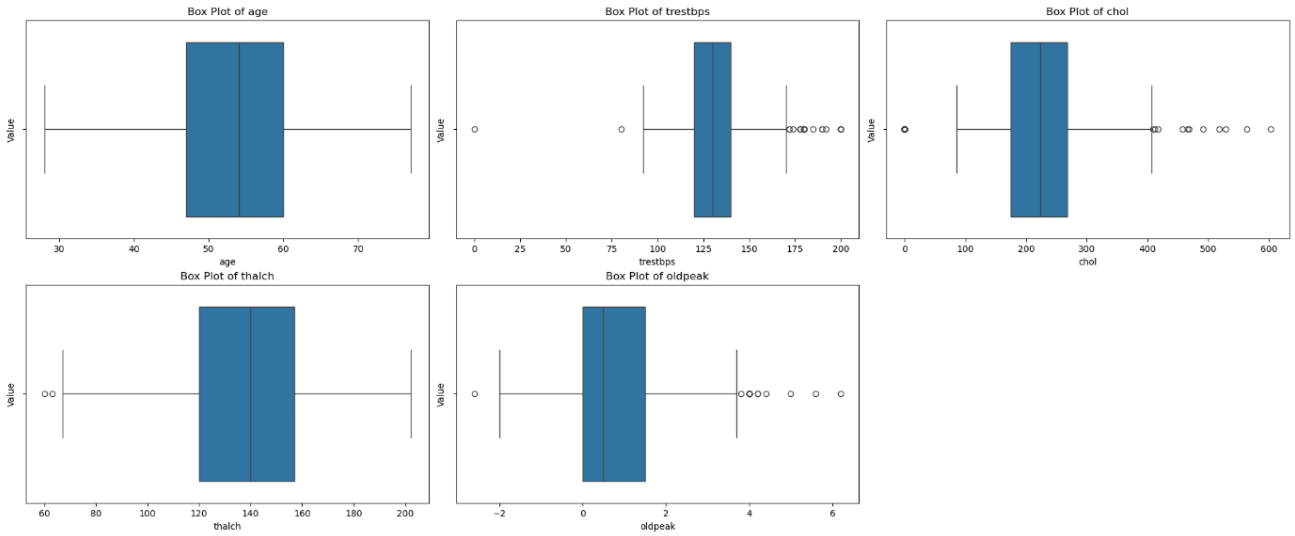

Figure 2 Box Plot for Outlier detection

Subsequently, outlier values were removed using Python, resulting in the box plot data shown in Figure 3. This process removed many of the outliers

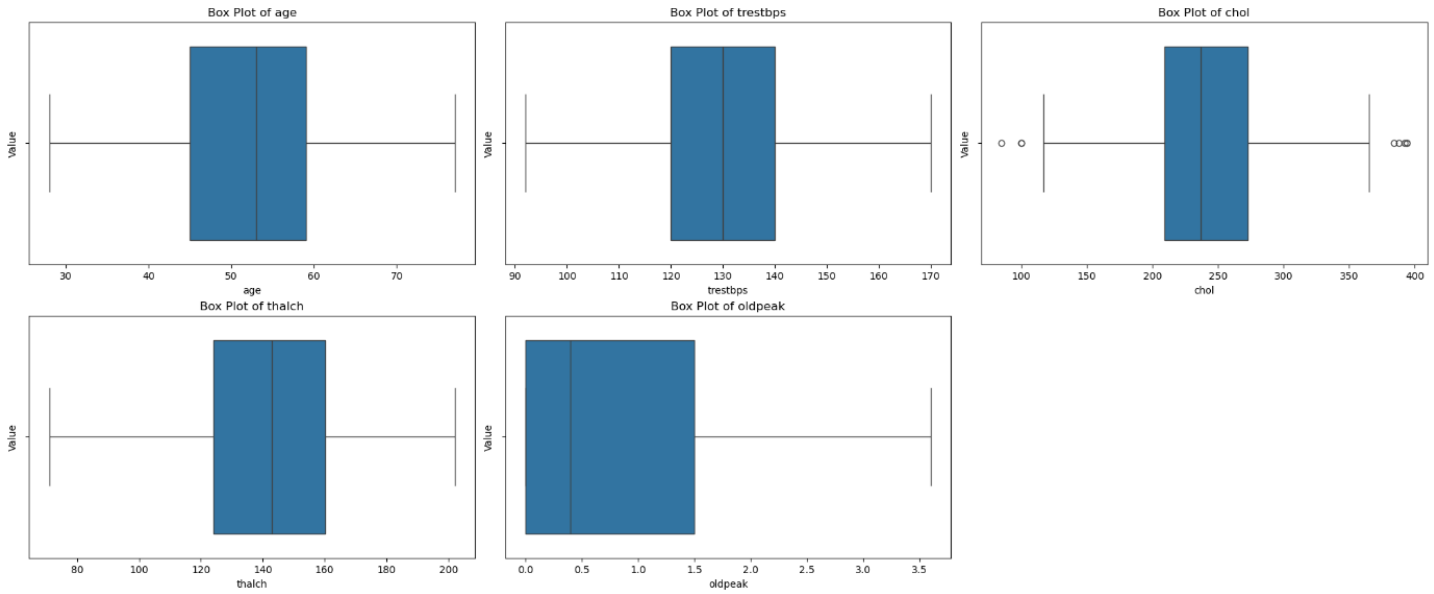

Figure 3. Boxplot After Outliers

Subsequently, histograms and a Python code I wrote to return the skewness values of the data were used to check whether the data follows a normal distribution. The histogram of the data is shown in Figure 4. For the columns that may pose a problem in statistical analysis—'age', 'trestbps', 'chol', 'thalch', 'oldpeak'—the skewness values indicating whether they adhere to a normal distribution are provided below. (Note : The range of skewness for a fairly symmetrical bell curve distribution is between -0.5 and 0.5; moderate skewness is -0.5 to -1.0 and 0.5 to 1.0; and highly skewed distribution is < -1.0 and > 1.0. In our case, we have ~1.7, so it is considered highly skewed data)

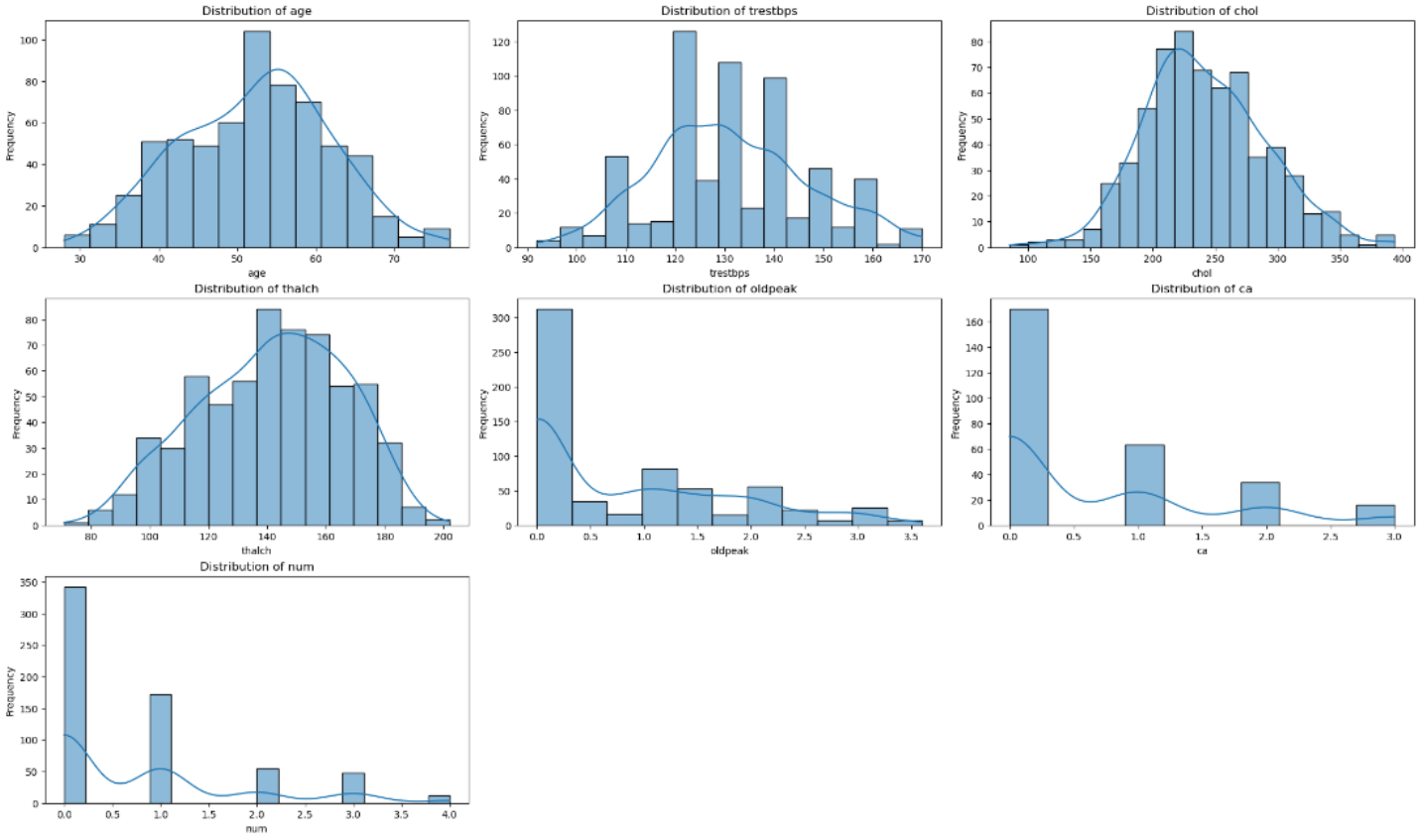

Figure 4. Histogram Plot of Numeric Columns

- age: Skewness = -0.06, Type = Fairly Symmetrical (-0.5 to 0.5): Data distribution is close to symmetrical, but might have slight deviations.
- trestbps: Skewness = 0.27, Type = Fairly Symmetrical (-0.5 to 0.5): Data distribution is close to symmetrical, but might have slight deviations.
- chol: Skewness = 0.26, Type = Fairly Symmetrical (-0.5 to 0.5): Data distribution is close to symmetrical, but might have slight deviations.
- thalch: Skewness = -0.25, Type = Fairly Symmetrical (-0.5 to 0.5): Data distribution is close to symmetrical, but might have slight deviations.
- oldpeak: Skewness = 0.92, Type = **Moderate Positive Skew** (0.5 to 1.0): Data has a noticeable right tail, but not extreme.

Following this, for the 'oldpeak' column, which exhibited **moderate positive skewness**, a **log transformation** was applied as a feature engineering technique. However, the skewness value remained at Skewness = 0.92. Later, a **square root transformation** was performed using the sqrt_transformed command, and the skewness value of the attribute decreased to Skewness = -0.27, Type = **Fairly Symmetrical**. This adjustment improved the data's adherence to a normal distribution, making it more suitable for our analysis (figure 5)

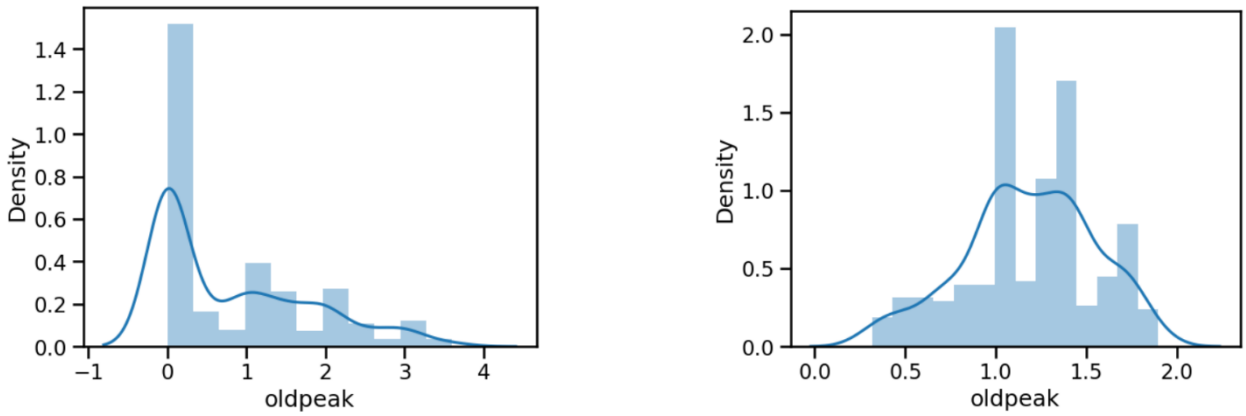

Figure 5. The data distribution of the oldpeak column

Before exploring the relationships between features, one-hot encoding was applied to the non-numeric data in our raw dataset, converting these categorical variables into dummy variables. This transformation allowed us to represent categorical features numerically. To visualize the relationships among the resulting variables, a correlation heatmap was generated using the Seaborn library. This heatmap clearly displays the correlation values between the features, providing a comprehensive overview of their interdependencies (Figure 6).

In preparation for the next step, we classified the obtained correlation values based on their absolute values. Specifically, if the value was less than 0.1, it was categorized as 'Very Weak'; if less than 0.3, as 'Weak'; if less than 0.5, as 'Moderate'; if less than 0.7, as 'Strong'; and all remaining values were categorized as 'Very Strong.' While no 'Strong' or 'Very Strong' correlations were observed, the other correlations and their classifications are as follows

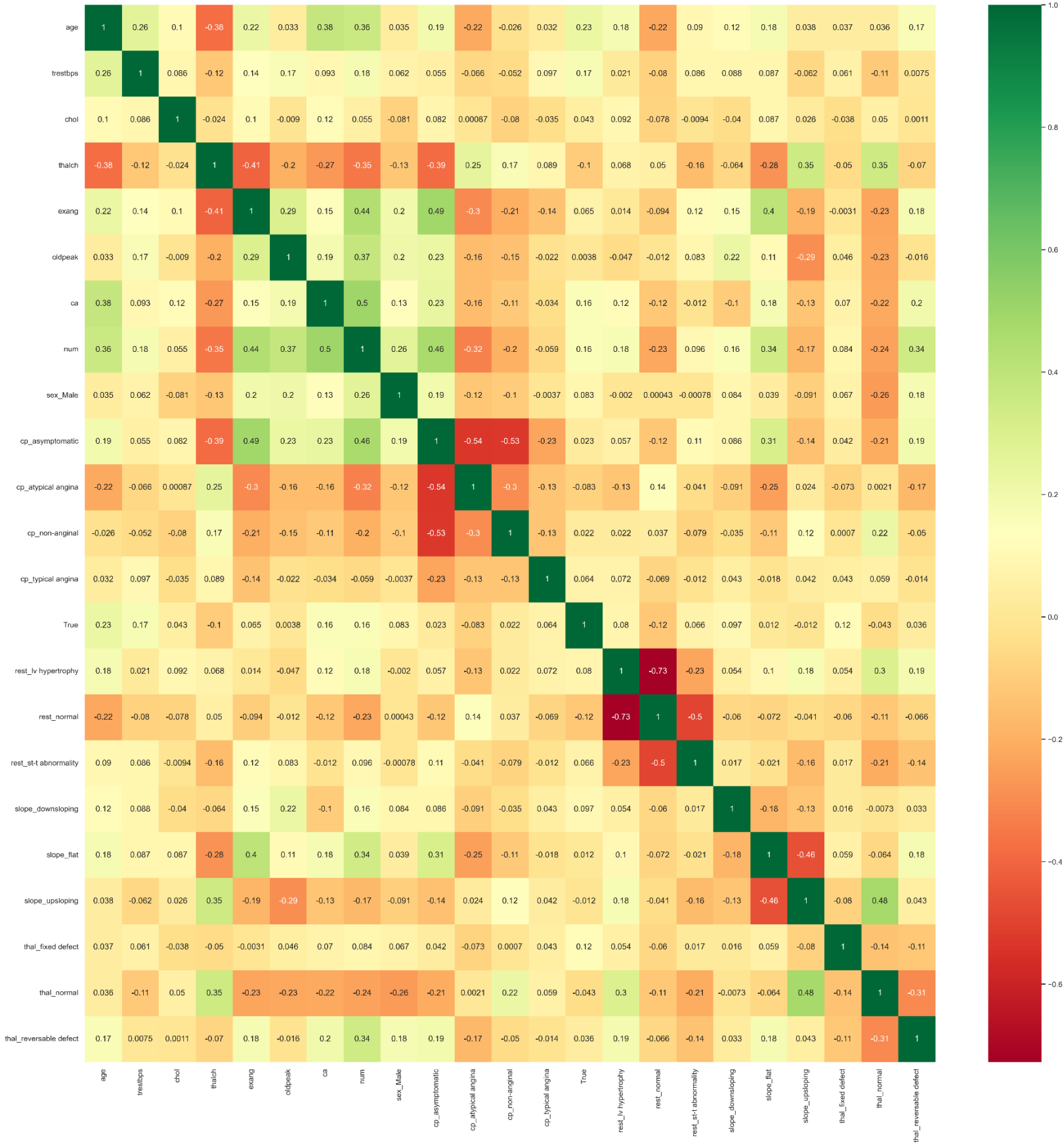

Figure 6. Correlation and Heatmap of the Complete

Table 2. Classified Correlation Values

Very Weak	Weak	Moderate
chol: 0.05	trestbps: 0.18	age: 0.36
cp_typical angina: -0.06	sex_Male: 0.26	thalch: -0.35
rest_st-t abnormality: 0.10	cp_non-anginal: -0.20	exang: 0.44
thal_fixed defect: 0.08	rest_lv hypertrophy: 0.18	oldpeak: 0.37
	rest_normal: -0.23	ca: 0.50
	slope_downsloping: 0.16	cp_asymptomatic: 0.46
	slope_upsloping: -0.17	cp_atypical angina: -0.32
	thal_normal: -0.24	slope_flat: 0.34
		thal_reversable defect: 0.34

In order to determine the significance of the obtained correlation values, a t-test was conducted with a p-value threshold of 0.05 (values less than 0.05 are considered significant). The results are shown in the table below. Here, the H1 hypothesis is the alternative hypothesis, indicating that there is a significant relationship between the two variables, while the H0 hypothesis is the null hypothesis, suggesting that there is no significant relationship between the variables. In Table 3, the features identified as significant are highlighted in green, whereas those considered non-significant are marked in red. Additionally, box plots of the significant variables based on disease severity are shown in Figure 7.

Table 3. Significance of Correlation Values

Column	t-statistic	p-value	Conclusion
age	-5.2814	3.03e-07	The correlation is statistically significant.
thalch	0. 8657	0.3876	The correlation is not statistically significant.
exang	-1.0404	0.2993	The correlation is not statistically significant.
oldpeak	-1.6075	0.1099	The correlation is not statistically significant.
ca	-2.4372	0.0169	The correlation is statistically significant.
cp_asymptomatic	-1.7207	0.0867	The correlation is not statistically significant.
cp_atypical angina	1.6215	0.1063	The correlation is not statistically significant.
slope_flat	0.0813	0.9353	The correlation is not statistically significant.
thal_reversable defect	-2.3402	0.0202	The correlation is statistically significant.

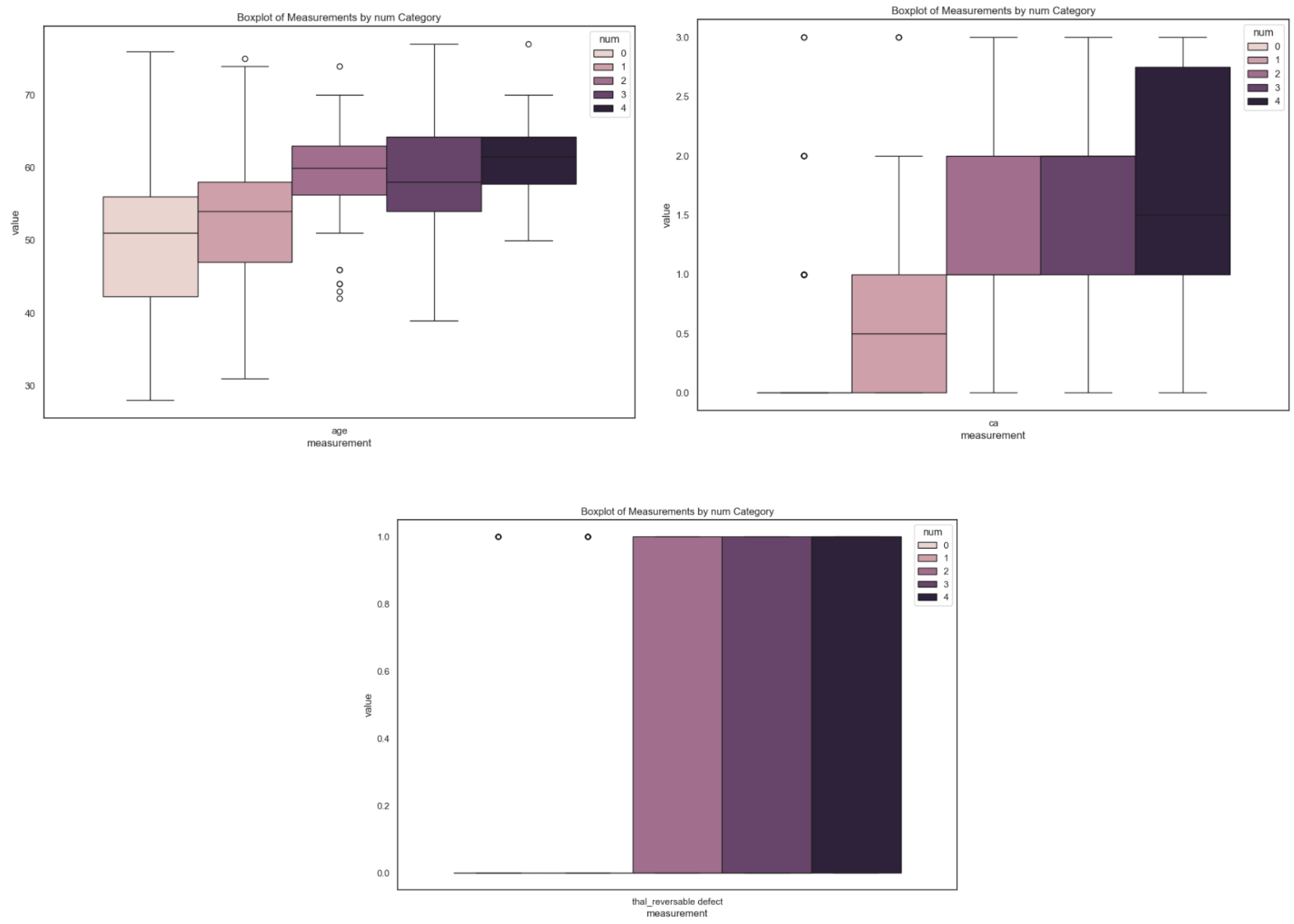

Figure 6. Box Plots of the age, ca, and thal_reversable defect Columns Based on the num Column"

Conclusion

Upon reviewing all of this data, it can be scientifically accepted that the values for age, ca, and thal_reversable defect are particularly relevant to heart disease. These variables provide significant insights into the disease's condition, indicating a strong association with heart disease status. Specifically, age reflects the impact of aging on heart disease, while ca and thal_reversable defect are indicators of heart disease severity and progression. Thus, these variables are critical for understanding and analyzing the relationship between heart disease and its various clinical manifestations