Predicting Heart Disease

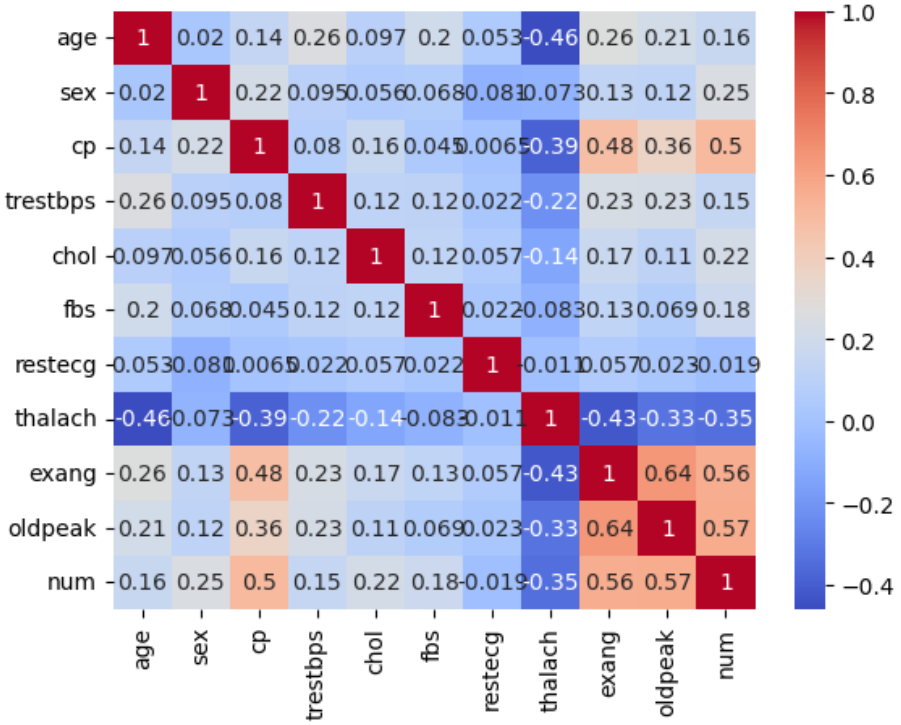
Overview of the dataset

Here I have chosen "processed.cleveland.data", "processed.hungarian.data", "processed.switzerland.data", "processed.va.data" datasets for analysis. I have done the most part of the analysis with hungarian dataset. Here is a quick overview of the data in above processed datasets.

No.	Features	Description	Value
1	Age	Age is an important aspect of health care.	Its value is an integer.
2	Sex	Gender	Female = 0, Male = 1
3	Chest pain(cp)	The patient is suffering from chest pain.	Asymptomatic = 4, typicalangina = 1, atypicalangina = 2, non-anginal pain = 3
4	RestingBloodPressure (trestbps)	High blood pressure ensues with some other factors which increase the risk.	It has either an integer or float value.
5	Cholesterol(Chol)	Serum cholesterol	It has either an integer or float value
6	FastingBloodSugar(Fbs)	Fasting blood sugar is more than 120 mg/dL	0 = false; 1 = true
7	RestingECG (restech)	ElectroCardioGraphic Resting	ST-T wave abnormality =2, Normal =0, Left ventricular hypertrophy =1,
8	Max Heart Rate Achieved (thalach)	This is the highest heart rate you have ever had.	It has either an integer or float value.
9	Exercise-Induced Angina (exang)	Angina instigated by exercise	no = 0, yes = 1
10	Oldpeak	Exercise-tempted ST depression compared to rest	It shows the value as either an integer or a float.
11	Slope	slope of peak exercise ST segment	flat = 1, downsloping = 2, Upsloping =0
12	Coronary Artery (ca)	Fluoroscopy has colored a large number of major vessels.	It has either an integer or float value.
13	Thalassemia (thal)	Normal, reversible defect, fixed defect,	Measuring scales: 3 = normal; 7 = reversable defect; 6 = fixed defect
14	Num(target: Heart Disease predicting attribute)	Heart disease diagnosis (angiographic disease status)	0 indicates a diameter narrowing of less than 50%, 1 indicates a diameter narrowing of more than 50%.

Corelation Matrix

Correlation is a statistical feature that describes the strength and route of a linear relationship among two quantitative variables. A correlation matrix with heatmap is shown below. Using a heatmap, you can see how dependent values are affected by independent features. Furthermore, it is easy to see which features are greatest associated with the additional features variable.

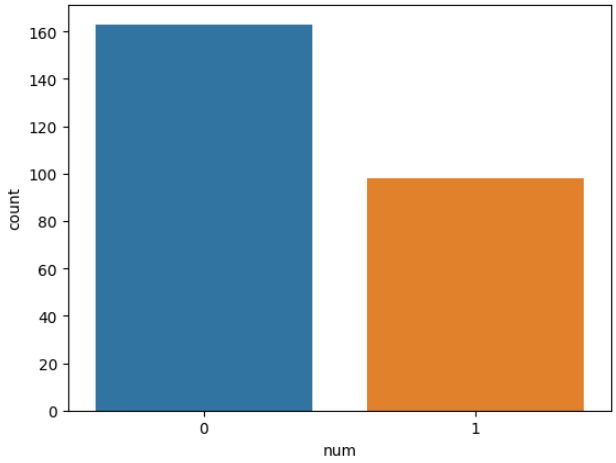


Results and Analysis

We can plot the feature of the heart disease dataset vs. num (predictive attribute) for data visualization. Exploratory Data Analysis (EDA) is a technique used for analyzing datasets to summarize their main characteristics, which is frequently accomplished through the use of statistical graphics and other data visualization methods.

Disease Status

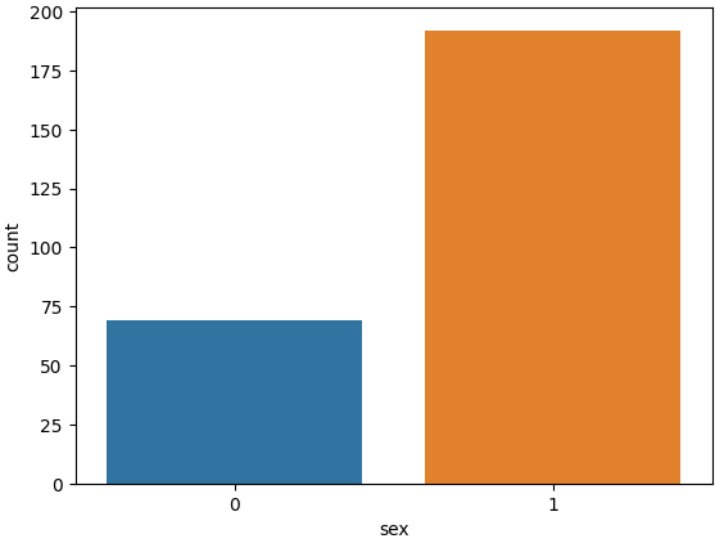
Here we can see that less than half people are diagnosed with disease. We can represent ‘diseased’ with 1 and ‘normal’ with 0.



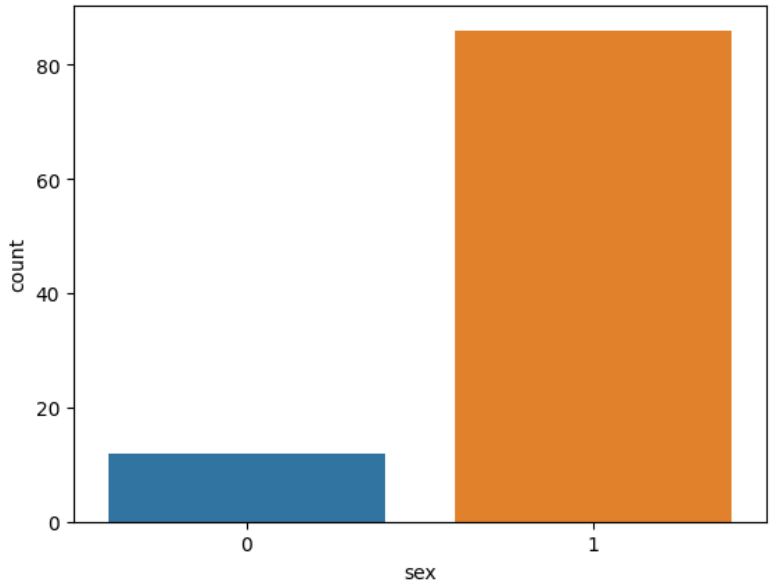
We can also analyse the other dataset attributes such as Age, Chest Pain, Sex, Exercise-Induced Angina, Fasting Blood Sugar, Resting ECG, Slope, Coronary Artery, and Thalassemia features.

Sex

Here are total number of males and females.



In the sex attribute, we have two values, male and female: 0 is used for females, an 1 is used for males. Males are additional likely to have heart problems than females, according to the findings.

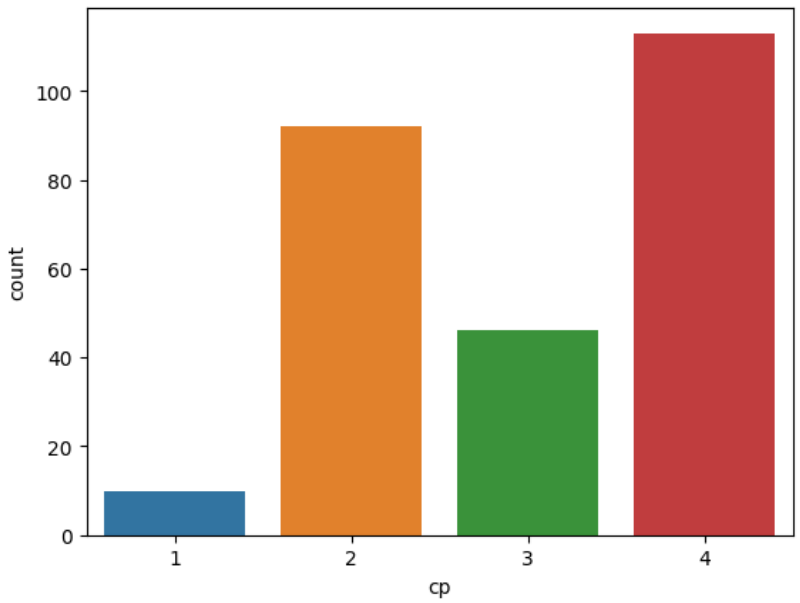


Analysing Chest Pain

Patients with heart disease may experience chest pain. As shown in following figure, chest pain in the subsequent categories: non-anginal pain = 3, asymptomatic = 4, atypical angina = 2, typical angina = 1. We have noticed that people who have '0' chest pain,

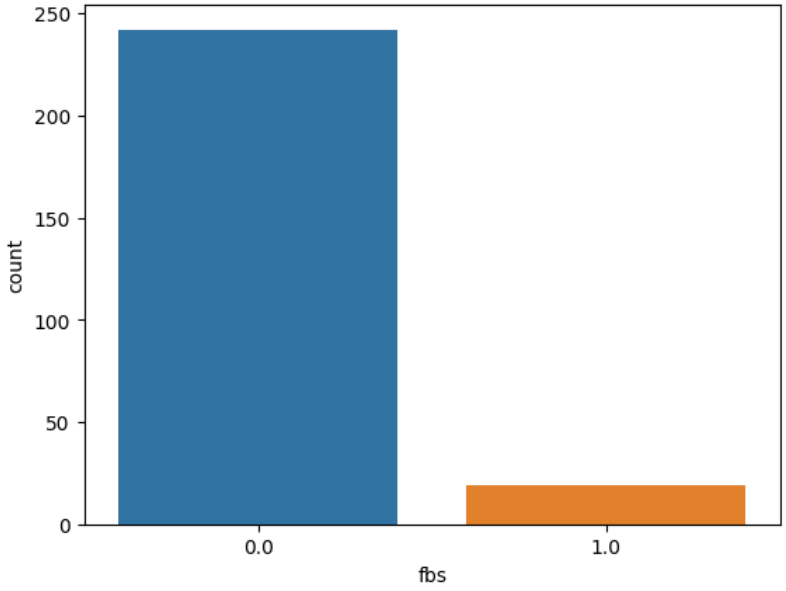
i.e., those who have typical angina, are considerably less likely to have heart difficulties.

Patients who have asymptomatic angina have increased chances of heart disease occurrence.



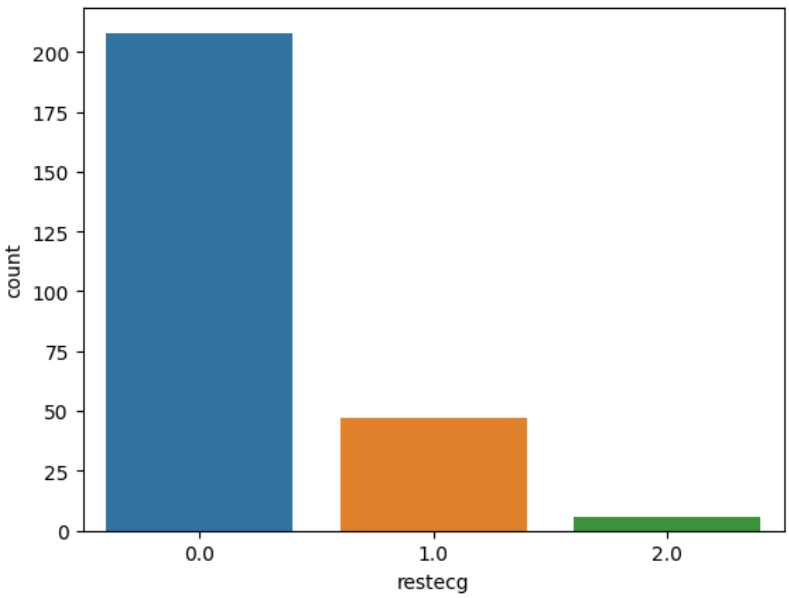
Analysing Fasting Blood Sugar

Fasting blood sugar (FBS) cannot play many roles in heart disease occurrence. We can analyse the dataset in which if the patient's fasting blood sugar level exceeds 120 mg/dL, it means that they are facing it, and we represent it by the value 1 (True); the other case is represented by the value 0 (False), as shown in following Figure. The outcome shows that there is nothing extraordinary here for predicting the presence of heart disease.

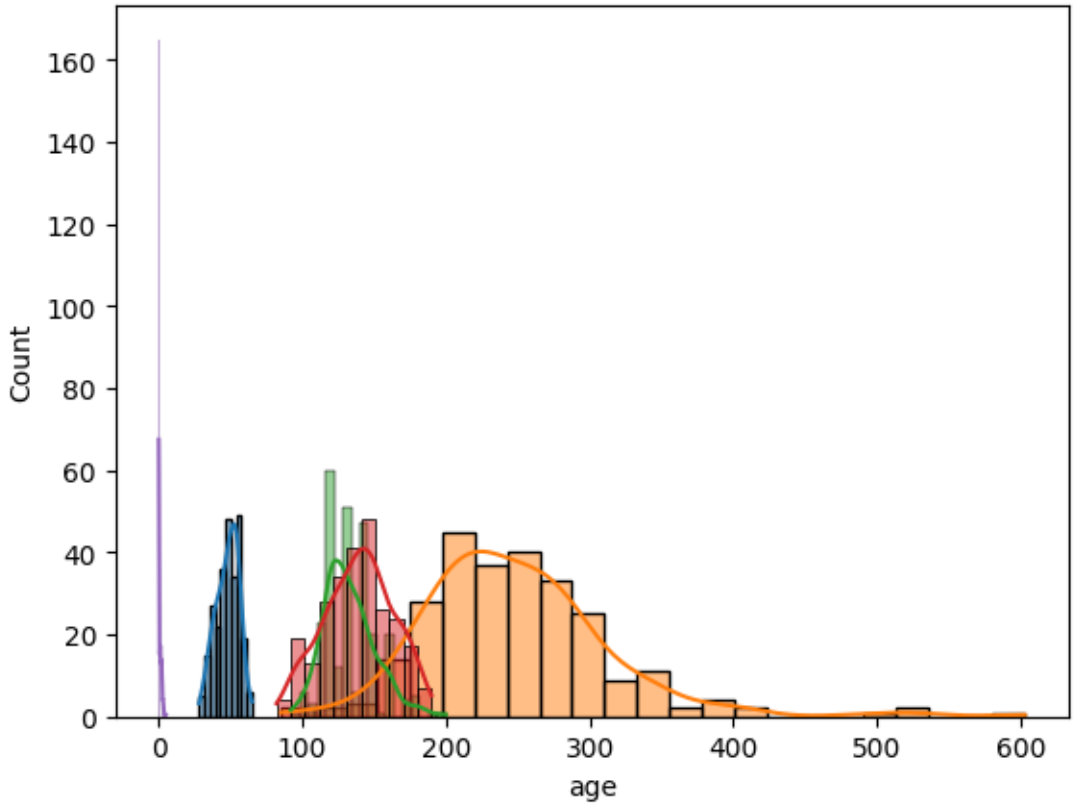


Analysing Resting Electrocardiographic

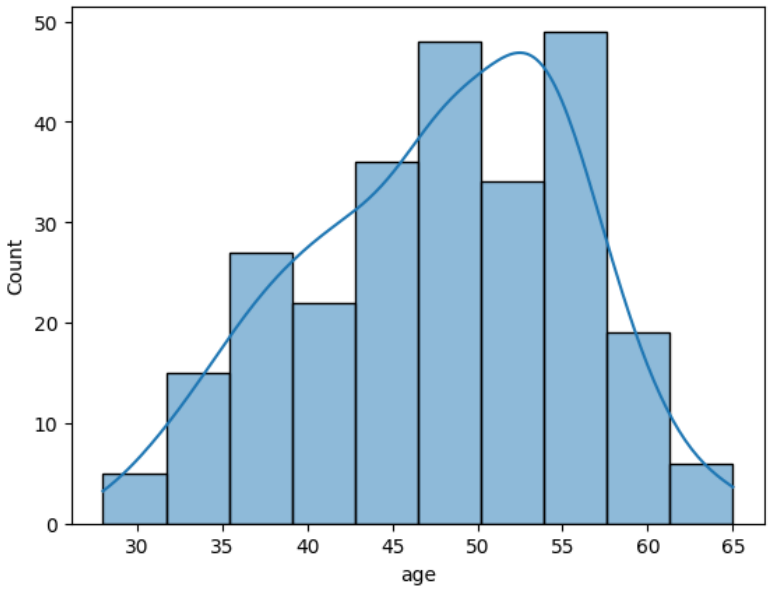
Resting Electrocardiographic values are 0, 1, and 2. The outcome shows that individuals with Resting ECG values of '0' have increased chances of heart disease as compared to Resting ECG value '1' and '2'.



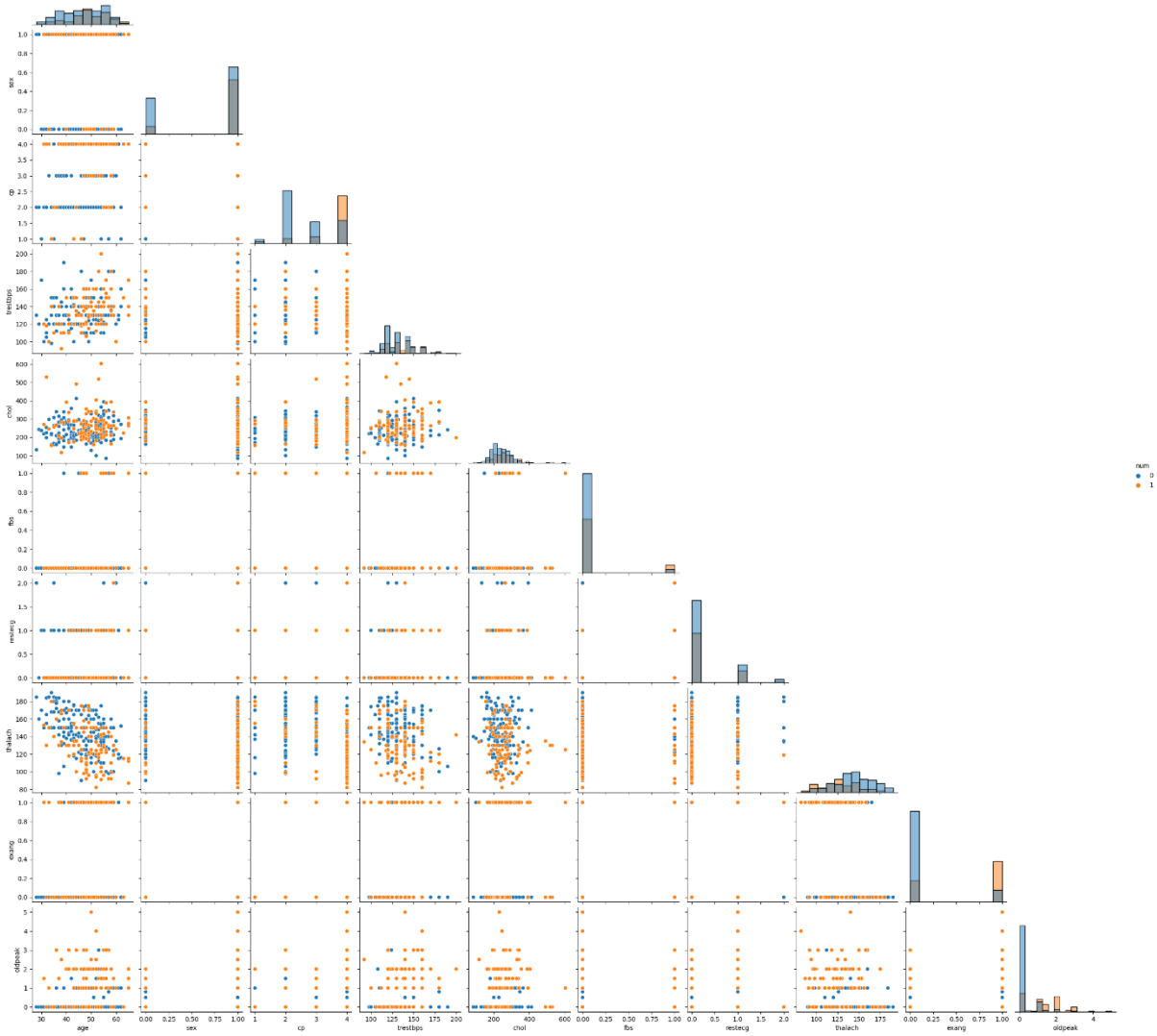
Numeric Variable Distribution



Age



Numerical Feature Pairplot



Feature Selection

It is better to drop the columns with higher number of "not available" values. Therefore I have omitted “slope”, “ca”, “thal” columns in “processed.hungarian” dataset.

We can see that all the parameters in cleveland dataset has values. Therefore we can use all the parameters.

In “processed.switzerland” dataset, we can see that "ca" column has only 5 values. Therefore we can ignore that column. Even though "fbs" and "thal" columns have half of the values, we cannot use those columns in analysis because we don't have enough data in those columns.

In “processed.va” dataset we can see that "ca" column has only 2 values. Therefore we only consider other columns for analysis. We can remove the records with not enough values just like we did in the previous dataset.

Logistic Regression

Here we can visualize relationships between features and target.

We can use logistic regression because the response variable ("num") is binary. Logistic regression is specifically designed to model binary response variables, making it a good choice for this type of dataset.