**Heart Attack Possibility Prediction**

**Milestone: Data collection and preprocessing**

Group 18

Student 1 Karan Paresh

Student 2 Tianyu Yang

857-294-7660 (Tel of Karan Paresh)

617-368-0332 (Tel of Tianyu Yang)

[samani.k@northeastern.edu](mailto:samani.k@northeastern.edu)

[yang.tianyu@northeastern.edu](mailto:yang.tianyu@northeastern.edu)

**Percentage of Effort Contributed by Student 1: 50%**

**Percentage of Effort Contributed by Student2: 50%**

**Signature of Student 1: Karan Paresh**

**Signature of Student 2: Tianyu Yang**

**Submission Date: 10/4/2021**

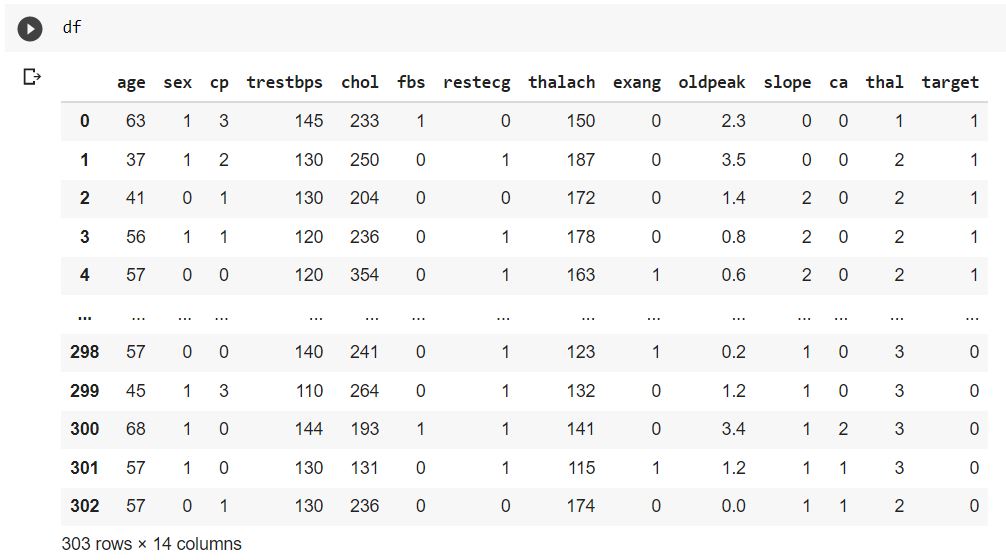
**Data Collection and Pre-processing**

**Data description: -**

The dataframe has 14 variables and 303 records, the data has both continuous and discrete variables.

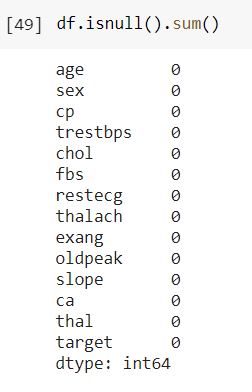
Discrete variables = sex, cp (chest pain type), fbs (fasting blood sugar), restecg, exang (exercise induced angina), slope, ca (major vessels colour by flouropsy) , thal, target

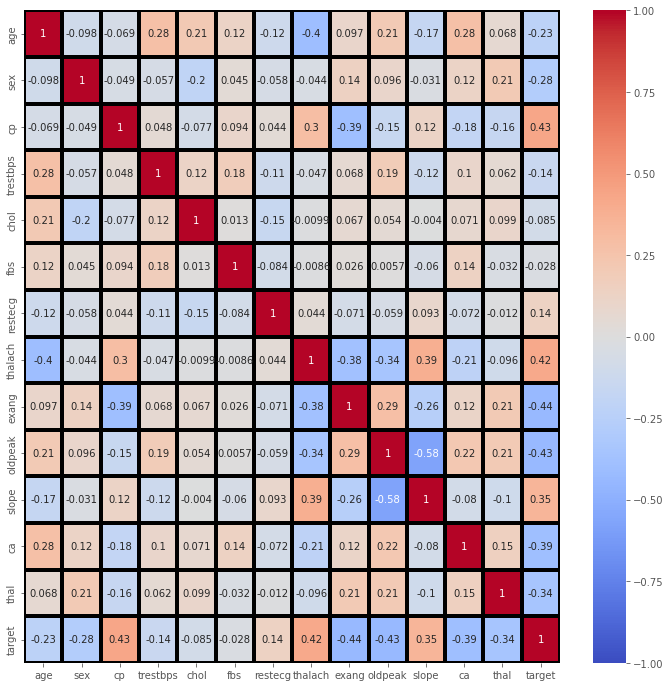
Continuous variables = age, trestbps (resting blood pressure), chol (serum cholesterol), thalach, oldpeak (ST depression induced by angina).



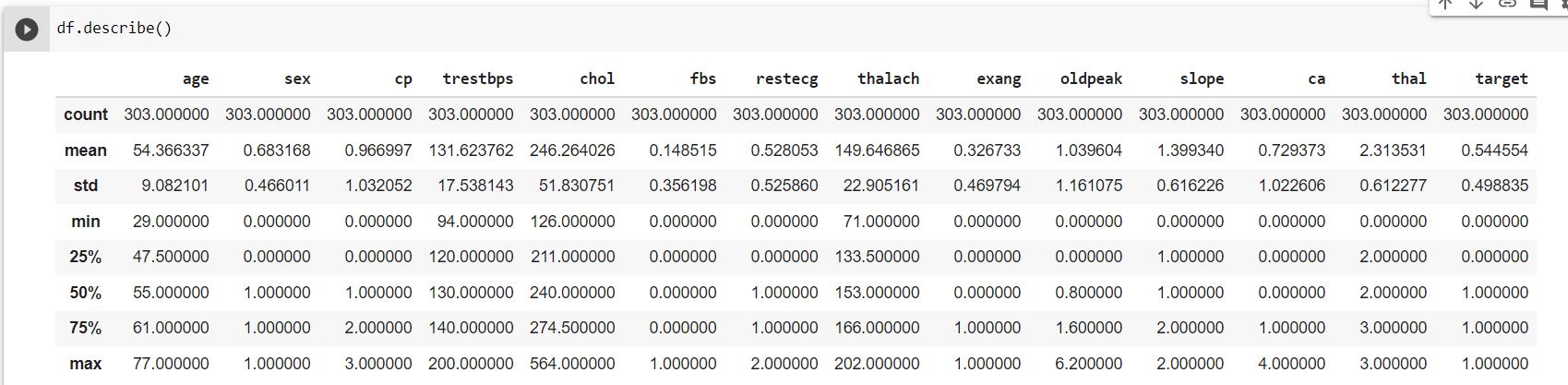
**Null Values: -**

The heart attack dataset contains zero null values and the data imputation on the null values aren’t required. The label encoding is also not required for any of the features.



**Feature correlation matrix: -**

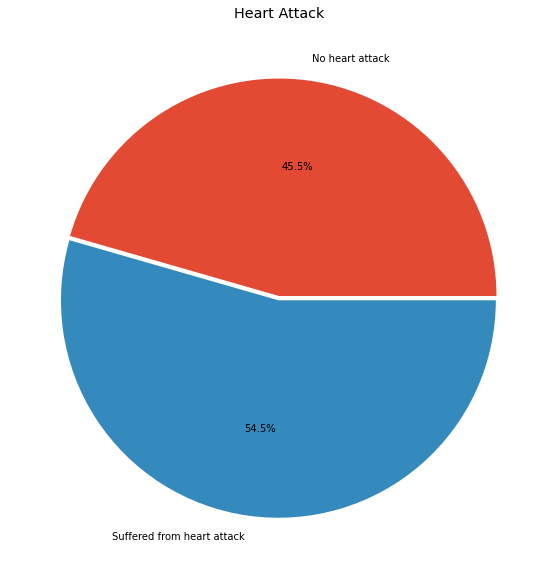
We can see there is a **positive correlation** between chest pain (cp) & target (our predictor). This makes sense since, the greater amount of chest pain results in a greater chance of having heart disease. Cp (chest pain), is a ordinal feature with 4 values. In addition, we see a **negative correlation** between exercise induced angina (exang) & our predictor. This makes sense because when you exercise, your heart requires more blood*,* but narrowed arteries slow down blood flow*.* This matrix shows the correlation of every variable with each other and there by we can decide on how many features are needed to predict the target variable. The feature selection can also be done using backward and forward selection.

**Summary of the data: -**

**Initial insights of the data: -**

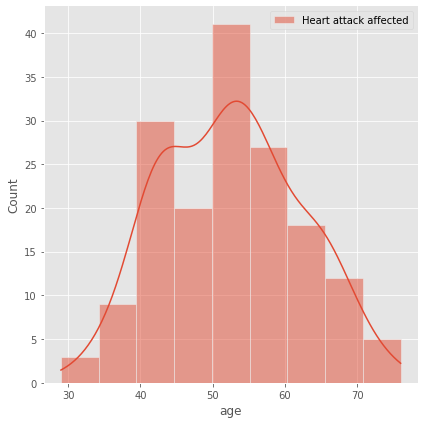
Figure 1: -

The pie chart represents the percentage of people who may suffer from the heart disease and percentage of those who may not. The result shows that 45.5% people do not suffer from heart attack and 54.5% people suffer from heart attack.



*Figure: 1*

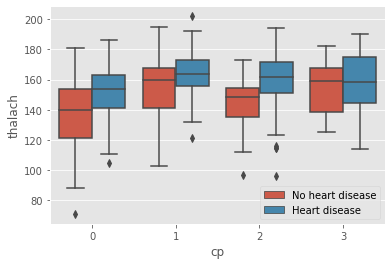
Figure 2: -

The displot shows that the Heart attack is very common in the seniors which is composed of age group 60 and above and common among adults which belong to the age group of 41 to 60, but it’s rare among the age group of 19 to 40 and very rare among the age group of 0 to 18.

*Figure: 2*

Figure 3: -

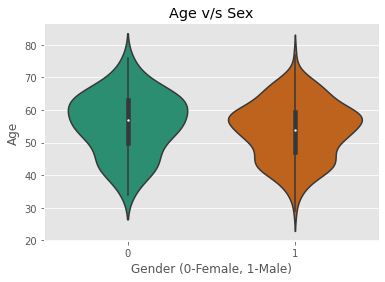
The boxplot shows that the heartrate (thalach) is gradually higher for those who are suffering from the heart disease and on the other hand as the severity of chest pain (cp) increases the mean heartrate is also increasing.



*Figure: 3*

Figure 4:

Visual illustrates the age distribution of male and female and the mean age of both the genders is almost same and the kernel density is higher for the age groups between 50 to 70 years.



*Figure: 4*