So, from the dataset we can see we have 14 attributes along with the target attribute which tells us if this person/row have the heart disease or not. The shape of dataset is 303x14. The sex of person is 1 = male and 0 = female, the cp (chest pain) attribute has 4 values (1=typical angina, 2=atypical angina, 3=non-anginal pain, 4=asymptomatic). The trestbps is a person resting blood pressure in mm Hg when he was admitted in hospital. chol is the person cholesterol measurement in mg/dl. fbs is the person fasting blood sugar whose values are (1= greater than 120 mg/dl and 0=false). The restecg is "resting electrocardiographic measurement" (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or "definite left ventricular hypertrophy by Estes' criteria"). Exang is the exercise induced angina (1=yes and 0=no). oldpeak is the ST depression induced by exercise relative to rest ("ST is the flat section between S and T wave. It represents the interval between ventricular depolarization and repolarization.). The slope column represents the slope of peak exercise ST segment, it has 3 values: 1=unsloping, 2=flat, 3=down-sloping. Attribute ca as mentioned above the number of major vessels (0-3). The second last attribute is thal which represents a blood disorder known as "thalassemia" and values it can have are mentioned above.

so, we want to find out what are the causes of a heart disease we did some research online and we found out that the diagnosis of heart disease is done by doing different clinical tests and analysing symptoms. The test done for diagnosis depends on the condition a doctor thinks the patient have. The different type of tests includes: Electrocardiogram (ECG), Holter monitoring, Echocardiogram, Stress test etc[[]](https://www.mayoclinic.org/diseases-conditions/heart-disease/diagnosis-treatment/drc-20353124)

The high-risk factors that we found out online after research for a heart disease are: high blood pressure, smoking, high cholesterol, weight, family history and diabetes [[]](https://www.bhf.org.uk/informationsupport/risk-factors), this was from one source, from another source we found out major risk factors that can’t be changed includes: male gender (they have a higher risk of a heart attack than females.), heredity and increasing age. As we mentioned heredity as one of the major risk factors that can’t be changed so from our attributes that we have in our dataset thalassemia is one.

Now as we know a little about our dataset and what attributes we have and which ones are important, we will perform some experiments and play with our dataset so we can have a general idea. We will use three python libraries numpy pandas and matplotlib. "numpy" is fundamental package for scientific computing with Python and linear algebra. "pandas" is data analysis and manipulation library. "matplotlib" as from the name you can have an idea that its used for plotting, we will use this for data visualisation, for general analysis we will mostly use matplotlib an pandas.

As our dataset is csv("comma separated values") file we will use pandas read\_csv function to load the data into a dataframe.

Now as we have dataset loaded in dataframe what we want to check is how it looks like, for this we can use pandas head(n) function, it returns the n rows and n is an integer.

Now as we have seen how our dataset looks like what we need to do further is use pandas describe() function. This is a kind of descriptive statistics function which will tell us about the number of rows we have in each column mean, standard deviation, minimum and maximum value in the rows and the percentiles. By looking at the result of age column we say that we have 303 entries, the mean age is 54.36, the youngest person is 29 and the eldest person is 77. From the standard deviation we can have an idea how is the spread around the mean value. From percentiles we can have an idea how much percentage of values we have in 25, 50 and 75 percentiles.

Now before visualisation an important step is to clean the data of any duplicates so what we will do is remove the duplicate rows by using pandas drop\_duplicates() function. As there was only one row that was duplicate, we removed it.

Now we want to find out the number of males and females we have in our dataset. As we have already mentioned that one risk factor that can't be changed for a heart disease is if you are a male. But now we just want to know numbers of males and females. We will use pandas value\_counts() function on "sex" attribute, it returns a series containg counts of unique values in descending order in our case its only 1 and 0 (1=male, 0=female). We will verify this by another function as well called .count() and a condition where ['sex']==1. We get same results and found out that in dataset we have more males than females this also verifies one of the risk factor that can't be changed is being male.

Now we will use bar-plot and pie-plot from matplotlib to visualise our findings and it will give us an idea how much percentage of population is male and female. As from the pie chart we found out that 68.2% of the population is male that is colour blue and 31.8% of the population is female.

To continue our analysis what we now want to do is look at the distribution of population with respect to age and after that we will look at the distribution with respect to gender/sex as well. We will do this by making histogram of age column and for this we will use matplotlib hist() function. The bin size is 2. From the results we can see that we can fit a bell curve through this distribution if we remove one of the middle spike, the other thing is that there is a great spread across the mean thats what we also analysed from the standard deviation in start. From the second distribution where we have labelled female by orange and male by blue shows that the start from 29 to 31 we have no female in our dataset and after that the count of females is nearly half as that of males till 47-49(age group) where we have much less females as compared to males. We can see also that in certain number of bins/age-group we have nearly same number of males and females that is for 61-67. Another thing we noticed is that for age group 71-77 we only have females and there are no males in that age group but at the extreme 77-79 we have a smaller number of males and no female in that range. By these distributions we now have idea about our population, and we know much more in detail that most of the population is between 41-69 and in which age group we have no males and no females and where the count is less.

Now we will investigate distributions of people that have a heart disease and people not suffering from a heart disease. For this we will use the age and target attributes, age because we want to look at the population's age relation to target attribute distribution. By looking at the plots we analysed that for age group 41-45, 51-55 and 57-59 the count of people suffering from heart disease is higher than the mean count, but there is no relation it’s kind of random. For the second plot people not suffering from heart disease we can see that age group 57-59, the count is high. For this plot there is also no clear relation that we can talk about. We can say that probability of suffering from heart disease decreases after 60, because there is a decreasing trend.

As our main problem is to find the causes of a heart disease so one other thing, we could do to find the correlation between all the attributes and the target by this we will have a rough idea that how a specific attribute is more related to the heart disease. For this we used pandas corr() function it will return us a dataframe where we can see the correlation of every attribute with other attributes, but for now we just want to see with the "target". We have sorted the output in descending order and colour labelled with intensities accordingly. It can be analysed that 'cp' chest pain and 'tahlach' maximum heart rate achieved are more correlated to target as compared to others. For most of the attributes there is negative correlation as well.