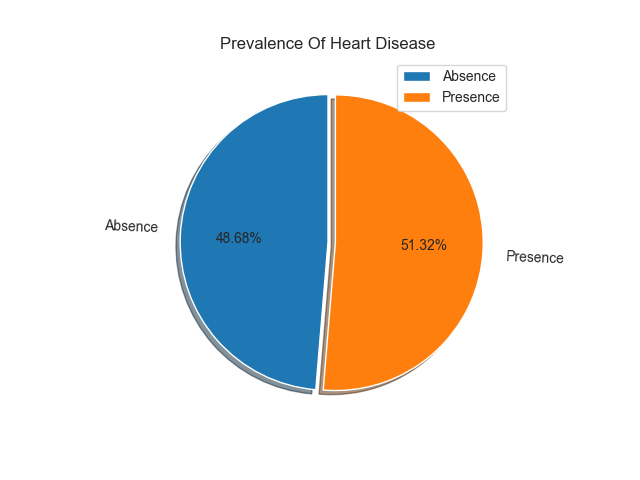
**Heart disease Diagnostic analysis**

**Problem statement**

Health is real wealth, in the pandemic time we all realized the brute effects of covid-19 on all irrespective of any status. In recent times, it has been observed the rise in death attributed to heart diseases, even among prominent individuals. Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year1 .

In the scope of this project, it was observed that more than 50% of sample size has been diagnosed with heart disease.



Thus, the relevance of the problem can not be ignored and it demands for all the insights one can get regarding heart disease.

**Research Question**

What factors correlate with the presence or absence of heart disease, and which metrics are pivotal in understanding these associations?

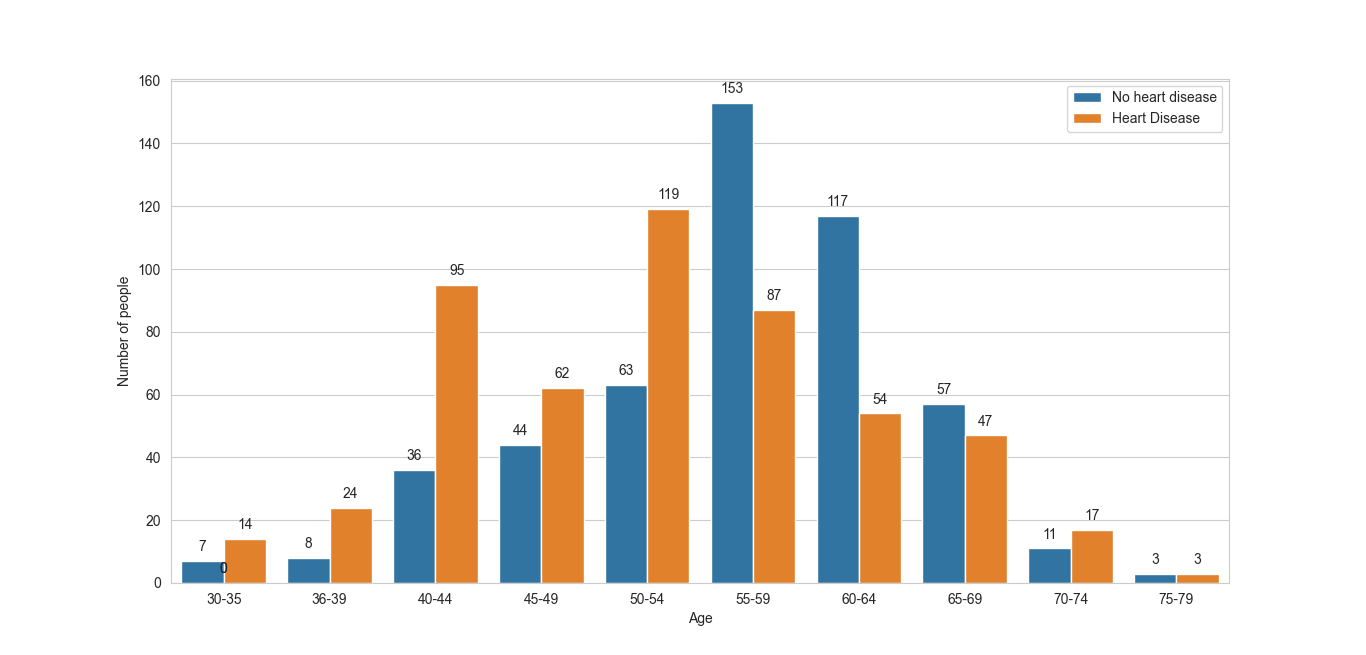
**Assumptions**

For the scope of this project, there are handful of assumptions has been made for analyzing purposes

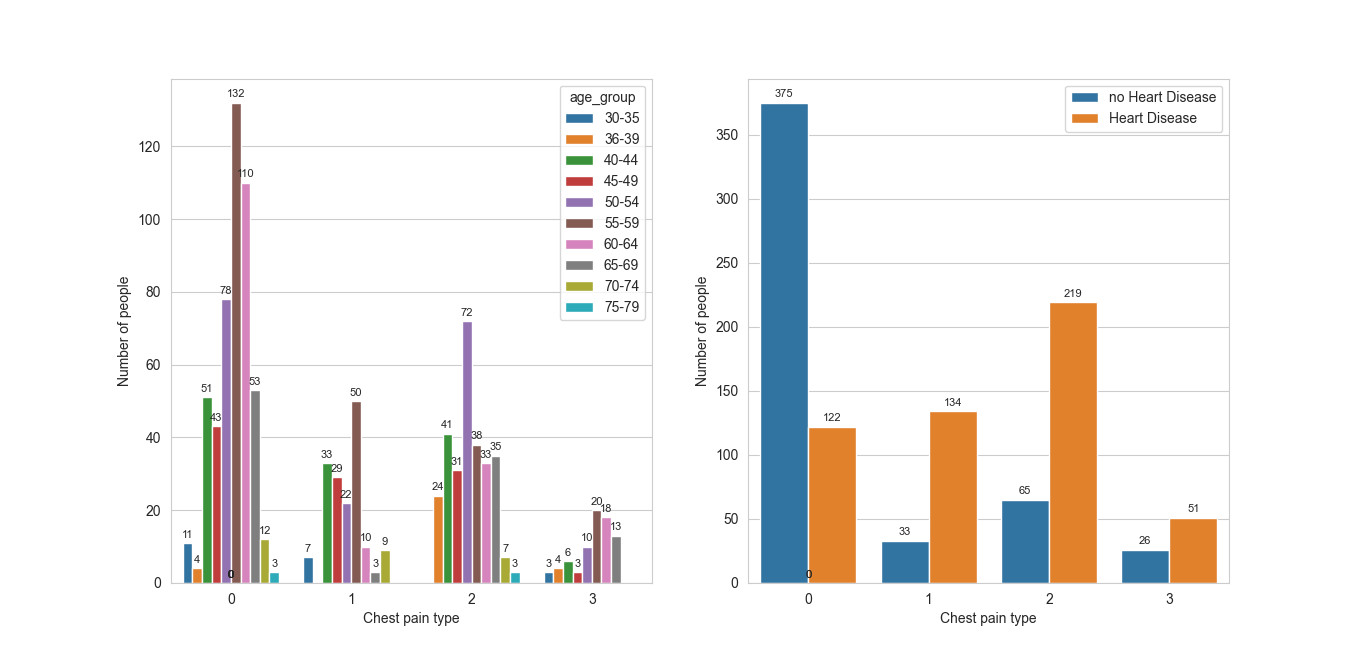
* The data has been assumed to not have many outliers which will affect the analysis drastically
* The data is valid in the current status quo .

**Subgroup analysis**

When we observe the age distribution with respect to age:

Although, age group 50-54 has highest number of heart patients I.e. 119. When we compare the risk factor which is no. Of heart patients divided by total patients of that age group, it comes highest for 40-44, that is , 72.52%. But it will be hurry to conclude anything without comparing it with chest pain distribution.

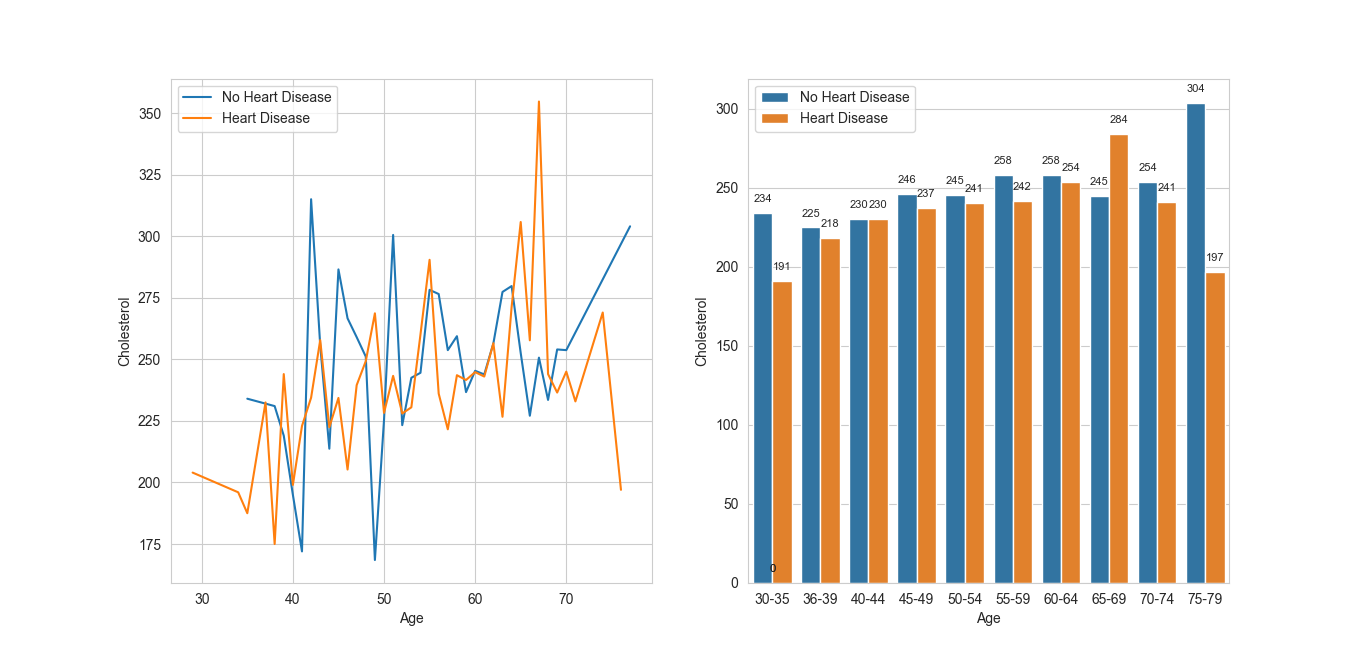
While when it was observed the chest pain type distribution across age group and target,



It’s safe to conclude chest pain type 0 is most common among the people. And CP type 1 is most prominent with risk factor of **80.24%** and type 2 with **77.11%**.

**Key findings**

1. **High Cholesterol can not solely be reason for heart disease:** The data suggests that cholesterol, at least, can not be sole reason leading towards heart disease.



The bar-graph and line-plot analysis demonstrates for many age groups the avg cholesterol level of people having no heart disease is more than people having heart disease. However, the possibility that it may impact other factors which potential cause of heart disease is still open.

1. **The fasting blood sugar has independent relation to heart disease:**  While having blood sugar can potentially lead to diabetes but the statistical test done proves no such relation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | No-Heart Disease  (Observed) | Expected Value | Heart Disease  (Observed) | Expected Value | Total  (Observed) |
| FBS=0 | 229 | 234.26 | 202 | 196.74 | 431 |
| FBS=1 | 52 | 46.74 | 34 | 39.25 | 86 |
| Total | 281 |  | 236 |  | 517 |

FBS=1: FBS>120 mg/dl

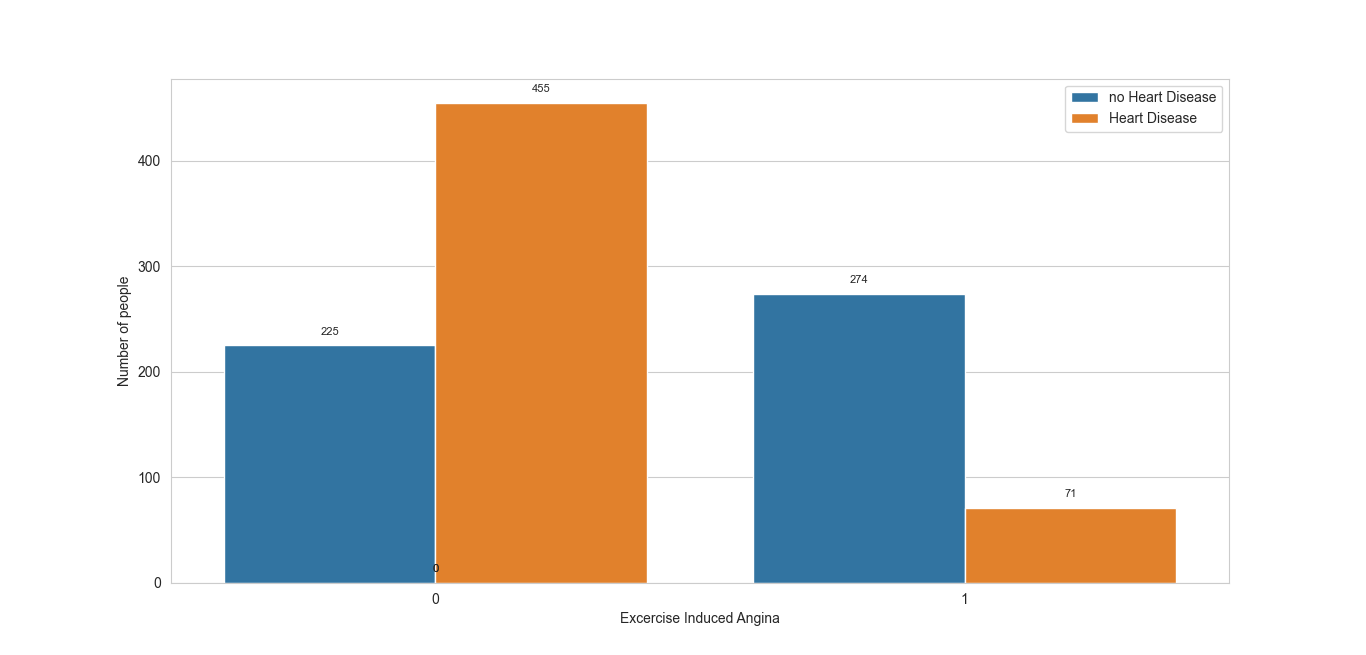
FBS=0:FBS<120 mg/dl

**The chi-squared value for the following contingency relation came: 1.272** (which is less than the critical value 3.841 for alpha=0.05 and DOF=1 )

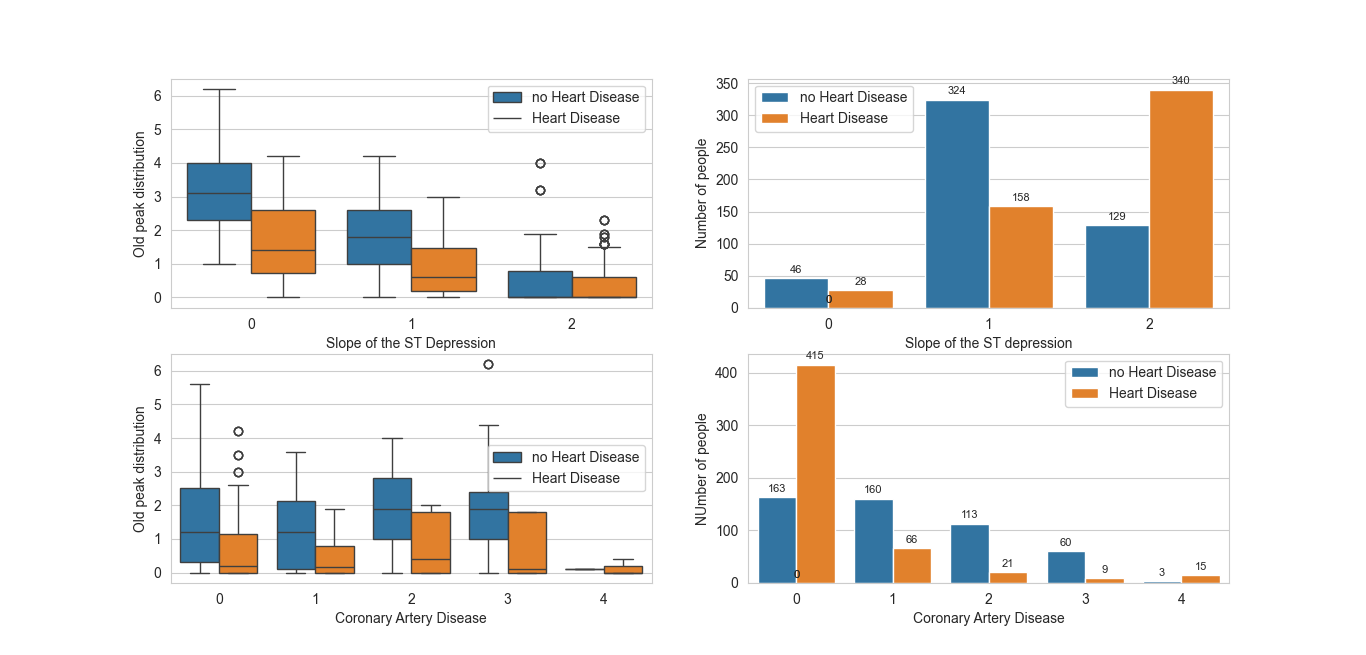
**And the corresponding p-value came out to be: 026** (which is greater than the reference value 0.05)

Thus, it’s safe to conclude that the table passes passes the chi-square test. Thus, Null hypothesis has been proven, and they have apparent independence.

1. **The exercise induced angina(exang) is a good sign:** The analysis suggest people having positive exercise induced angina are in lower risk of heart disease.

 Here it can be observed that people experiencing no exang have higher chance of heart disease.

1. **The old peak distribution [0,1] is more likely to fall in risky category with respect to heart disease:** People having old peak lies between 0,1 are at higher risk for heart disease



Here it can be observed that slope of ST depression type 2 and Coronary Artery Disease type 0 are most common among heart patient, comparing it to old peak distribution concludes, heart patients resides in the range 0-1.

**Limitations of the analysis:**

1. **Sample Size:** The analysis is done for the sample size of approx 1000 individuals. Thus, there might be apparent misguiding trend which affected the analysis.
2. **Potential confounding Variables:** There might be parameters not included in this analysis which can have impact on the trend or pattern derived here like the environment patient lives in, diet etc.
3. ****Data Quality and Bias:**** The analysis relies on the quality and accuracy of the available data. While efforts were made to ensure data integrity, there may be inherent biases or inaccuracies in the data collection process that could affect the validity of the results