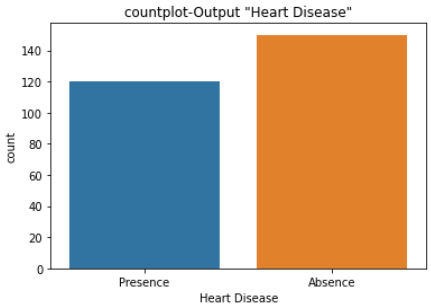
Exploratory Data Analysis of Heart Disease

This assignment focuses on exploratory data analysis of heart disease. The data is obtained from kaggle.com. The link for the data is given below.

https://www.kaggle.com/johnsmith88/heart-disease-dataset

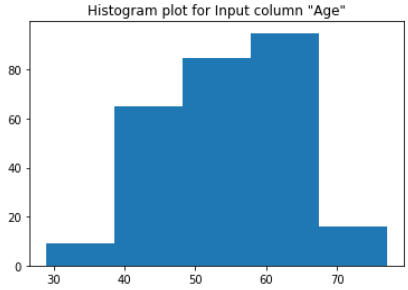
A detailed understanding of the data is carried out by using multiple plotting methods which results in understanding both numerical and categorical data. A shorter Research question is formulated and plotting is done to address the research question.

Q 1: Is the heart disease data contains equal number of classes?



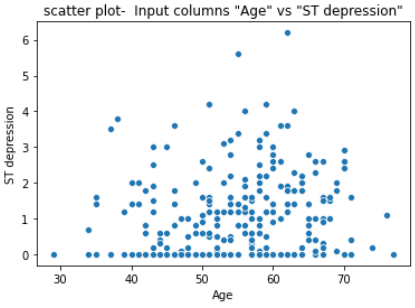
Based on the count plot of dependent variable heart disease, it is understood that the data suffers unbalanced data problem. The count of absence of heart disease is greater than presence of the heart disease.

Q 2: Are the data in the column age is normally distributed?



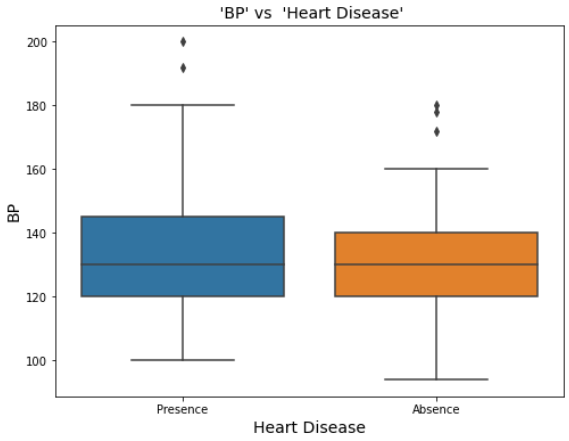
It is clear from the plot histogram of the column Age that the data does not follow a normally distributed distribution, as shown in the figure. The data has a skewed distribution to the right.

Q3 : Is there a correlation between the column age and ST depression?



It is clear from the scatter plot that there is no linear relationship between the two variables. Neither there is a positive nor a negative correlation between the two variables.

Q4: Is there a significant difference between population means of column BP?



Box plot is plotted between numerical data BP and categorical data Heart Disease. From the plot, it is inferred that there is no significant difference between populations of the column BP.