Task to be performed:

1. Preliminary analysis:

a. Perform preliminary data inspection and report the findings on the structure of the data, missing values, duplicates, etc.
b. Based on these findings, remove duplicates (if any) and treat missing values using an appropriate strategy.

- In preliminary data analysis, It was observed that, the data is consist of : 303 rows & 14 columns.
- No null values were present in the data.
- Row ID: 163, and 164 were duplicates, Hence, one of the row data was removed.
- Column names were replaced with standard variable names.

2. Prepare a report about the data explaining the distribution of the disease and the related factors using the steps listed below:

a. Get a preliminary statistical summary of the data and explore the measures of central tendencies and spread of the data.

- Statistical summary was created to get the mean, STD and Min & Max values of continuous data.

b. Identify the data variables which are categorical and describe and explore these variables using the appropriate tools, such as count plot

- Statistical summary was created to get the mean, STD and Min & Max values of continuous data.

c. Study the occurrence of CVD across the Age category

- Highest number of CVD patients were observed in the dataset between 40-55 age.

d. Study the composition of all patients with respect to the Sex category.

- Overall, higher number of data is available for male patients then female.
- Count plots were plotted for major data variables such as chest pain, fasting blood sugar etc. with respect to gender.
- Number of male patients with typical angina, a type of chest pain was significantly higher than female patients.

- Similarly, reversable defect in thalassemia was observed higher in Male patients compared to Female.

e. Study if one can detect heart attacks based on anomalies in the resting blood pressure (trestbps) of a patient

- With correlation matrix data, it was observed that resting blood pressure is not corelated with heart disease condition.
- However, a simple logistic regression method can predict the heart attacks only on based on resting blood pressure with 60% accuracy.

f. Describe the relationship between cholesterol levels and a target variable

- Data was divided in three different group based on cholesterol levels (As per standard limits). <200 = Normal, 200-240 = borderline cholesterol and >240 = High cholesterol.
- It was observed that people with borderline were higher in heart disease condition compared to no CVD.

g. State what relationship exists between peak exercising and the occurrence of a heart attack

- People with 'downsloping' st_slope in peak exercising have higher risk of heart attack.

h. Check if thalassemia is a major cause of CVD

- Thalassemia is marginally negatively co-related with heart disease and from the count plot, frequency of people with normal thalassemia level with heart disease were higher.

i. List how the other factors determine the occurrence of CVD.

- From Correlation matrix, chest pain, max heart rate and st_slope is positively corelated with occurrence of heart disease.
- Whereas, exercise induced angina, and number of vessels and st_depression is negatively corelated with occurrence of heart disease.

j. Use a pair plot to understand the relationship between all the given variables

- A pair-plot was plotted with continuous columns to check their positive or negative correlation between the variables.

3. Build a baseline model to predict the risk of a heart attack using a logistic regression and random forest and explore the results while using correlation analysis and logistic regression (leveraging standard error and p-values from statsmodels) for feature selection.

- First scaling of numerical data and for categorical data dummy variables were obtained with standard scaler and get dummies method.
- Accuracy of predicting heart disease with logistic regression model was 86.81%.
- Whereas, we obtained accuracy of predicting heart disease with random forest at 84.61%.