**Inference from Data Analysis of CVD Dataset**

1. **Age**

Age Distribution for Diseased and Healthy Patients

**Female**: high risk band - 50-55 (Steady increase from 40)

**Male**: Even though count is maximum at 50-55, depending on Healthy to Diseased Count ratio, high risk at 40-45(Similar distribution to Whole Dataset)

**Whole**: Even though count is maximum at 50-55, depending on Healthy to Diseased Count ratio, high risk at 40-45

1. **Gender**

Almost 3/4th of female Population is diseased and lesser no of male % affected by positivity of CVD. Women at more risk in this dataset.

1. **Chest Pain Type**

Angina contributes the highest to CVD followed by non-anginal pain and asymptomatic, typical angina contributes to negative in CVD.

1. **Cholesterol and Blood Pressure**

Cholesterol and Blood pressure have no significant variation in range for diseased and healthy and hence doesn’t have a direct co-relation in this dataset.

1. **Diabetes**

Diabetes doesn’t have a direct co-relation individually in the data set, because spread of diabetes data remains almost the same in Diseased and Healthy.(85%:15% ratio is maintained for Non-Diabetic and Diabetic data spread respectively across Diseased and Healthy)

1. **ECG**

Abnormal ECG contributes significantly to CVD, 63% of Abnormal ECG values leads to positivity(hyper can be neglected since the value is too less) and normal ECG does indicate CVD negative 54% of Normal ECG data leads to negativity in CVD.

1. **Max heart rate, ST depression and Major Vessels**

The Max heart rate does seem to play a role, with a higher range of values from 114-202 with bulk of its values from 148-161 range as compared to a lower range in healthy 125-156.

The ST depression too has a different range for Diseased and Healthy, the diseased has a lower range, ranging from 0-2.6, whereas health patients’ range is from 0-4.4, so value > 2.6 might help in negativity in CVD.

Major Vessels value too plays a critical role in CVD positivity, Lower range of Value 0 and 1 contribute significantly to CVD, Negative Co- relation exists.

1. **Exercise Induced Angina**

‘No’ value in Exercise Induced angina contributes to CVD heavily nearly 86% of diseased has ‘no’ value and hence it is a deciding factor.

1. **ST SLOPE**

Down sloping plays a huge role in contribution to CVD ,65% in Diseased has Down sloping and it reduces to 25% in Healthy.

1. **Thalassemia**

Significant relation between Thalassemia value and CVD Positivity, Fixed Defect is an important player, 80% of Diseased data has Fixed Defect.

**Linear Regression Model Explanation**

Three Models has been build in this project:

* **Basemodel –** Consists of all the columns, Accuracy from Confusion Matrix is **82**%

Columns Used: All, No of Columns: **14**

More columns are used.

* **model1 –** Variables suggested by Base R for Model, Accuracy from Confusion Matrix is **78**%

**Columns Used:** sex+chest\_pain\_type+resting\_blood\_pressure+resting\_ecg+max\_heart\_rate+thalessimia+major\_vessels

No of Columns: **7**

**Accuracy is Less**

* **myModel –** Model build from the columns selected by me after Data Analysis. Accuracy is **81**%

Columns Used:

sex+age+chest\_pain\_type+resting\_ecg+max\_heart\_rate+exercise\_induced\_angina+st\_depression+st\_slope+thalessimia+major\_vessels

**No of Columns: 10**

This model gives almost the same accuracy as BaseModel but number of columns used has been significantly lesser than the BaseModel and hence making the Model more effective and faster.