**Heart Disease**

*Apurba Adhikari (202118007)*

*Harsh Doshi (202118038)*

*Siraj Shaikh (202118041)*

* Introduction: -

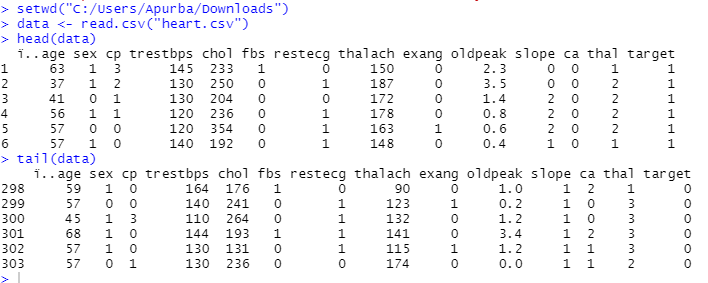
In India, more than 17 Lakh people die every year due to heart diseases and by 2030, the figure is expected to increase with 2.3 crore deaths. Around 26 percent of total deaths in India occur due to non-communicable diseases (NCD), largely referred to the heart diseases.

Heart Disease is a major health threat, that is why it is important to know risk factors so that proactive measures can be taken before it’s too late. Several health conditions such as lifestyle, age and family history can increase risk for heart disease. This report aims to highlight factors that contribute to heart disease by analysing this dataset.

* Description of DATASET: -

The “target” field refers to the presence of heart disease in the patient. It is integer-valued 0 = no disease and 1 = disease. The first four rows and all the dataset features are shown in Table [1](https://www.hindawi.com/journals/cin/2021/8387680/tab1/) without any pre-processing. Now the attributes which are used in this research purpose are described as follows and for what they are used or resemble:

* Age—age of patient in years, sex—(1 = male; 0 = female).
* Cp—chest pain type.
* Trestbps—resting blood pressure (in mm Hg on admission to the hospital). The normal range is 120/80 (if you have a normal blood pressure reading, it is fine, but if it is a little higher than it should be, you should try to lower it. Make healthy changes to your lifestyle).
* Chol—serum cholesterol shows the amount of triglycerides present. Triglycerides are another lipid that can be measured in the blood. It should be less than 170 mg/dL (may differ in different Labs).
* Fbs—fasting blood sugar larger than 120 mg/dl (1 true). Less than 100 mg/dL (5.6 mmol/L) is normal, and 100 to 125 mg/dL (5.6 to 6.9 mmol/L) is considered prediabetes.
* Restecg—resting electrocardiographic results.
* Thalach—maximum heart rate achieved. The maximum heart rate is 220 minus your age.
* Exang—exercise-induced angina (1 yes). Angina is a type of chest pain caused by reduced blood flow to the heart. Angina is a symptom of coronary artery disease.
* Oldpeak—ST depression induced by exercise relative to rest.
* Slope—the slope of the peak exercise ST segment.
* Ca—number of major vessels (0–3) colored by fluoroscopy.
* Thal—no explanation provided, but probably thalassemia (3 normal; 6 fixed defects; 7 reversible defects).
* Target (T)—no disease = 0 and disease = 1, (angiographic disease status).
* **Reading dataset from CSV file: -**

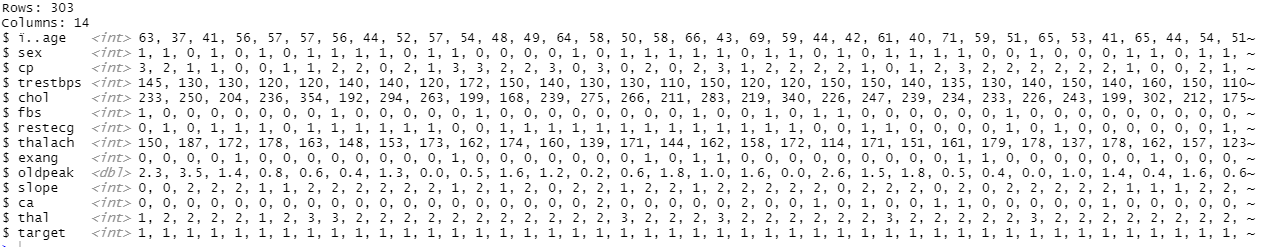
****

* glimpse() is like a transposed version of print() : **columns run down the page, and data runs across**. This makes it possible to see every column in a data frame.
* The tidyverse package actually contains other packages (dplyr, ggplot2, etc.) and you'll see that when you load the tidyverse package using library() .

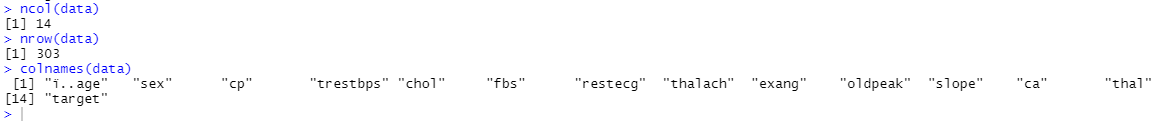
**install.packages('tidyverse')**

**library(tidyverse)**

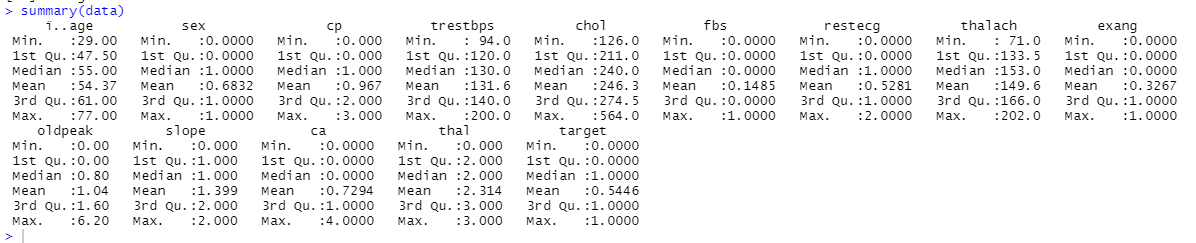
**glimpse(data)**

****

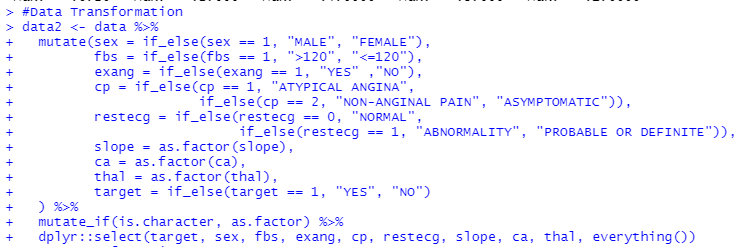
* **Number of columns, number of rows and name of all columns:-**

****

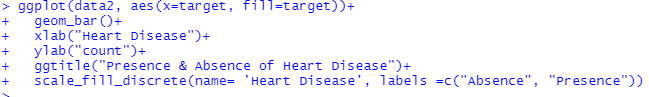
* **Summary of data:-**

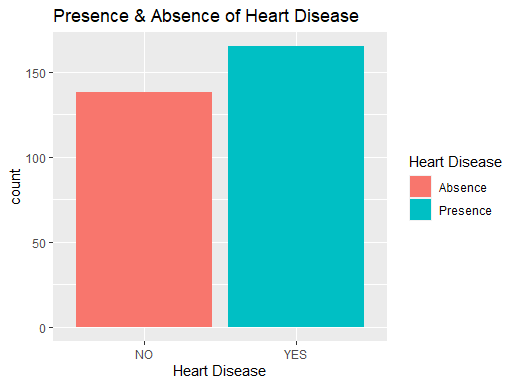
****

* **Data transformation OR clean the data:-**
* **mutate() adds new variables and preserves existing ones; transmute() adds new variables and drops existing ones**.

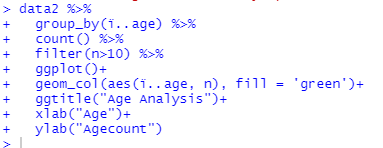
****

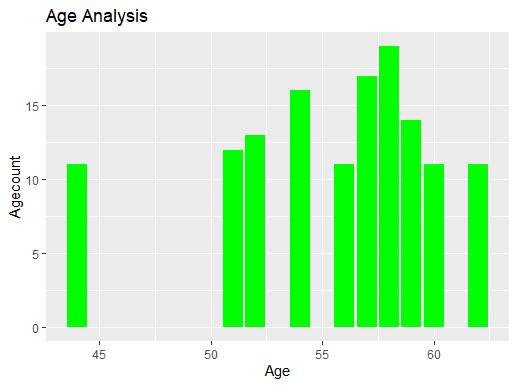
* **Bar plot for target (Heart disease):-**

****

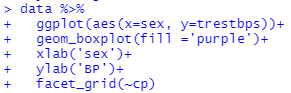
****

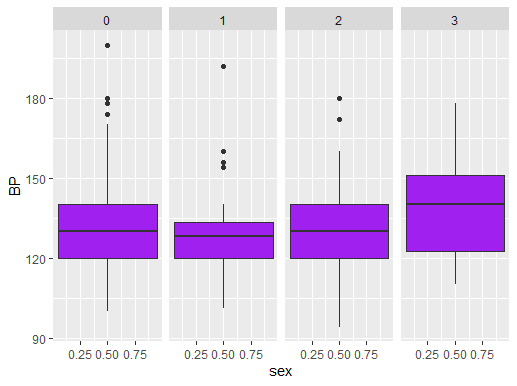
* **Count the frequencies of the values of age:-**

****

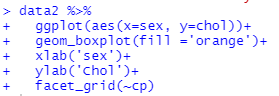
****

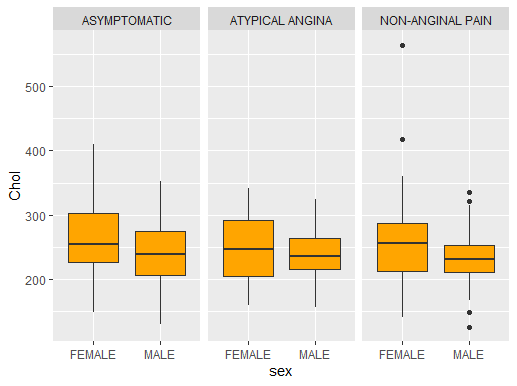
* **Comparison of SEX across BP:-**

****

****

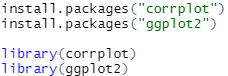
* **Comparison of Cholesterol across pain type:-**

****

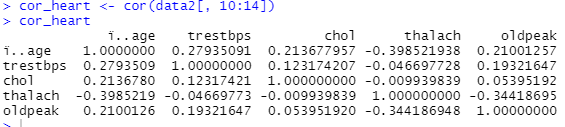
****

* **R package corrplot provides a visual exploratory tool on correlation matrix that supports automatic variable reordering to help detect hidden patterns among variables.**

**ggplot2 is a plotting package that provides helpful commands to create complex plots from data in a data frame.**

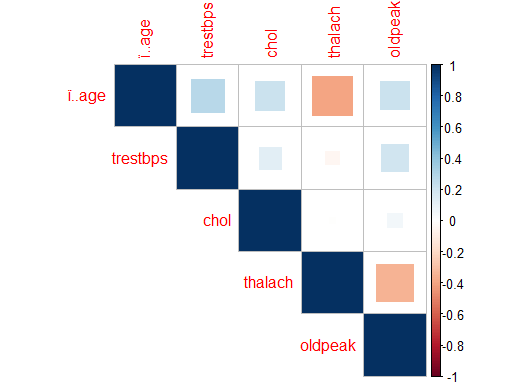
****

* **Corelation between different variables:-**

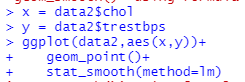
****

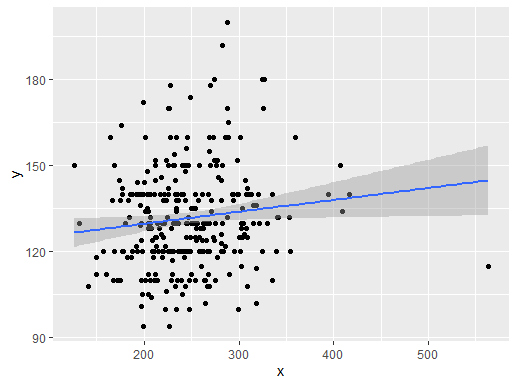
* **Visualization of corelation between variables:-**

****

****

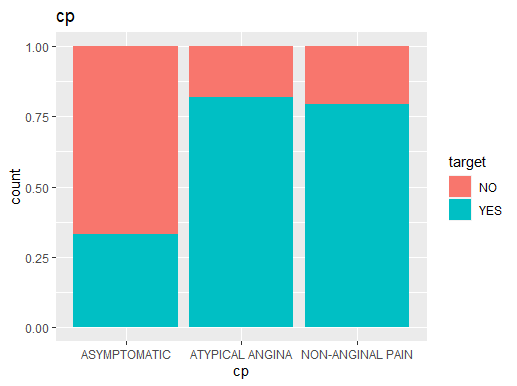
* **Linear regression:-**

****

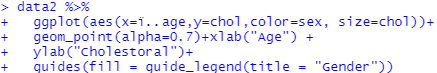
****

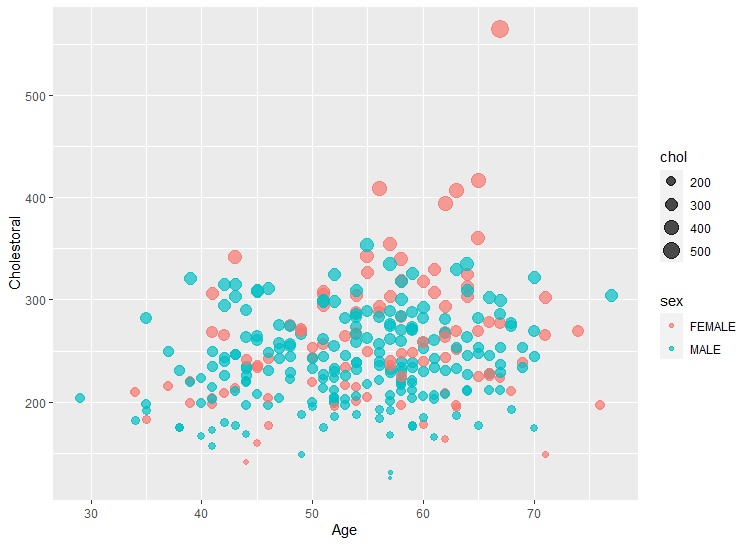
* **Analysis of different kind of chest pains (CP):-**

****

****

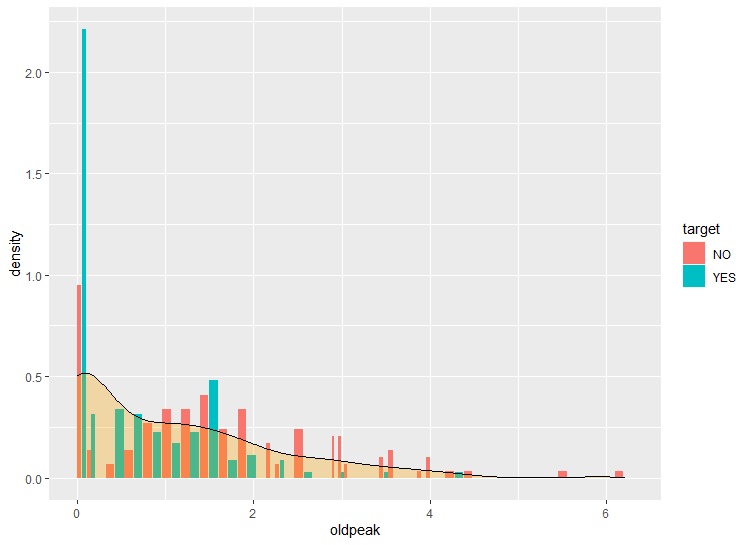
* **Representation of cholesterol level:-**



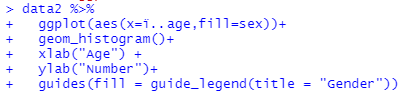


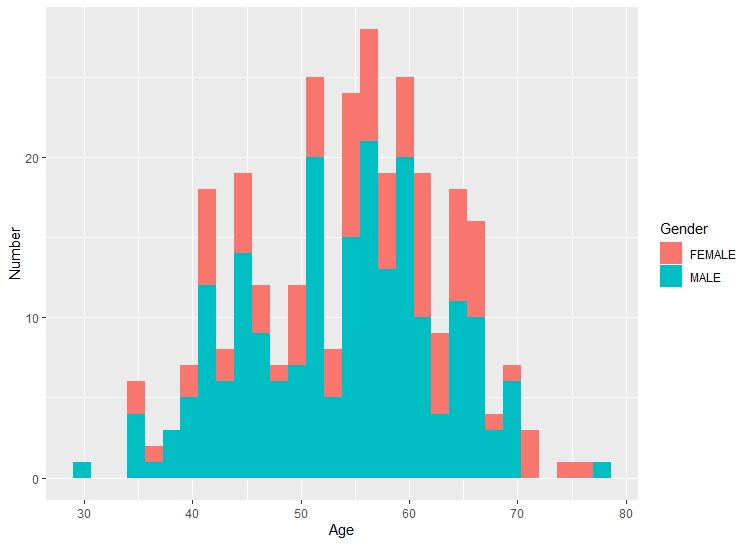
* **Histogram of old peak with density:-**





* **Histogram of gender wise age: -**





* **Conclusion: -**

To conclude, we can say the feature selection methods are multiple and will mainly depend on your data and what you seek to do with it. In the case of data with little variance, a regression ridge can give very satisfactory results. If you are looking to select variables, the lasso estimator can be your ally and help you identify the variables that are really important, if they are not correlated. More generally, the elastic-net method is probably the one that will give you the best results in many cases, but it is not necessarily the one that you should always consider first. It is important to look at more basic methods which may very well help you manage the bias-variance dilemma.