**Examining Factors Responsible for Heart Attacks**

Project 3

DESCRIPTION

Cardiovascular diseases are one of the leading causes of deaths globally. To identify the causes and develop a system to predict potential heart attacks in an effective manner is necessary. The data presented has all the information about relevant factors that might have an impact on cardiovascular health. The data needs to be studied in detail for further analysis.

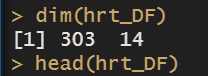
There is one dataset data that has 14 attributes with more than 4000 data points.

You are required to determine and examine the factors that play a significant role in increasing the rate of heart attacks. Also, use the findings to create and predict a model.

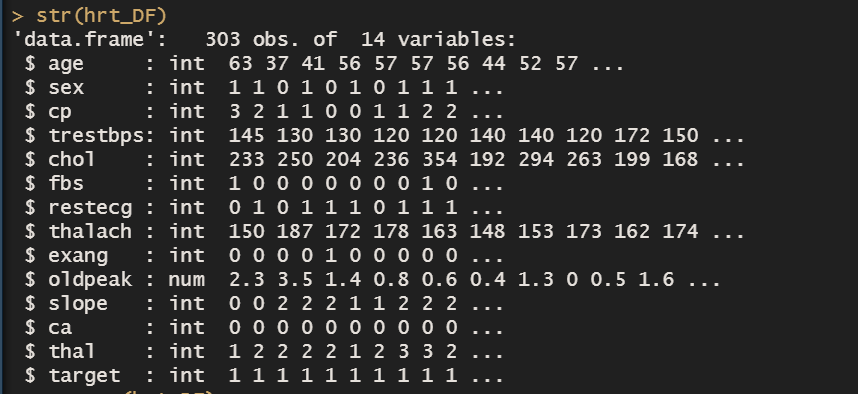
**Steps to perform:**

* **Importing, Understanding, and Inspecting Data :**

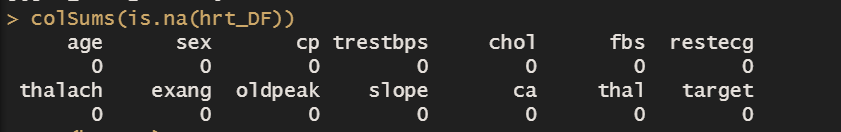
**Perform preliminary data inspection and report the findings as the structure of the data, missing values, duplicates, etc.**



Here we can see in this dataset the numbers of columns are 14 and there are 303 numbers of rows



The str function give us a structure of the heart dataset there are three numerical variable and rest of them are categorical nature.

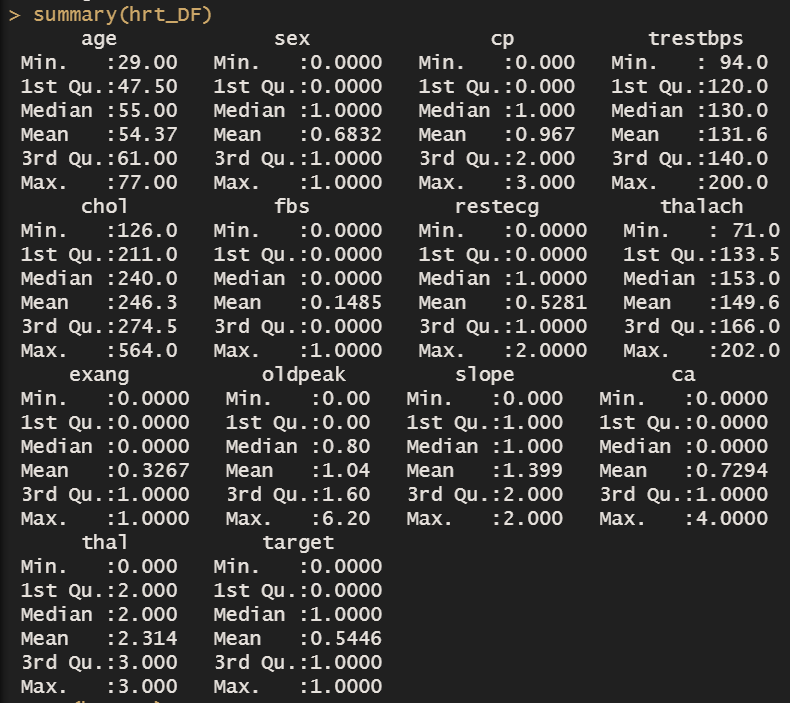


We can see there is no null values or missing values in this dataset, because the sum of null values is appeared as zero for every column. And as far as the duplicate values for concern here in dataset there is no as such identical variable on basis of which we can the different values for each row that’s why we can’t find the duplicate values for variable.

**Based on the findings from the previous question, remove duplicates (if any) and treat missing values using an appropriate strategy.**

Based on the findings of the previous question there is no missing values in the dataset and due to that there is no involvement of identical variable in dataset that’s why we can’t find a duplicate values.

**Get a preliminary statistical summary of the data. Explore the measures of central tendencies and the spread of the data overall.**



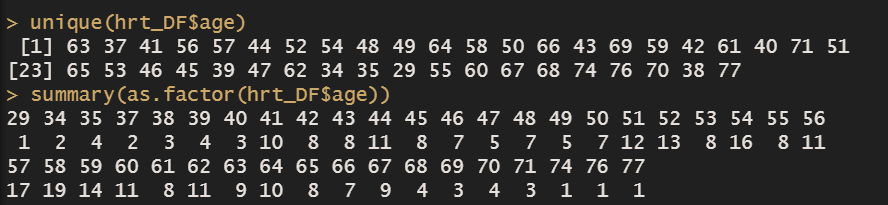
The most of the variable class is a factor and therefor the measure of dispersion is not clearly evident. There are only three variable which is numerical in nature which are thalach, chol and trestbps. To understand how the all variables are distributed for that dashboard has been created please kindly go through it.

**Performing EDA and Modelling:**

**Identify the data variables which might be categorical in nature. Describe and explore these variables using appropriate tools. For example: count plot.**

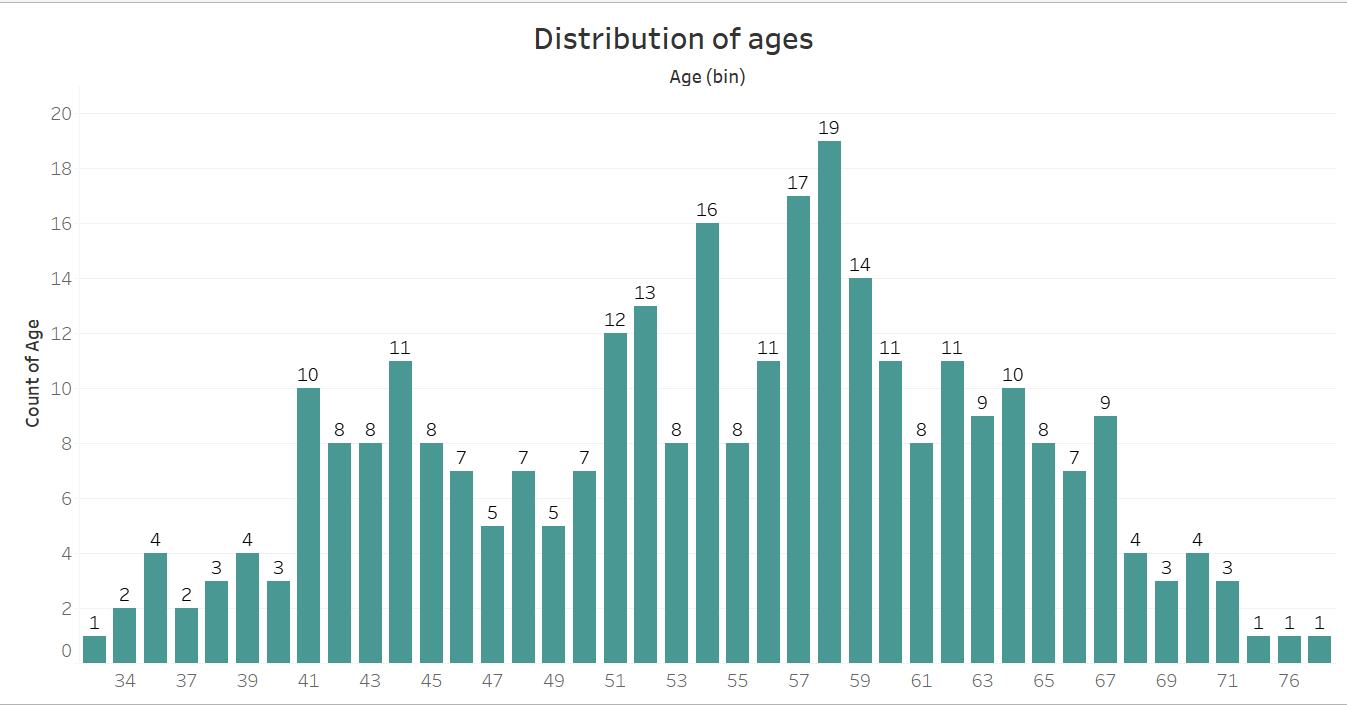
To understand the complete overview of data type I have created the dashboard which represent the visualising format of the data type and it will show the picture about how the data is distributed with the help of count plot and other required chart and plots.

**Study the occurrence of CVD across different ages.**



There are 41 age groups from the range of 29 to 77. The unique function tells the detail unique values lies in this particular variable, from which we get the unique values which is located on the top of this picture. To understand which group contains how much number of values we used the summery function, it shows the total count of numbers contained by a particular group.

As per the query the agency wants to find the occurrence of CVD by every age group and the bar chart given below shows the clear answer for this query.

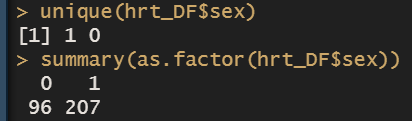


The count of ages is mention on the y axis and the bins plotted on the x axis. The age group of 57 to 59 is having the highest records compare to other age group.

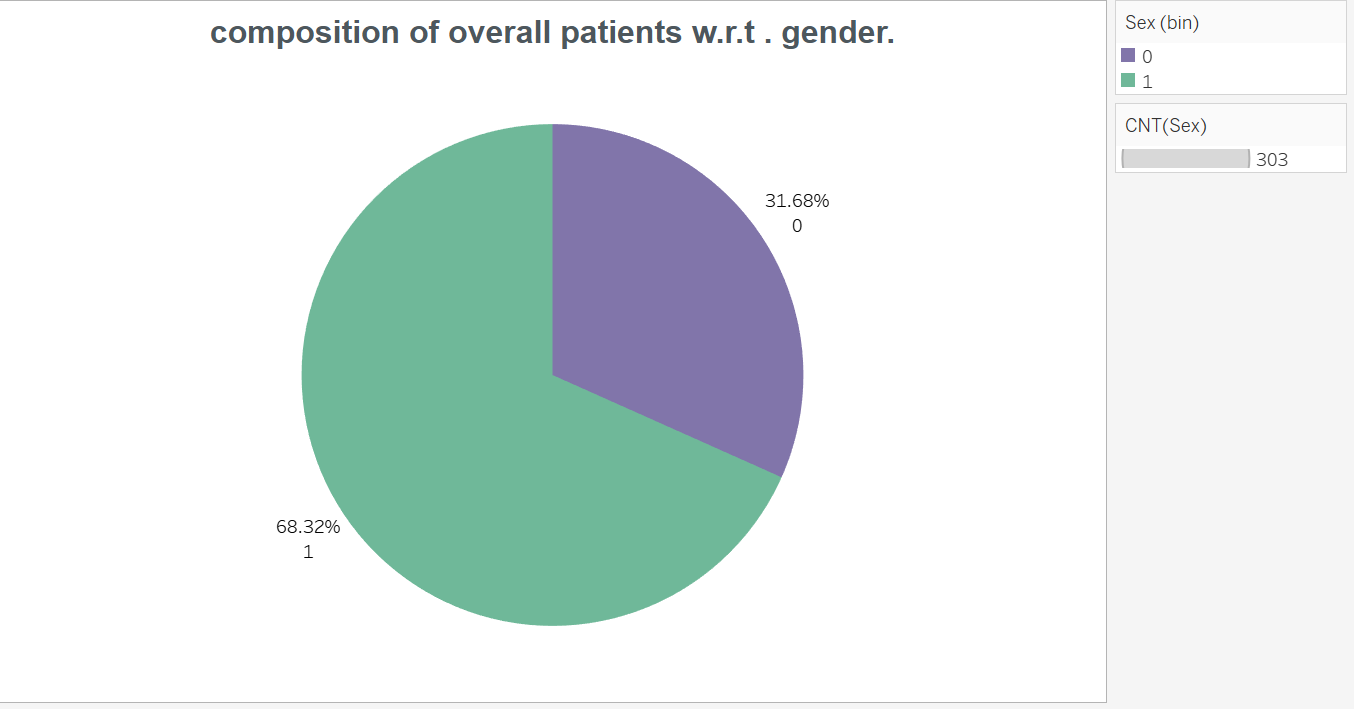
**Can we detect heart attack based on anomalies in resting blood pressure of the patient?**

The correlation between these two variables is very low thus the resting blood pressure is not the variable which is sufficient for measuring the heart attacks across the patients. The correlation between the target variable and trestbps is highly negative because the value of the correlation -0.144 that means the higher the chol decreases the target variable move likely towards zero. Please go through the logistic regression model in that model we can see that the trestbps affecting the target variable in negative manner because the beta coefficient of cholesterol is -0.0267 and on other hand the p-value is lesser than the 0.05 which means this variable is significant in nature. but when we look at it carefully, we can understand that there are other variables also which is more significant than the trestbps that’s why only on the basis of trestbps we can not detect the heart attack.

**Study the composition of overall patients w.r.t. gender.**

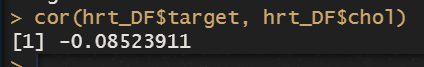


Here the 0 represent for female and 1 represent for the male according to the number which we get from summary function the data has highest numbers records registered for the male patients. The pie chart given below tells the patients records as per percentage format.



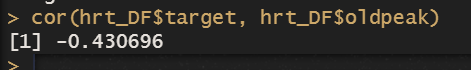
As mention earlier the male patient’s records have higher value compare to female patients. The 68.32% of patients are male whereas the remaining patients are female.

**Describe the relationship between cholesterol levels and our target variable.**

****

The correlation between the target variable and cholesterol level is highly negative because the value of the correlation -0.08 that means the higher the chol decreases the target variable move likely towards zero. Please go through the logistic regression model in that model we can see that the cholesterol level affecting the target variable in negative manner because the beta coefficient of cholesterol is -0.0028 and on other hand the p-value is greater then the 0.05 which means this variable is not significant nature.

**What can be concluded about** **the relationship between peak exercising and occurrence of heart attack?**

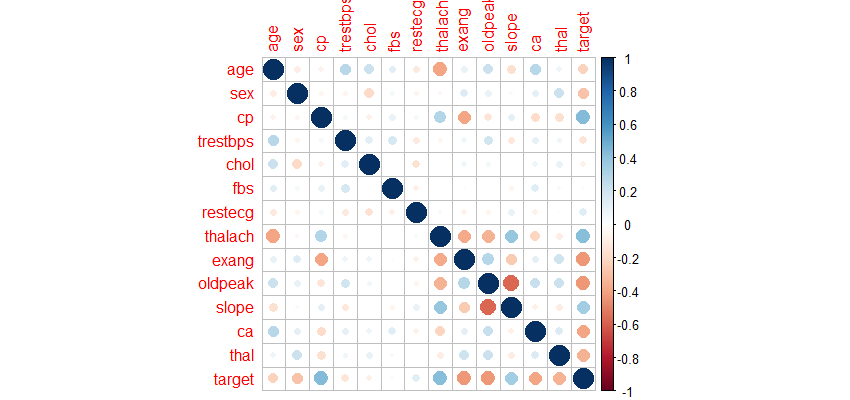


the relationship between peak exercising and occurrence of heart attack can be explain by the help of correlation method, The correlation between the target variable and oldpeak is highly negative because the value of the correlation -0.430 that means the higher the chol decreases the target variable move likely towards zero. Please go through the logistic regression model in that model we can see that the cholesterol level affecting the target variable in negative manner because the beta coefficient of old peak is -0.319 and on other hand the p-value is greater than the 0.05 which means this variable is not significant nature.

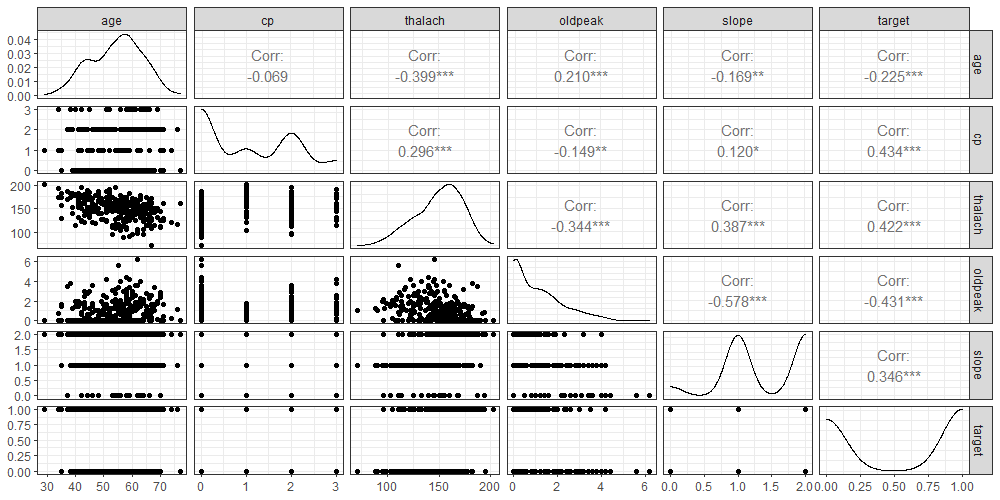
**Is thalassemia a major cause of CVD?** **How are the other factors determining the occurrence of CVD?**

In this dataset the thal variable represent the thalassemia, due to its negative correlation with target variable, the impact of thal to occurrence of CVD represents the inverse relation whereas the coefficient for the thalassemia is having negative value, this means that an increase in thalassemia’s coefficient which is negative in nature will be associated with a decreased probability for occurrence of heart attack. Therefor the thalassemia is not a major cause of CVD. To understand the how are the other factors determining the occurrence of CVD, kindly go through the analysis which is made by using the logistic regression model, through that model I have tried to solve the second problem statement of the above query.

**create a pair plot to understand the relationship between all the given variables.**



To create a pair plot I only used few variable because majority of variables are not correlated with each other as we can see there are only few variable like target, cp, thalach, slope, old peak, and ages are having some kind of relation with each other that’s why I took only these variables for pair plot.



The above plot tells the results of correlation in numerical format and on the other hand the relation of the variable is plotted with the help of scatter plot. From the above results the CP has a higher correlation with Target variable and rest of other variable has less correlated with Target variable values or weak correlation.

The above plot tells the results of correlation in numerical format and on the other hand the relation of the variable is plotted with the help of scatter plot. From the above results the LOS has a higher correlation with TOTCHG and rest of other variable has a negative values or weak correlation

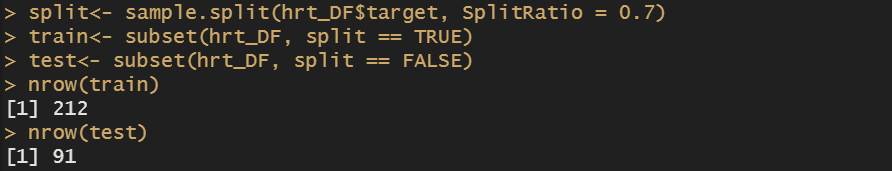
By considering the above results the cp and Target variable has correlation of 0.434 which means the relation between these two variables is positive in nature and the correlation between them is highest among all other correlation so it is clearly visible that the Target variable has been highly affected by the chest pain.

**Perform logistic regression, predict the outcome for test data, and validate the results by using the confusion matrix.**

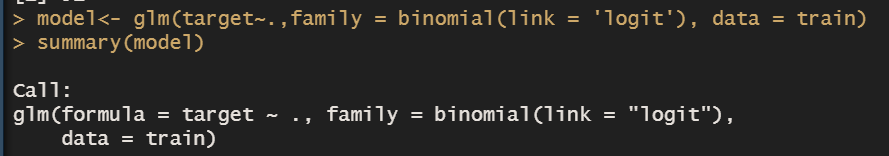
**Logistic regression** is used to predict the class (or category) of individuals based on one or multiple predictor variables (x). It is used to model a binary outcome, that is a variable, which can have only two possible values: 0 or 1, yes or no, diseased or non-diseased. And here we have to analyse a prediction of heart diseased and for that calculation we have target variable which has outcome as 0 and 1, here 0 represents to no and one represents to yes, therefor to calculate a end result the logistic regression model is perfect for this case.

Logistic regression works for a data that contain continuous or categorical predictor variables. here we are using heart dataset for predicting the probability of occurrence of heart attacks based on multiple clinical variables.

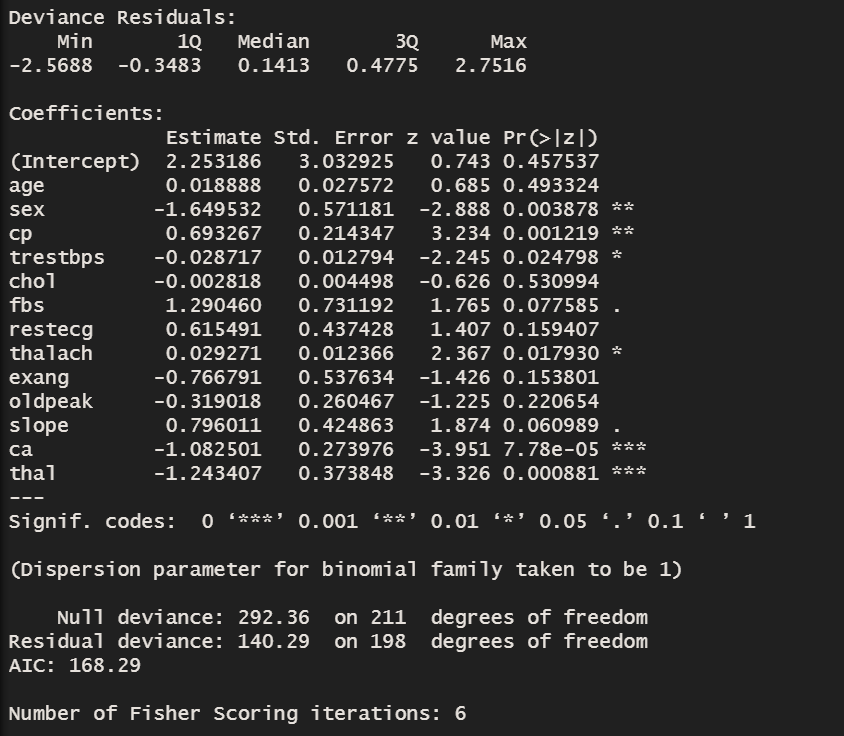
And for that I have randomly split the data into training set (70% for building a predictive model) and test set (30% for evaluating the model).



After splitting the data into 2 part the model has been created for that I used the glm function and took the dependent variable as target and for response variable I took the training dataset



Here, I have included all the predictor variables available in the data set. This is done using ~.: The following are the outcome we have got after applying the above formula.



From the output above, the coefficients table shows the beta coefficient estimates and their significance levels. Columns are:

* Estimate: the intercept (b0) and the beta coefficient estimates associated to each predictor variable
* Std.Error: the standard error of the coefficient estimates. This represents the accuracy of the coefficients. The larger the standard error, the less confident we are about the estimate.
* z value: the z-statistic, which is the coefficient estimate (column 2) divided by the standard error of the estimate (column 3)
* Pr(>|z|): The p-value corresponding to the z-statistic. The smaller the p-value, the more significant the estimate is.

Interpretation:

It can be seen that only 6 out of the 13 predictors are significantly associated to the outcome. These include: sex, cp, trestbps, thalach, ca, thal.

The coefficient estimate of the variable cp is b = 0.693267, which is positive. This means that an increase in cp is associated with increase in the probability for occurrence of heart attacks. However, the coefficient for the variables like trestbps, ca, thal, sex is having negative coefficient. This means that an increase in these variable coefficients which is negative in nature will be associated with a decreased probability for occurrence of heart attack.

Making prediction:

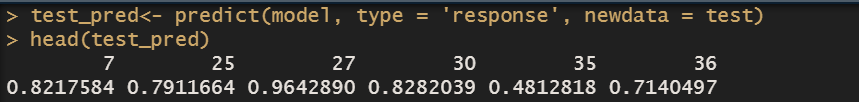
I have made a prediction using the test data in order to evaluate the performance of our logistic regression model.

The procedure is as follow:

1. I have predicted the target variable (probabilities for occurrence of heart attacks) based on predictor variables
2. I have assigned the observations to the target variable with highest probability score (i.e above 0.5)

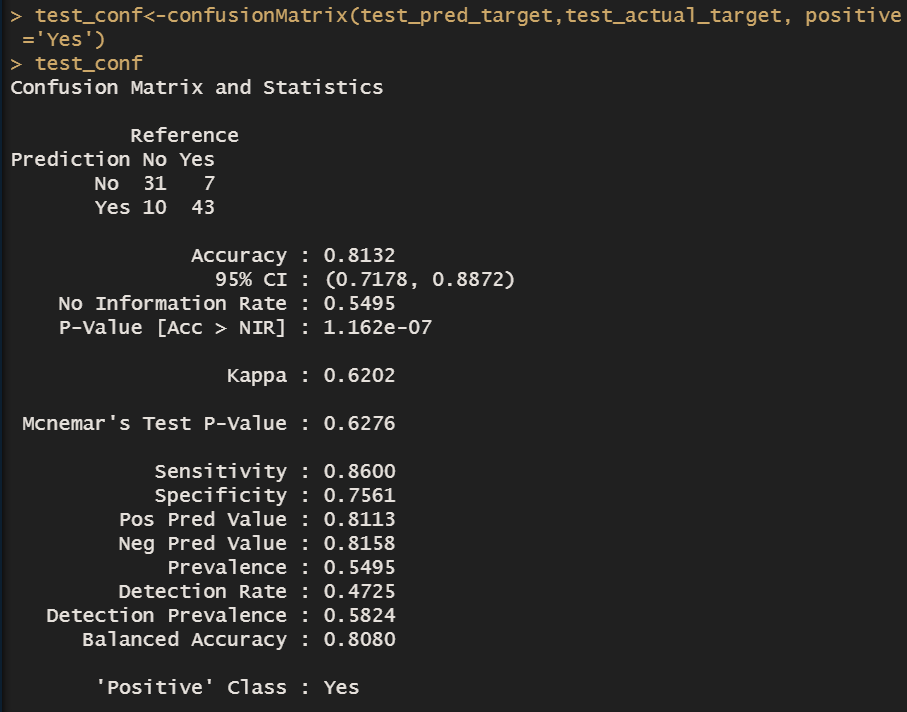
The R function predict() can be used to predict the probability of being diabetes-positive, given the predictor values.

**Prediction of probabilities** for occurrence of heart attack:



Evaluation of model by using confusion matrix:

Evaluating the accuracy of a classification model starts with summarizing the results into the following manner.



The above matrix represents the results of a model predicting if a patients suffer from heart attack based on the other response variable. The test data consists of 91 observations.

The matrix show us the prediction of truly positive which is 43 out off 91 and on the other hand it also gives the value of truly negative which is 31 out off 91 observation.  So, is this a good model? Let’s figure that out by examining the other aspect of the result.

Accuracy:

In the simplest terms, accuracy tells you generally how “right” your predictions are. It is the sum of true positives and negatives divided by the total population. In this case that’s 0.8132, meaning that 81.32% of the model’s predictions were correct.

Sensitivity:

Also known as recall or the true positive rate, sensitivity tells you how often the model chooses the positive class when the observation is in fact in the positive class. It is calculated by dividing the number of true positives in the matrix by the total number of real positives in the data.

In our example, sensitivity is 0.860, meaning that the model correctly predicts that a patient will suffer from heart attack 86% of the time.

Specificity:

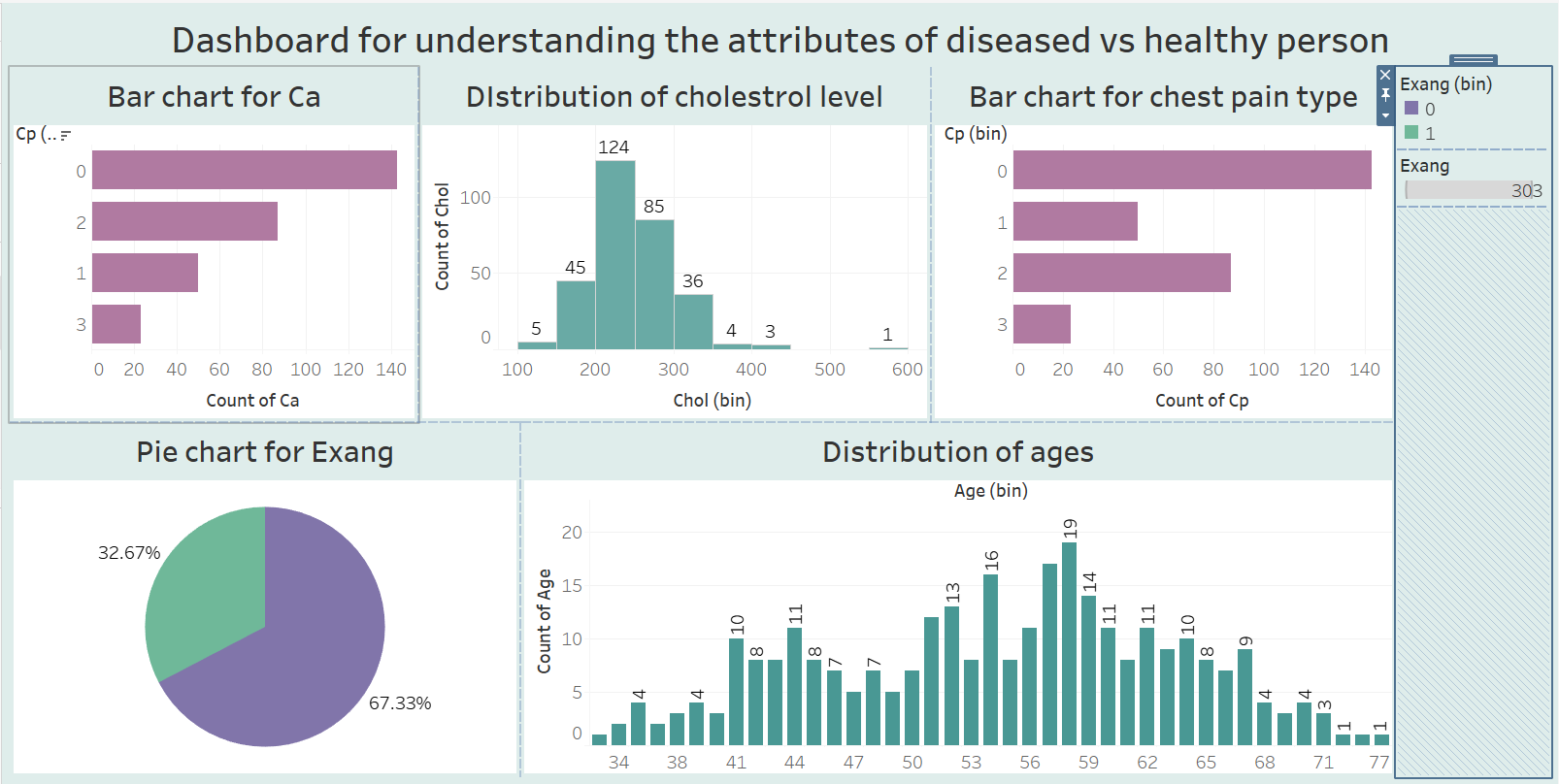
Also known as the true negative rate, specificity measures how often the model chooses the negative class when the observation is in fact in the negative class. It is calculated by dividing the number of true negatives by the total number of real negatives in the data.

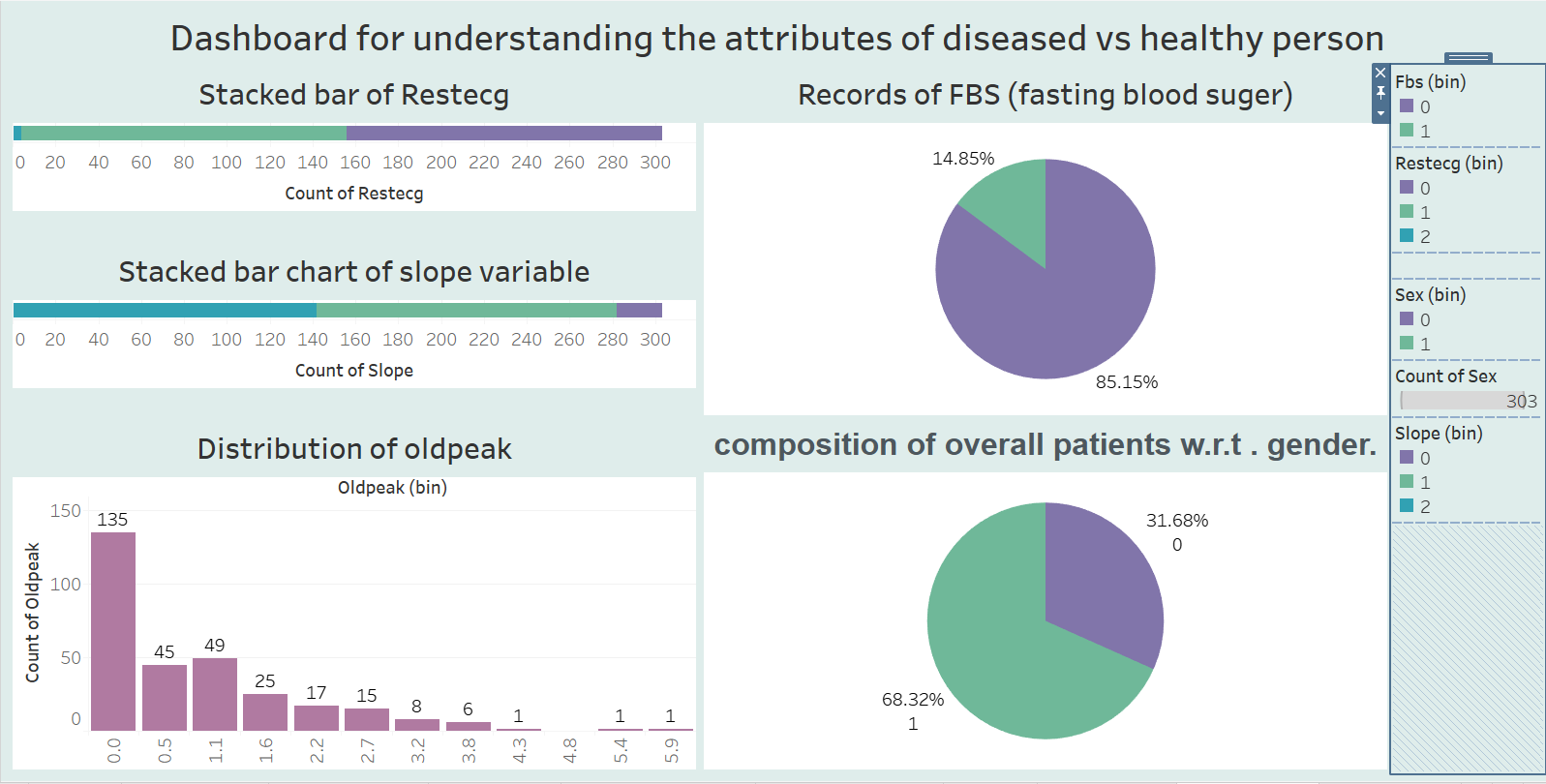
The true negative rate in our example is 0.7561, meaning that the model correctly predicts that a patient will not suffer from heart attack 75.61% of the time.

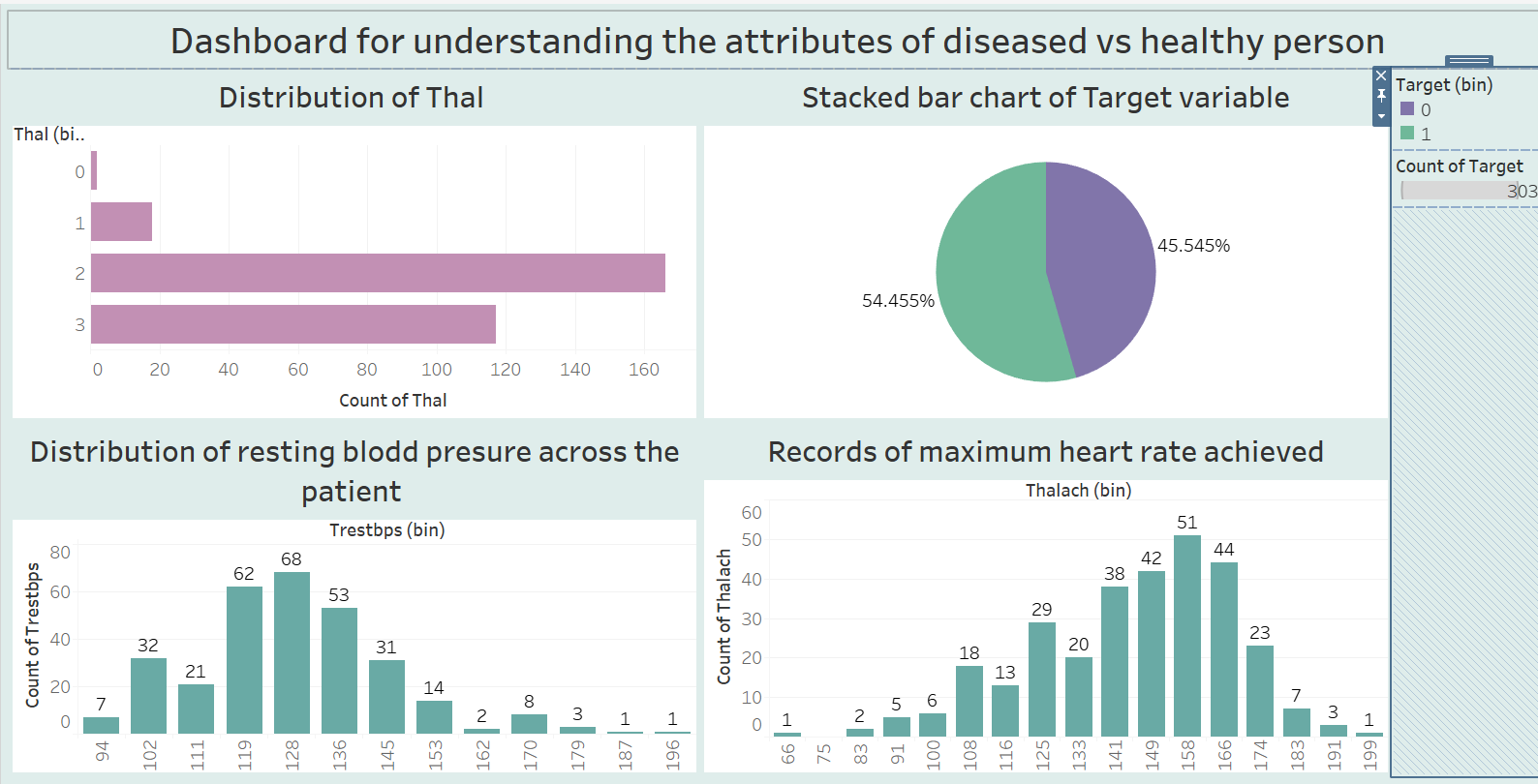
* **Dashboarding:**

**Visualize the variables using Tableau to create an understanding for attributes of a Diseased vs. a Healthy person.**

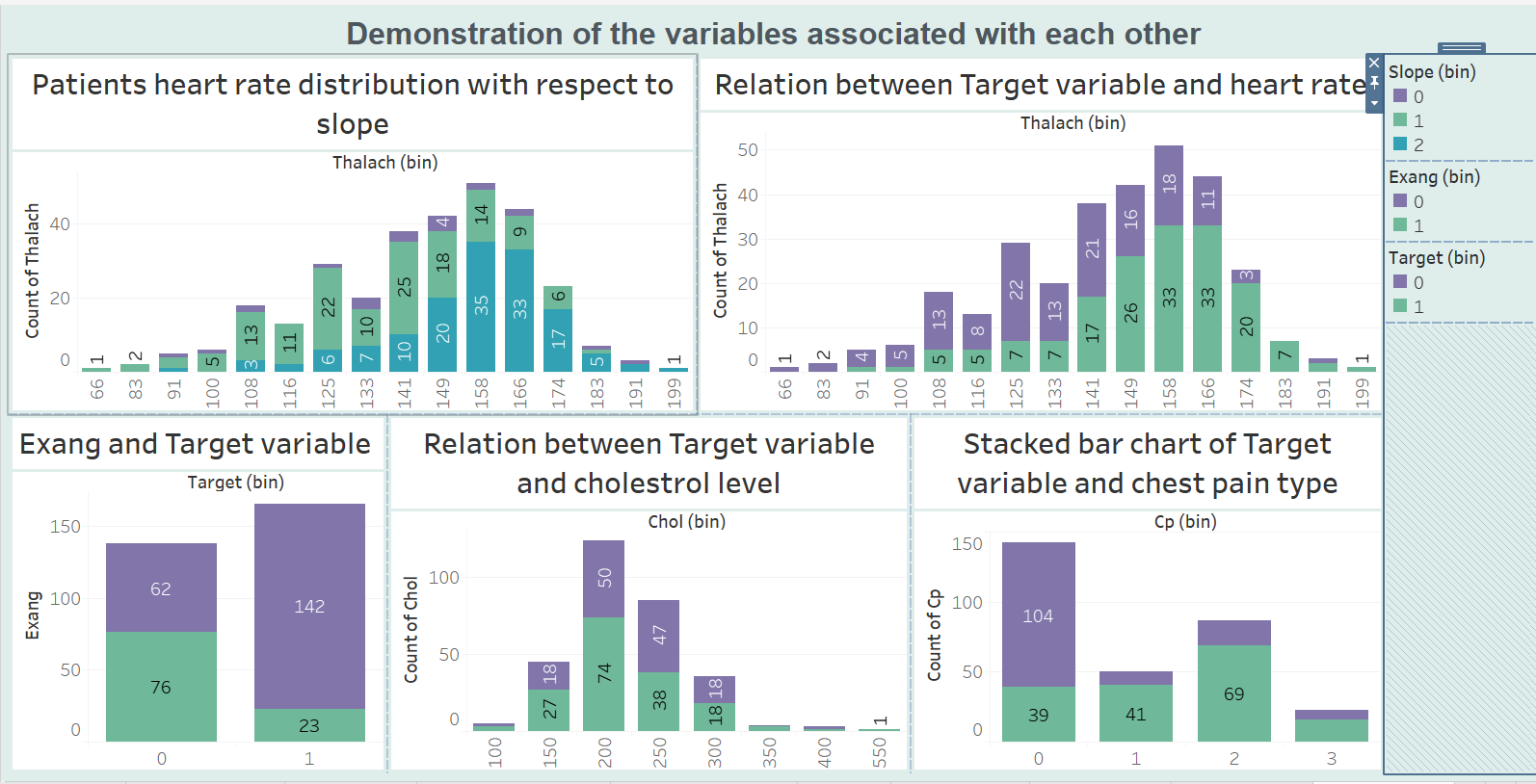
Three dashboards are created which tells the story of entire dataset the agency wants to find the healthy vs diseased person comparison. The following dashboard provide all the visualization about how the particular variable is distributed, which group contains the highest information. Through the EDA we can have an answer for the queries like composition of overall patient with respect to gender, CVD occurrence as per the ages. Please kindly go through the following dashboard to get an answer for queries mention above







**Demonstrate the variables associated with each other and factors to build a dashboard.**



The dashboard shown above is created by the variable which is correlated with each other and apart from that the remaining chart are created to give answer to those queries which was mention earlier. For instance, agency wants to find a relation between the exang and target variable, and here is the bar chart which represent the relation between them. The second question was asked frequently which variable is affecting the target variable and we saw through the corrplot that the cp is affecting the target variable. The dashboard also has one chart which shows the count of chest pain as per its category with respect to target variable, here the green represents to yes and purple represent to no. and third and last one chart which is created as per query requirement that is target variable and cholesterol level, rest of other chart has been created because the variable which is included in the diagram is highly correlated with each other.

.