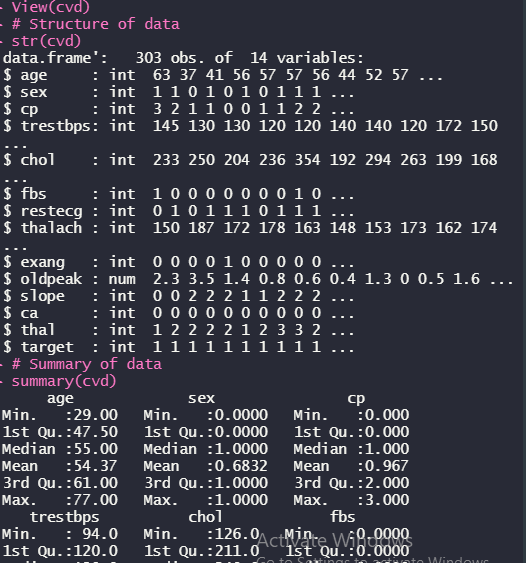
* Uploading the file and getting the structure of file:

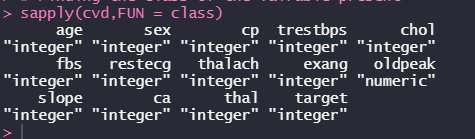
**Project 3: Examining Factors Responsible for Heart Attacks**



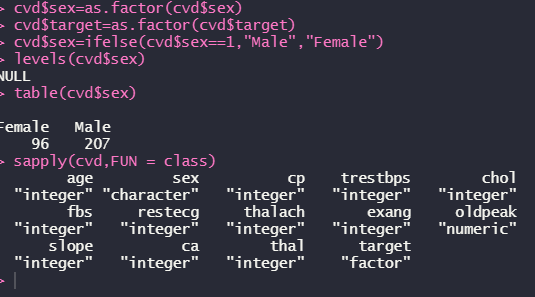
* Finding the mission value if any:

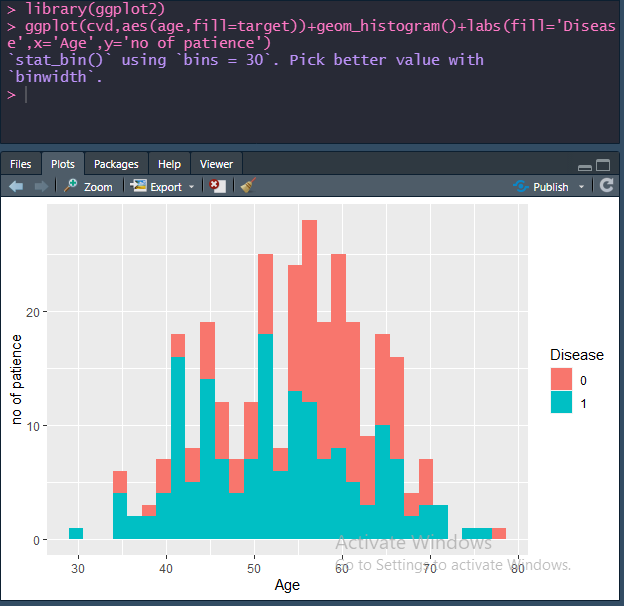


Finding the functions of the variable:



* Converting the categorical target variable into factor.



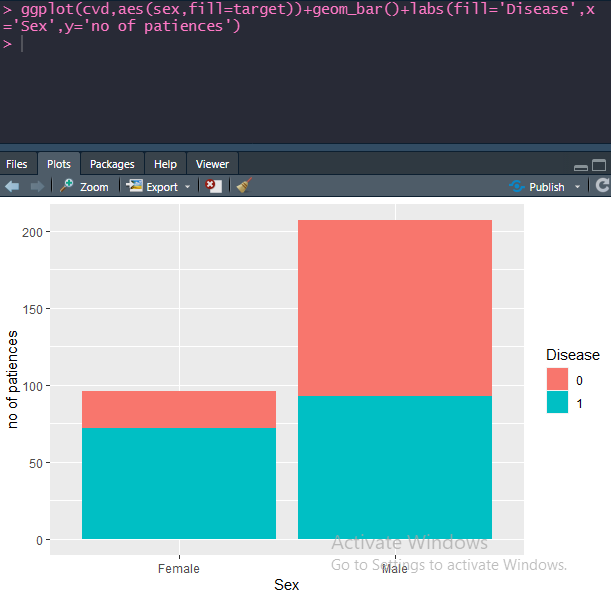
* **Studying and  examine the factors that play a significant role in increasing the rate of heart attacks**.
* **Age:**  Patient age in years. In the data we can see, as expected, that age is a risk factor. In other words, the higher the age, the more likely that the patient has a heart disease.

**Sex:**

Patient sex

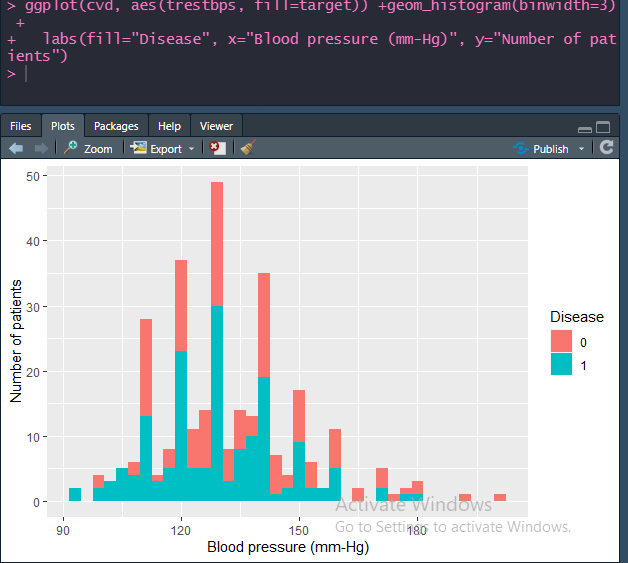
* Value 0: female
* Value 1: male

There are approximately half the observation of women than men. We can also see that sex is a risk factor, like some of the references indicate, men are more likely to have a heart disease than women.

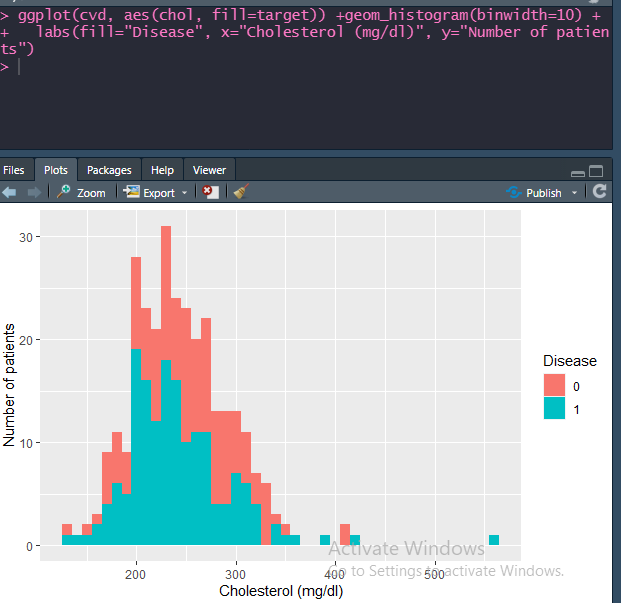


### Trestbps: Resting blood pressure in millimeters of mercury (mm Hg) when the patient was admitted to the hospital.

By the different peaks, looks like most people tend to have a normal blood pressure inside certain groups (could be healthy adults, adults that take medication, seniors...). It also looks like very high pressures can indicate that there is a heart disease.



### Chol: Cholesterol level in mg/dl. This is a variable that we can control to prevent the disease. Looks like the majority of the people in the dataset have high levels of cholesterol. It also looks like up to a certain level, the presence of a heart disease is slightly higher on higher cholesterol levels. Though the cases that have the highest levels of cholesterol don't have a heart disease, it could be that these people weren't fasting when the blood sample was taken.



### Oldpeak: Decrease of the ST segment during exercise according to the same one on rest.The ST segment is a part of the electrocardiogram of a heart beat that is usually found at a certain level in a normal heart beat. A significant displacement of this segment can indicate the presence of a heart disease as we can see in the plot.



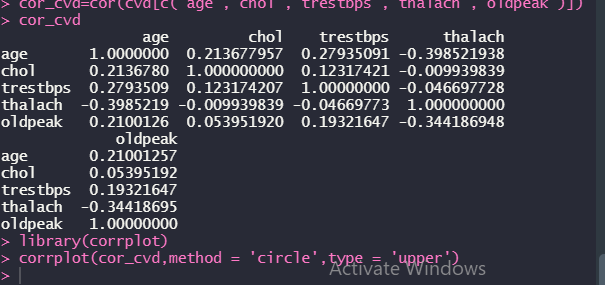
### Thal : Results of the blood flow observed via the radioactive dye.

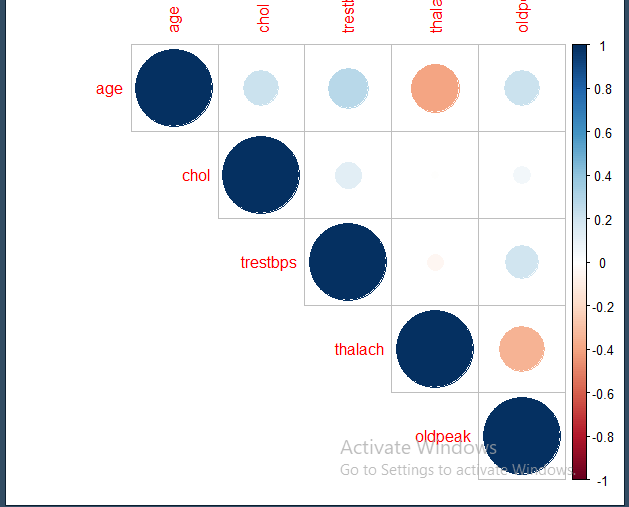
* Value 0: NULL (dropped from the dataset previously)
* Value 1: fixed defect (no blood flow in some part of the heart)
* Value 2: normal blood flow
* Value 3: reversible defect (a blood flow is observed but it is not normal)

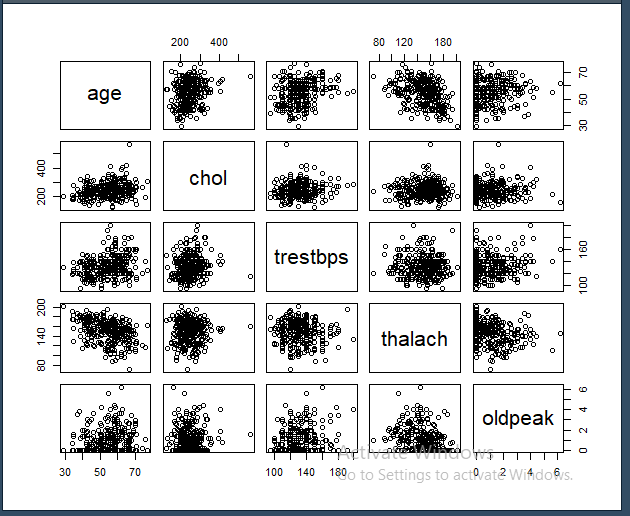
This feature and the next one are obtained through a very invasive process for the patients. But, by themselves, they give a very good indication of the presence of a heart disease or not.



**Creating a corrplot and pair plot to better understand the variables:**

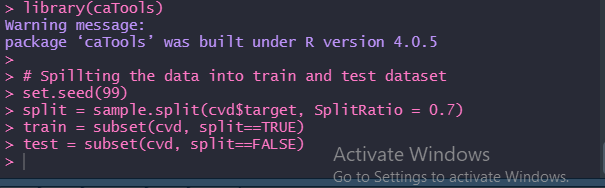




Ce

Creating the model:

Splitting the dataset into train and test.



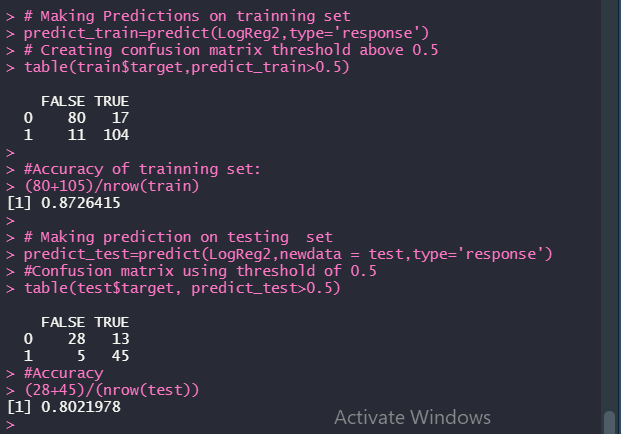
### Logistic Regression:

### 

### Here a lot of variable is not signification like age, so again tuning the Regression model with the significant variable.

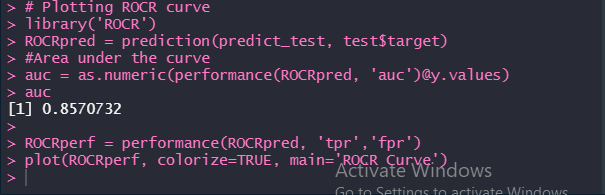
* All variable are significant in this model.
* Predictions:

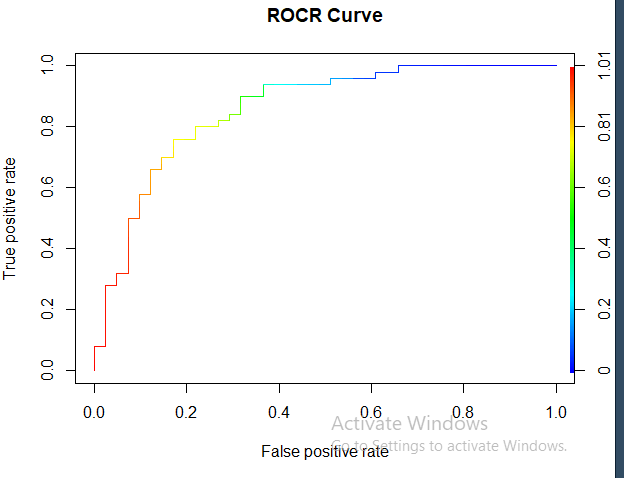
-🡪 Predictions on Training Set and Confusion matrix using threshold of 0.5



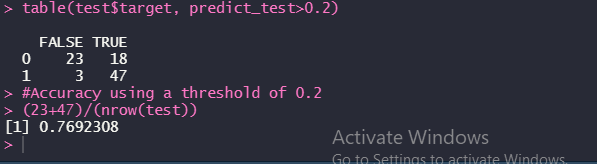
* **Accuracy on training set :** ***0.867924528301887***
* **Accuracy on test test: *0.802197802197802***

**Plotting ROCR curve**

****

****

* From ROCR curve threshold of 0.2 seems to be okay so that true positives are maximized such that maximum number of patients with heart disease are not identified as healthy.
* Confusion matrix with threshold>0.2:

****