ALY6015

Intermediate Analytics

Group Members

Kinza Alam
Jargi Desai
Vrunda Patel

Final project: Report

Instructor – Rechard He

Date: 07/03/2021

College of Professional Studies, Northeastern University, Boston, MA
02215

Introduction

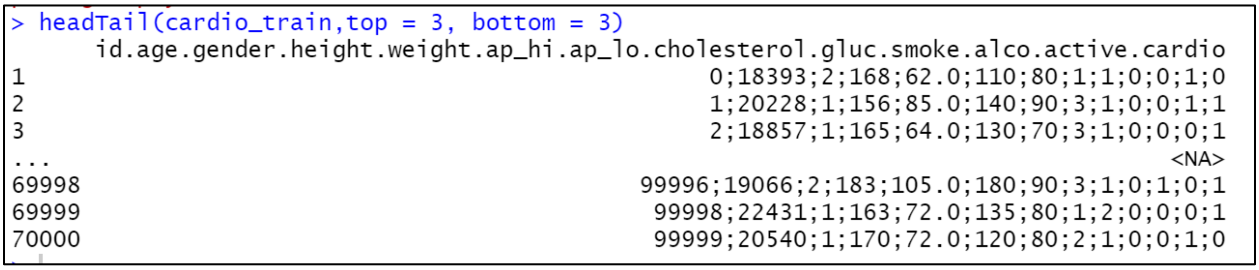
"Cardiovascular Disease Dataset" is dataset from Kaggle.com. There are 70,000 patients in record. This dataset also consists of 13 different variables which are- Id, Age, Height, Weight, Gender, Systolic blood pressure, Diastolic blood pressure, Cholesterol, Glucose, Smoking, Alcohol intake, Physical activity, Presence, or absence of cardiovascular disease(cardio).

In this project, we will be performing different methods of classification to analyze and process data which will help us to determine effect different factors on presence or absence of cardiovascular disease in patients.

Objective

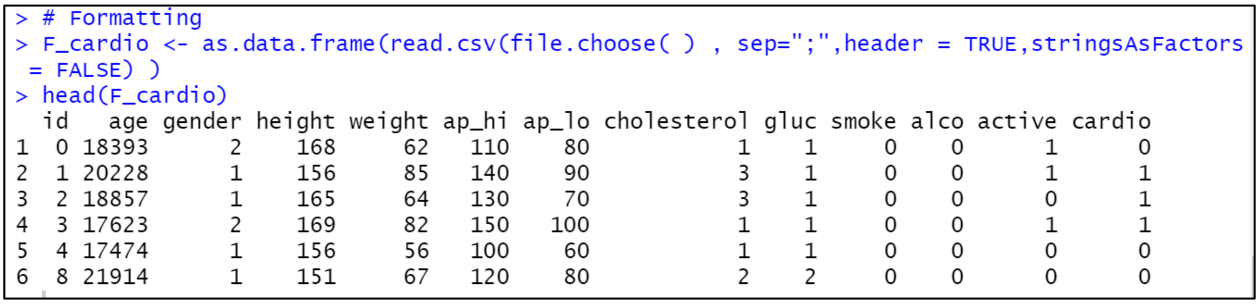
The goal of this research is to conduct an exploratory data analysis of cardiovascular patients using dataset. The following is a breakdown of the report's structure. The dataset and its properties are described in section 2. The data pre-processing is described in section 3. In section 4, we look at each attribute and how they relate to one another. Finally, in the final section, we conclude the analysis briefly.

First. We will load the data.



```
> headTail(cardio_train,top = 3, bottom = 3)
      id.age.gender.height.weight.ap_hi.ap_lo.cholesterol.gluc.smoke.alco.active.cardio
1                                                   0;18393;2;168;62.0;110;80;1;1;0;0;1;0
2                                                   1;20228;1;156;85.0;140;90;3;1;0;0;1;1
3                                                   2;18857;1;165;64.0;130;70;3;1;0;0;0;1
...                                                                                  <NA>
69998                                          99996;19066;2;183;105.0;180;90;3;1;0;1;0;1
69999                                           99998;22431;1;163;72.0;135;80;1;2;0;0;0;1
70000                                           99999;20540;1;170;72.0;120;80;2;1;0;0;1;0
```

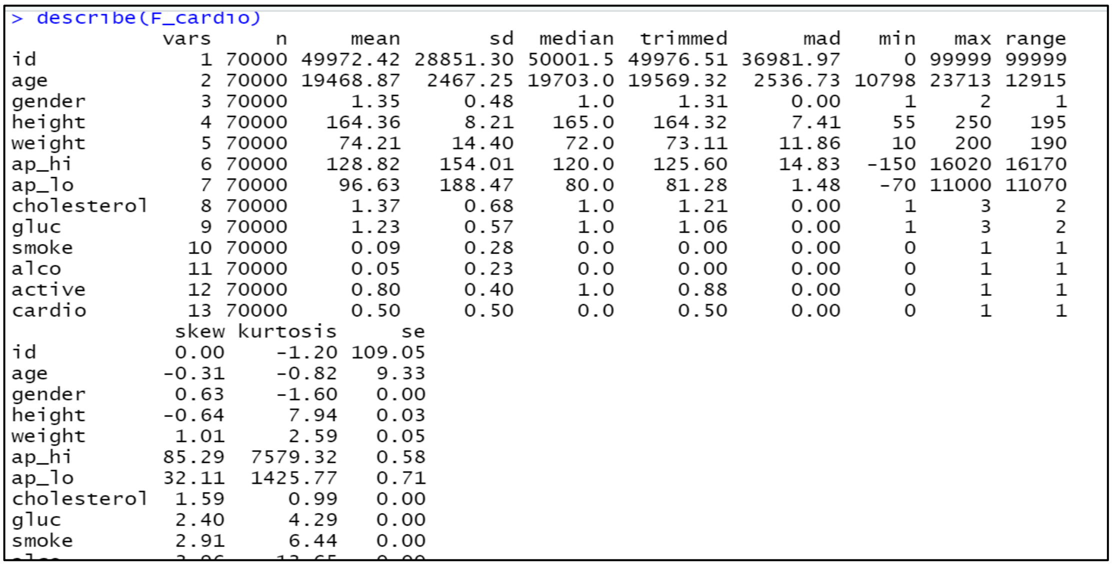
Since, the data is not in readable manner. We have done formatting. So, we can differentiate variables and its values.



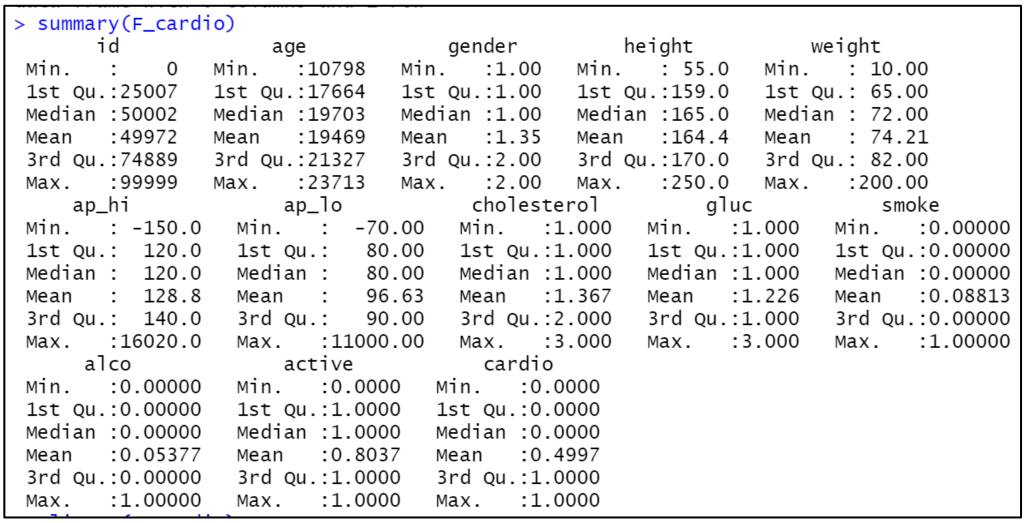
```
> # Formatting
> F_cardio <- as.data.frame(read.csv(file.choose( ) , sep=";",header = TRUE,stringsAsFactors
 = FALSE) )
> head(F_cardio)
  id   age gender height weight ap_hi ap_lo cholesterol gluc smoke alco active cardio
1  0 18393      2    168     62   110    80           1    1     0    0      1      0
2  1 20228      1    156     85   140    90           3    1     0    0      1      1
3  2 18857      1    165     64   130    70           3    1     0    0      0      1
4  3 17623      2    169     82   150   100           1    1     0    0      1      1
5  4 17474      1    156     56   100    60           1    1     0    0      0      0
6  8 21914      1    151     67   120    80           2    2     0    0      0      0
```

Data and its features

Checking the details of the variables.

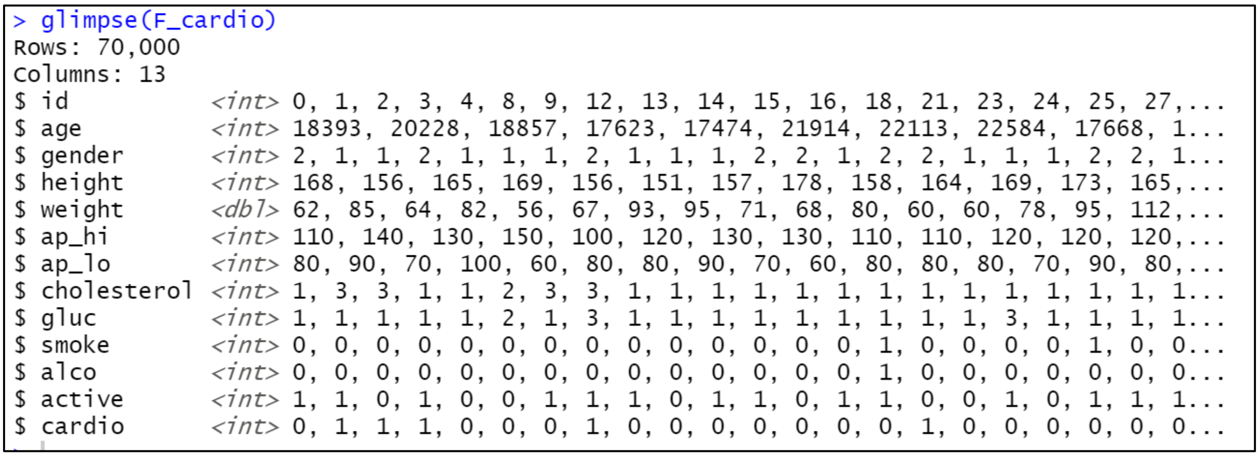


```
> describe(F_cardio)
            vars     n     mean       sd  median  trimmed      mad   min   max range
id             1 70000 49972.42 28851.30 50001.5 49976.51 36981.97     0 99999 99999
age            2 70000 19468.87  2467.25 19703.0 19569.32  2536.73 10798 23713 12915
gender         3 70000     1.35     0.48     1.0     1.31     0.00     1     2     1
height         4 70000   164.36     8.21   165.0   164.32     7.41    55   250   195
weight         5 70000    74.21    14.40    72.0    73.11    11.86    10   200   190
ap_hi          6 70000   128.82   154.01   120.0   125.60    14.83  -150 16020 16170
ap_lo          7 70000    96.63   188.47    80.0    81.28     1.48   -70 11000 11070
cholesterol    8 70000     1.37     0.68     1.0     1.21     0.00     1     3     2
gluc           9 70000     1.23     0.57     1.0     1.06     0.00     1     3     2
smoke         10 70000     0.09     0.28     0.0     0.00     0.00     0     1     1
alco          11 70000     0.05     0.23     0.0     0.00     0.00     0     1     1
active        12 70000     0.80     0.40     1.0     0.88     0.00     0     1     1
cardio        13 70000     0.50     0.50     0.0     0.50     0.00     0     1     1
             skew kurtosis     se
id           0.00    -1.20 109.05
age         -0.31    -0.82   9.33
gender       0.63    -1.60   0.00
height      -0.64     7.94   0.03
weight       1.01     2.59   0.05
ap_hi       85.29  7579.32   0.58
ap_lo       32.11  1425.77   0.71
cholesterol  1.59     0.99   0.00
gluc         2.40     4.29   0.00
smoke        2.91     6.44   0.00
```



```
> summary(F_cardio)
       id             age            gender          height          weight      
 Min.   :    0   Min.   :10798   Min.   :1.00   Min.   : 55.0   Min.   : 10.00  
 1st Qu.:25007   1st Qu.:17664   1st Qu.:1.00   1st Qu.:159.0   1st Qu.: 65.00  
 Median :50002   Median :19703   Median :1.00   Median :165.0   Median : 72.00  
 Mean   :49972   Mean   :19469   Mean   :1.35   Mean   :164.4   Mean   : 74.21  
 3rd Qu.:74889   3rd Qu.:21327   3rd Qu.:2.00   3rd Qu.:170.0   3rd Qu.: 82.00  
 Max.   :99999   Max.   :23713   Max.   :2.00   Max.   :250.0   Max.   :200.00  
     ap_hi              ap_lo             cholesterol         gluc            smoke        
 Min.   : -150.0   Min.   :  -70.00   Min.   :1.000   Min.   :1.000   Min.   :0.00000  
 1st Qu.:  120.0   1st Qu.:   80.00   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.00000  
 Median :  120.0   Median :   80.00   Median :1.000   Median :1.000   Median :0.00000  
 Mean   :  128.8   Mean   :   96.63   Mean   :1.367   Mean   :1.226   Mean   :0.08813  
 3rd Qu.:  140.0   3rd Qu.:   90.00   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:0.00000  
 Max.   :16020.0   Max.   :11000.00   Max.   :3.000   Max.   :3.000   Max.   :1.00000  
      alco             active           cardio      
 Min.   :0.00000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.0000  
 Median :0.00000   Median :1.0000   Median :0.0000  
 Mean   :0.05377   Mean   :0.8037   Mean   :0.4997  
 3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :1.00000   Max.   :1.0000   Max.   :1.0000  
```

Except for the ID column, this data contains 70000 observations with 12 descriptive attributes and 1 target. It is a binary classification problem because the target feature has two classes. It determines whether a person is suffering from cardiovascular disease.

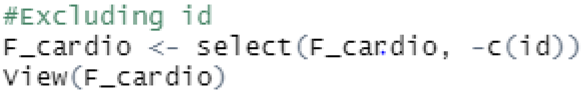


```
> glimpse(F_cardio)
Rows: 70,000
Columns: 13
$ id          <int> 0, 1, 2, 3, 4, 8, 9, 12, 13, 14, 15, 16, 18, 21, 23, 24, 25, 27,...
$ age         <int> 18393, 20228, 18857, 17623, 17474, 21914, 22113, 22584, 17668, 1...
$ gender      <int> 2, 1, 1, 2, 1, 1, 1, 2, 1, 1, 1, 2, 2, 1, 2, 2, 1, 1, 1, 2, 2, 1...
$ height      <int> 168, 156, 165, 169, 156, 151, 157, 178, 158, 164, 169, 173, 165,...
$ weight      <dbl> 62, 85, 64, 82, 56, 67, 93, 95, 71, 68, 80, 60, 60, 78, 95, 112,...
$ ap_hi       <int> 110, 140, 130, 150, 100, 120, 130, 130, 110, 110, 120, 120, 120,...
$ ap_lo       <int> 80, 90, 70, 100, 60, 80, 80, 90, 70, 60, 80, 80, 80, 70, 90, 80,...
$ cholesterol <int> 1, 3, 3, 1, 1, 2, 3, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
$ gluc        <int> 1, 1, 1, 1, 1, 2, 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 1...
$ smoke       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0...
$ alco        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0...
$ active      <int> 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1...
$ cardio      <int> 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0...
```

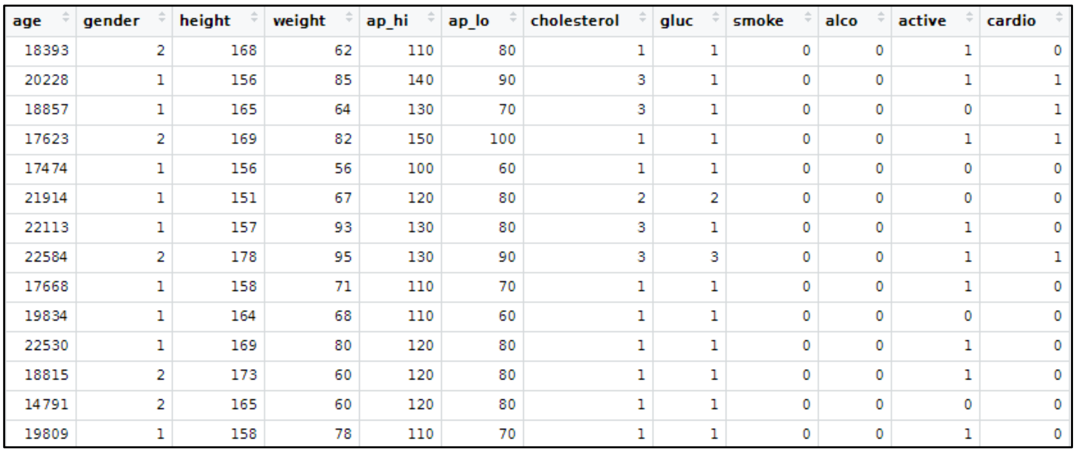
The following is the variable description:

Age : Age of the person in days Height : height of the person Weight : weight of the person Gender : gender of the person ap_hi : Systolic blood pressure ap_lo : Diastolic blood pressure Cholestrol : cholesterol level | 1: normal, 2: above normal, 3: well above normal | gluc : glucose level | 1: normal, 2: above normal, 3: well above normal | smoke : smoking | 0: No, 1: True | alco : Alcohol intake | 0: No, 1: True | active : Physical activity |0: No, 1: True |

Data Preprocessing



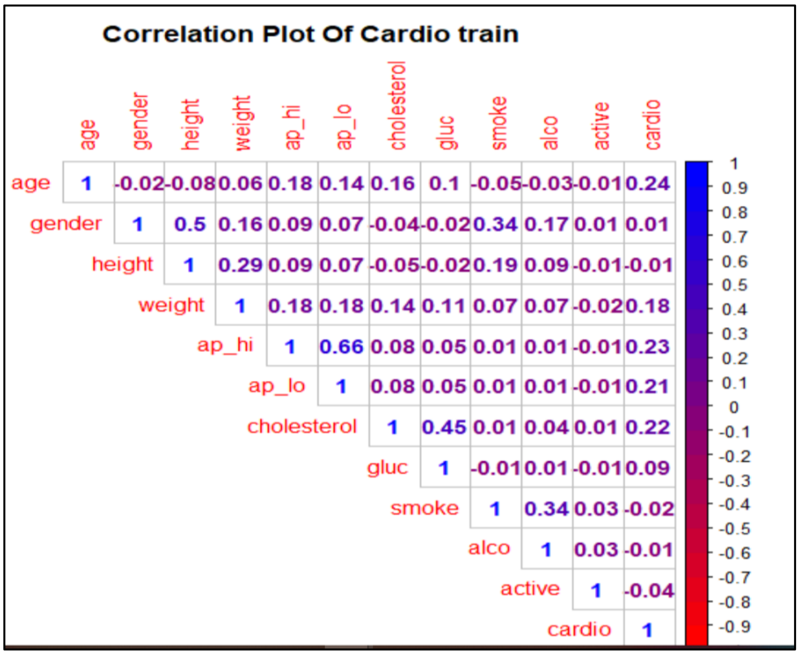
```
#Excluding id
F_cardio <- select(F_cardio, -c(id))
View(F_cardio)
```



age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
18393	2	168	62	110	80	1	1	0	0	1	0
20228	1	156	85	140	90	3	1	0	0	1	1
18857	1	165	64	130	70	3	1	0	0	0	1
17623	2	169	82	150	100	1	1	0	0	1	1
17474	1	156	56	100	60	1	1	0	0	0	0
21914	1	151	67	120	80	2	2	0	0	0	0
22113	1	157	93	130	80	3	1	0	0	1	0
22584	2	178	95	130	90	3	3	0	0	1	1
17668	1	158	71	110	70	1	1	0	0	1	0
19834	1	164	68	110	60	1	1	0	0	0	0
22530	1	169	80	120	80	1	1	0	0	1	0
18815	2	173	60	120	80	1	1	0	0	1	0
14791	2	165	60	120	80	1	1	0	0	0	0
19809	1	158	78	110	70	1	1	0	0	1	0

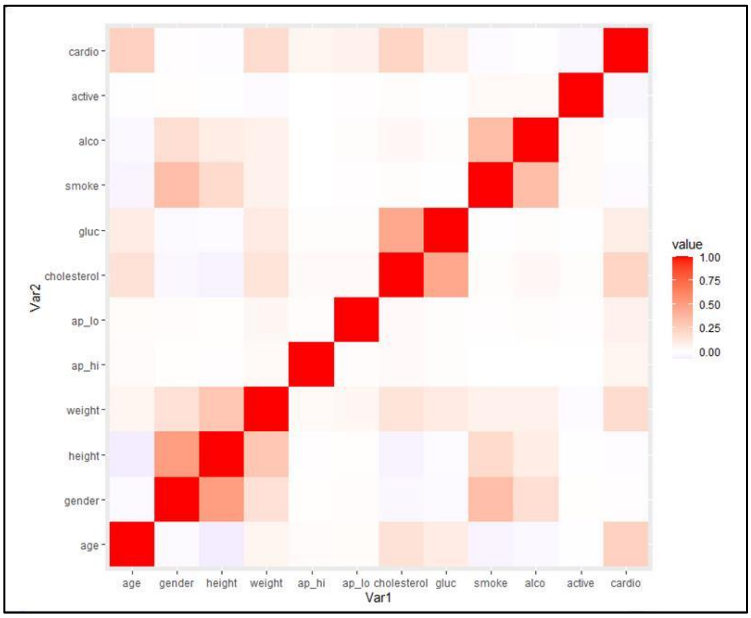
Correlation Matrix

We created a matrix plot of correlations, as shown in code chunck, to determine the correlation of each variable.



	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
age	1	-0.02	-0.08	0.06	0.18	0.14	0.16	0.1	-0.05	-0.03	-0.01	0.24
gender		1	0.5	0.16	0.09	0.07	-0.04	-0.02	0.34	0.17	0.01	0.01
height			1	0.29	0.09	0.07	-0.05	-0.02	0.19	0.09	-0.01	-0.01
weight				1	0.18	0.18	0.14	0.11	0.07	0.07	-0.02	0.18
ap_hi					1	0.66	0.08	0.05	0.01	0.01	-0.01	0.23
ap_lo						1	0.08	0.05	0.01	0.01	-0.01	0.21
cholesterol							1	0.45	0.01	0.04	0.01	0.22
gluc								1	-0.01	0.01	-0.01	0.09
smoke									1	0.34	0.03	-0.02
alco										1	0.03	-0.01
active											1	-0.04
cardio												1

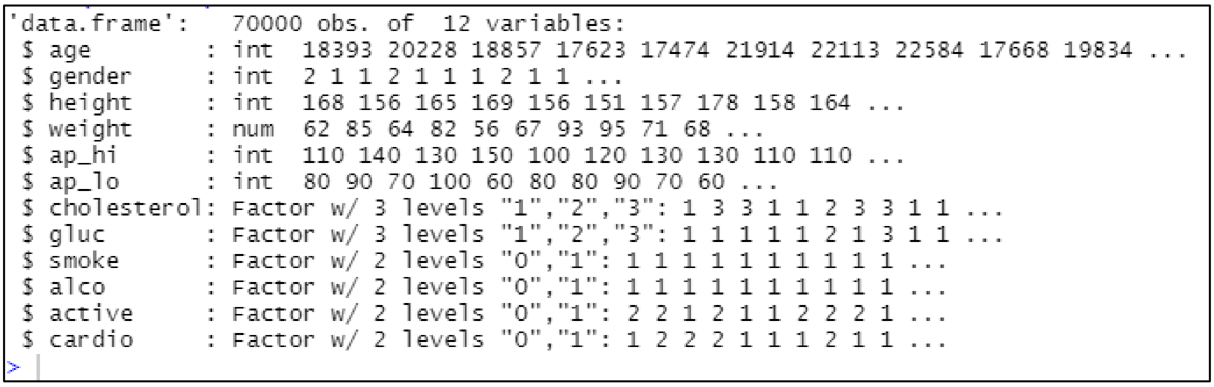
Correlation plot



From the plot above, we can see that there is high correaltion between cholestrol and glucous, height and gender, cardio and cholestrol, cardio and age, cardio and weight and alcohol and smoking which are highlighted through sligh red color. Light red color show low correlation such as between age and weight and lowest correlation is identified through shade of white such as gender and age, etc.

Factor	High correalation	Low Correalation
Cardio	Age, Weight, Cholesterol	Gender, Height, Smoke, Alcohol

The data collection includes categorical variables such as cholesterol, glucose, smoking, physical activity, and gender. As seen below, these variables are converted to factors.

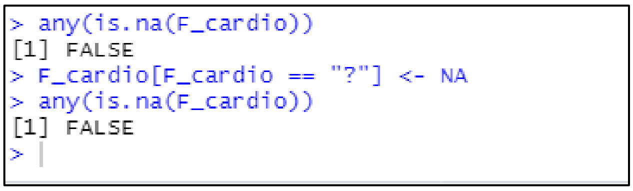


```
'data.frame':   70000 obs. of  12 variables:
 $ age        : int  18393 20228 18857 17623 17474 21914 22113 22584 17668 19834 ...
 $ gender     : int  2 1 1 2 1 1 1 2 1 1 ...
 $ height     : int  168 156 165 169 156 151 157 178 158 164 ...
 $ weight     : num  62 85 64 82 56 67 93 95 71 68 ...
 $ ap_hi      : int  110 140 130 150 100 120 130 130 110 110 ...
 $ ap_lo      : int  80 90 70 100 60 80 80 90 70 60 ...
 $ cholesterol: Factor w/ 3 levels "1","2","3": 1 3 3 1 1 2 3 3 1 1 ...
 $ gluc       : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 2 1 3 1 1 ...
 $ smoke      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ alco       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ active     : Factor w/ 2 levels "0","1": 2 2 1 2 1 1 2 2 2 1 ...
 $ cardio     : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 2 1 1 ...
> 
```

The most important aspect of the analysis is the data pre-processing stage. Missing values, impossible values, evident errors (in basic terms, typos), and outlier manipulations are all dealt with when cleaning data.

Missing Values

The any () function in R can be used to find missing values. In most circumstances, the only options for missing values in a dataset are 'NA' or '?' values. We can see that there are no missing values in the data set by looking at the chunk below.



```
> any(is.na(F_cardio))
[1] FALSE
> F_cardio[F_cardio == "?"] <- NA
> any(is.na(F_cardio))
[1] FALSE
> 
```

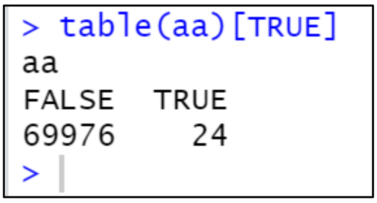
Impossible values

These are the values that can be deduced from two separate viewpoints. For starters, typos might result in illogical results, such as Systolic pressure with a negative sign, which is an evident mistake. We can use the abs () function to convert negative values to absolute values to deal with these kinds of issues.

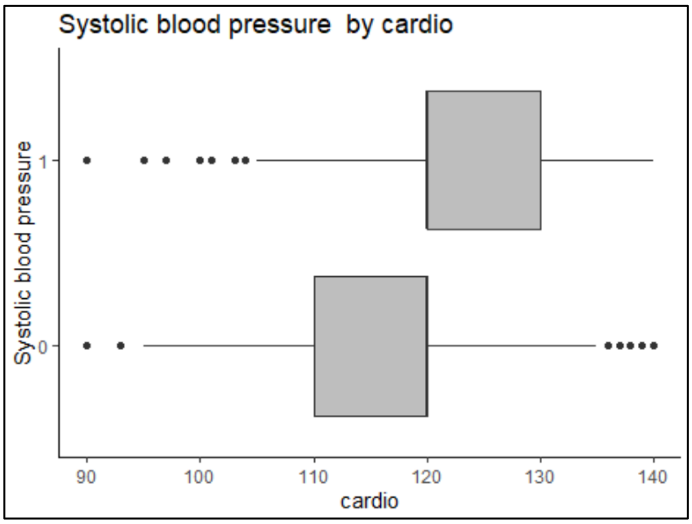
Second, these are also regarded as typos, such as the person's systolic blood pressure is zero, indicating that he or she is on the verge of passing away. In these circumstances, deleting the columns containing these values is the best option. In the dataset, there are only a few rows with these errors. As a result, we removed the rows where the systolic and diastolic pressures were both 0. There are also some outliers who are under 20 kgs in weight. However, the smallest adult (age range starts at 28 years) weight of a human has been reported at 20 kg. As a result, we opted to discard these numbers before addressing the outliers.

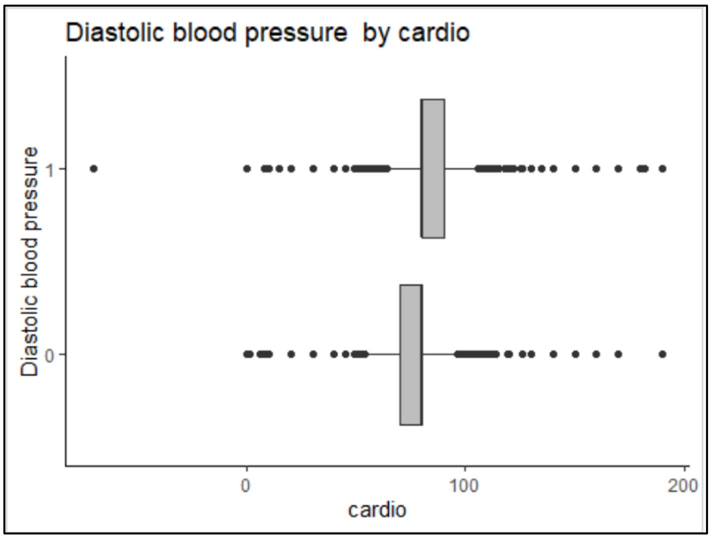
Outliers

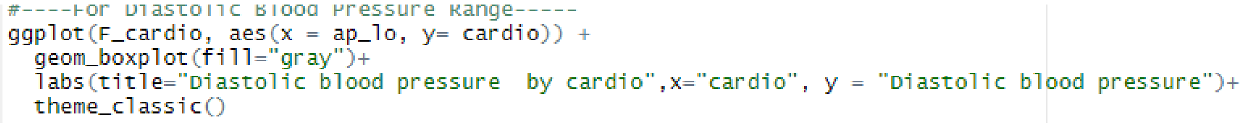
The influence of Outliers on data analysis is significant. Dealing with outliers prior to data analysis will be beneficial. In practice, several ways for dealing with outliers exist, including removing the row, imputing the value with the mean, utilizing the capping function, and even data manipulation. We choose to eliminate the values of systolic and diastolic pressures more than 360 and 370 in this project. These are the highest figures ever found in a research investigation. Then there are the outliers, which imply that these are possible but extreme quantities. These numbers are altered with a capping function and then substituted with 97.5 percent confidence intervals.



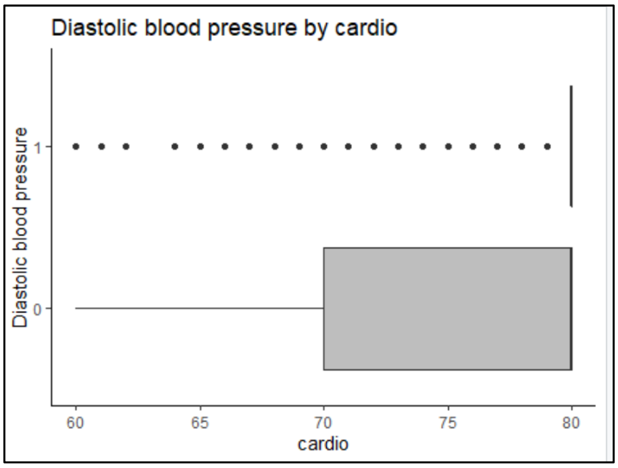
```
> table(aa)[TRUE]
aa
FALSE  TRUE
69976    24
> 
```



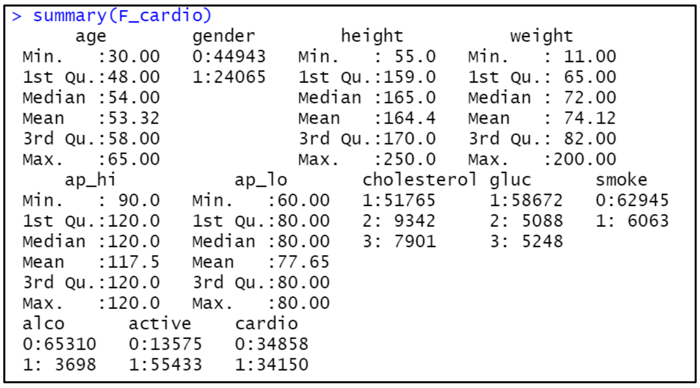




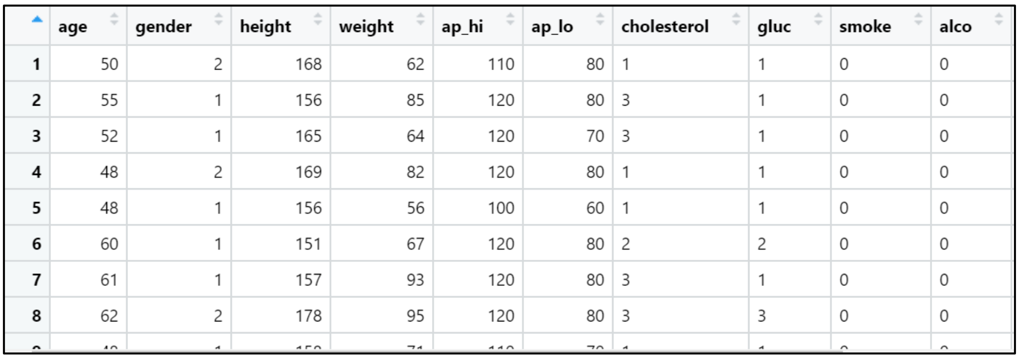
```
#----For Diastolic Blood Pressure Range-----
ggplot(F_cardio, aes(x = ap_lo, y= cardio)) +
  geom_boxplot(fill="gray")+
  labs(title="Diastolic blood pressure  by cardio",x="cardio", y = "Diastolic blood pressure")+
  theme_classic()
```



Except for the age, which is recorded in days, all five numerical variables are in one range, i.e., 1 to 300. As a result, the variable age in days is changed to years.



```
> summary(F_cardio)
      age          gender        height            weight      
 Min.    :30.00   0:44943   Min.   : 55.0   Min.   : 11.00  
 1st Qu.:48.00   1:24065   1st Qu.:159.0   1st Qu.: 65.00  
 Median :54.00             Median :165.0   Median : 72.00  
 Mean    :53.32             Mean   :164.4   Mean   : 74.12  
 3rd Qu.:58.00             3rd Qu.:170.0   3rd Qu.: 82.00  
 Max.    :65.00             Max.   :250.0   Max.   :200.00  
     ap_hi            ap_lo        cholesterol gluc       smoke    
 Min.    : 90.0   Min.   :60.00   1:51765     1:58672   0:62945  
 1st Qu.:120.0   1st Qu.:80.00   2: 9342     2: 5088   1: 6063  
 Median :120.0   Median :80.00   3: 7901     3: 5248            
 Mean    :117.5   Mean   :77.65                                 
 3rd Qu.:120.0   3rd Qu.:80.00                                 
 Max.    :120.0   Max.   :80.00                                 
 alco        active      cardio   
 0:65310    0:13575    0:34858  
 1: 3698    1:55433    1:34150  
```

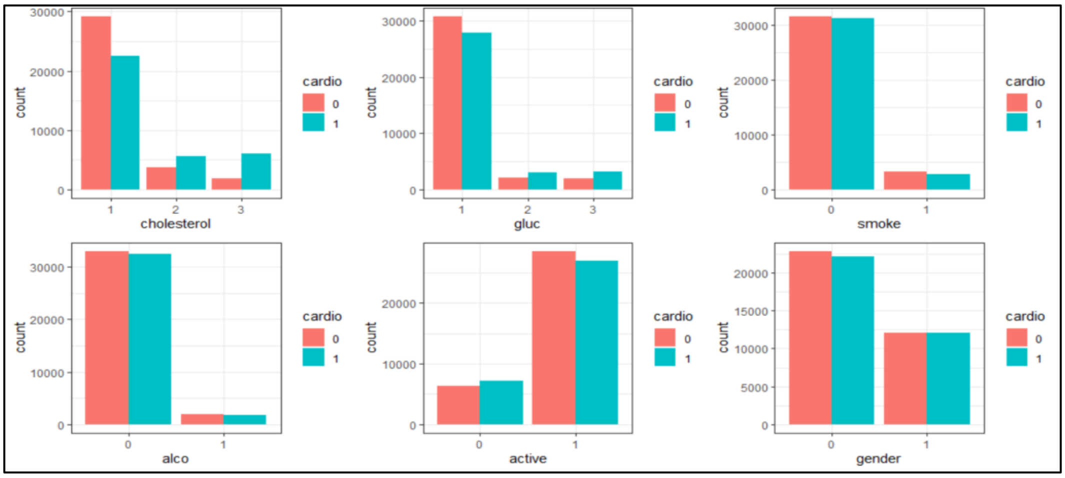


	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco
1	50	2	168	62	110	80	1	1	0	0
2	55	1	156	85	120	80	3	1	0	0
3	52	1	165	64	120	70	3	1	0	0
4	48	2	169	82	120	80	1	1	0	0
5	48	1	156	56	100	60	1	1	0	0
6	60	1	151	67	120	80	2	2	0	0
7	61	1	157	93	120	80	3	1	0	0
8	62	2	178	95	120	80	3	3	0	0

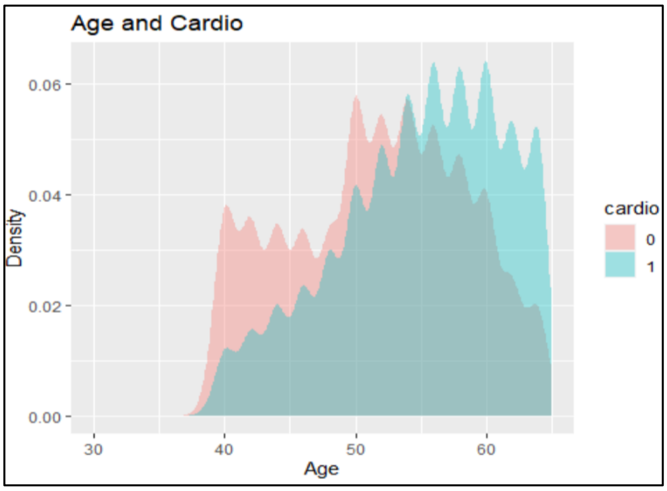
Once we are done with cleaning data variable, we see that age range for patients is from 30 – 65 of age group. We also see that there are unknown data as well. We see somewhat a pattern here. The younger generation is more indulged, and the older generation is less when compared.

3.4. Data Exploration

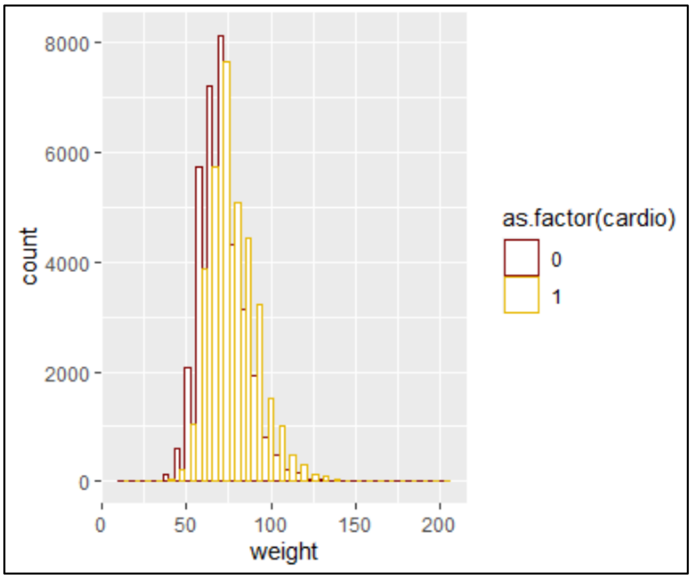
We decided to show cardio versus several alternatives because cardio is the project's target value.



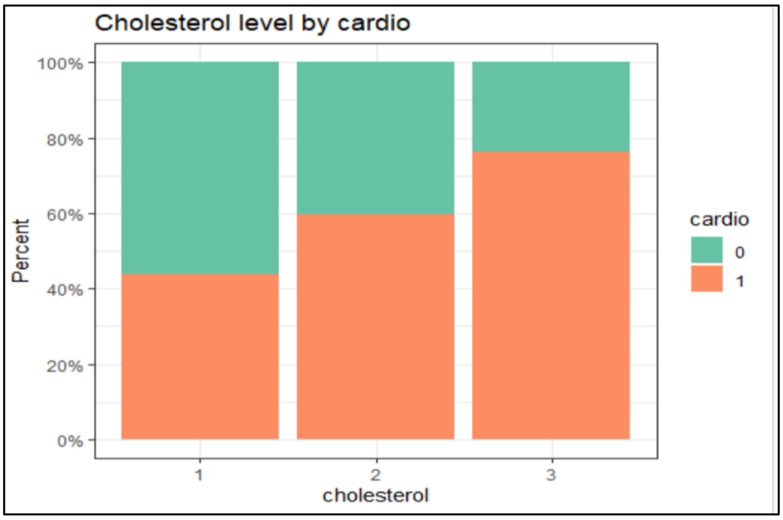
We have a nearly similar distribution of cardiovascular patients at all levels, but we find that more cardio vascular patients who are active, have normal cholesterol and glucose levels, are non-alcoholic, and do not smoke are more common. This is a shocking outcome that we did not anticipate. We come across the correlation factor to cope with this, which is really useful in subsequent investigation.



Age has a significant impact on a variety of category factors. As the density graphs above illustrate, as one gets older, cardio vacular diseases rise, indicating a significant risk.



Significant infact of weight can be seen on presence or absence of disease in male and femal. Patients with weight between 50- 80 kgs are more prone to disease.



We can also see that cholesterol levels have a good impact on cardiovascular health. Which implausibly demonstrates that an increase in cholesterol levels increases the risk of cardiovascular disease.

Calculating BMI

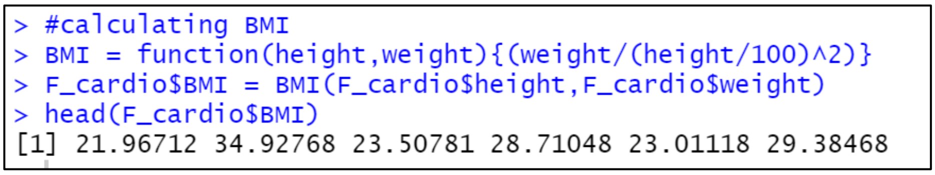
The Body Mass Index (BMI) is a quick way to assess your body size simply with your weight and height, regardless of your gender. Quickly calculate your BMI and find out which category you fall into.

Here we will be calculating BMI. Body mass index (BMI) is a measure of body fat based on height and weight that applies to adult men and women.

The formula for calculation of BMI is

BMI = *(Weight in kilograms) divided by (Height in meters squared)*

BMI is an indicator of total body fat in many individuals. Thus, it is considered as an indicator of health risk.



```
> #calculating BMI
> BMI = function(height,weight){(weight/(height/100)^2)}
> F_cardio$BMI = BMI(F_cardio$height,F_cardio$weight)
> head(F_cardio$BMI)
[1] 21.96712 34.92768 23.50781 28.71048 23.01118 29.38468
```

Analysis

Here, we proceed with identifying the methods we will be using, along with justification for those methods.

Method 1: Hypothesis Testing

Through correlation plot and matrix, we saw variables with high and low correlation. To check if presence or absence of cardiovascular disease is independent or dependent on those variables, we will be performing hypothesis testing.

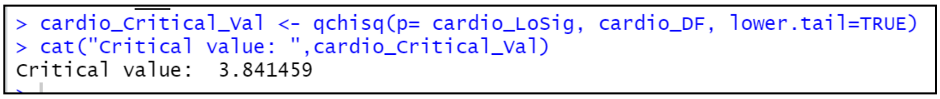
Age

Step 1: Stating Hypothesis:

- Null Hypothesis (H0): Age does not affect presence or absence of cardiovascular disease. (Age and Cardio are in-dependent)
- Alternate Hypothesis (H1): Age does affect presence or absence of cardiovascular disease.

Step 2: Computing Critical Value

- We will use chi- square test to calculate critical value to see difference in sample and population distribution.

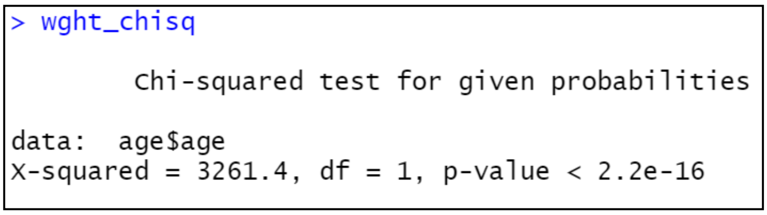


```
> cardio_Critical_Val <- qchisq(p= cardio_LoSig, cardio_DF, lower.tail=TRUE)
> cat("Critical value: ",cardio_Critical_Val)
Critical value:  3.841459
```

- The computed critical value for the given data is 3.84

Step 3: Calculating Test Value

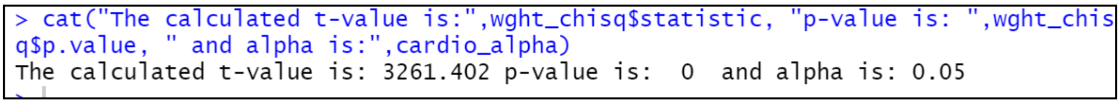
- Using chisq.test function we will compare Age and cardio.



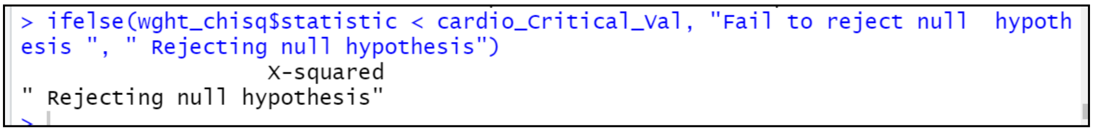
```
> wght_chisq

	Chi-squared test for given probabilities

data:  age$age
X-squared = 3261.4, df = 1, p-value < 2.2e-16
```

- From the Test:
  - X-squared i.e., test statistics is 3261.4,
  - Degree of Freedom = 1,
  - P-value of 2.2e-16

Step 4: Making a Decision



```
> cat("The calculated t-value is:",wght_chisq$statistic, "p-value is: ",wght_chis
q$p.value, " and alpha is:",cardio_alpha)
The calculated t-value is: 3261.402 p-value is:  0  and alpha is: 0.05
```



```
> ifelse(wght_chisq$statistic < cardio_Critical_Val, "Fail to reject null  hypoth
esis ", " Rejecting null hypothesis")
                   X-squared
" Rejecting null hypothesis"
```

- The Chi-square test result we calculated p-value of 2.2e-16, at alpha 0.05, which states that p-value < alpha. Hence, we will **reject null hypothesis**.

Step 5: Summarize Results

We will be rejecting null hypothesis as we have enough evidence to accept alternate hypothesis which states that age does affect presence or absence of cardiovascular disease at alpha 0.05.

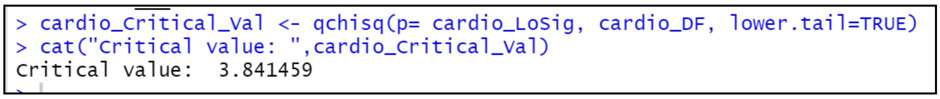
Cholesterol

Step 1: Stating Hypothesis:

- Null Hypothesis (H0): Cholesterol does not affect presence or absence of cardiovascular disease. (Cholesterol and Cardio are in-dependent)
- Alternate Hypothesis (H1): Age does affect presence or absence of cardiovascular disease.

Step 2: Computing Critical Value

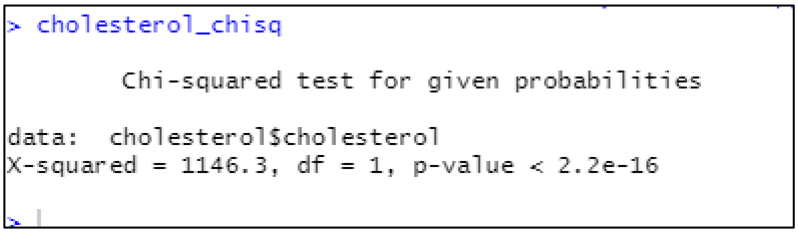
- We will use chi- square test to calculate critical value to see difference in sample and population distribution.



```
> cardio_Critical_Val <- qchisq(p= cardio_LoSig, cardio_DF, lower.tail=TRUE)
> cat("Critical value: ",cardio_Critical_Val)
Critical value:  3.841459
```

Step 3: Calculating Test Value

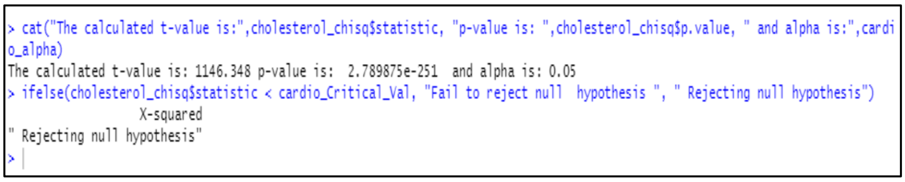
- Using chisq.test function we will compare Cholesterol and cardio.



```
> cholesterol_chisq

	Chi-squared test for given probabilities

data:  cholesterol$cholesterol
X-squared = 1146.3, df = 1, p-value < 2.2e-16

> 
```

- From the Test:
  - X-squared i.e., test statistics is 1146.3,
  - Degree of Freedom = 1,
  - P-value of 2.2e-16

Step 4: Making a Decision



```
> cat("The calculated t-value is:",cholesterol_chisq$statistic, "p-value is: ",cholesterol_chisq$p.value, " and alpha is:",cardi
o_alpha)
The calculated t-value is: 1146.348 p-value is:  2.789875e-251  and alpha is: 0.05
> ifelse(cholesterol_chisq$statistic < cardio_Critical_Val, "Fail to reject null  hypothesis ", " Rejecting null hypothesis")
               X-squared 
" Rejecting null hypothesis" 
> 
```

- The Chi-square test result we calculated p-value of 2.289875e-251, at alpha 0.05, which states that p-value < alpha. Hence, we will **reject null hypothesis**.

Step 5: Summarize Results

We will be rejecting null hypothesis as we have enough evidence to accept alternate hypothesis which states that cholesterol does affect presence or absence of cardiovascular disease at alpha 0.05.

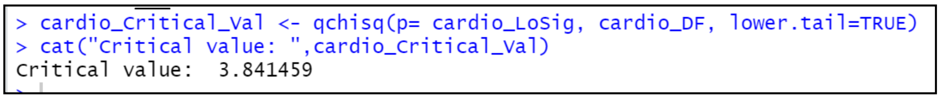
Weight

Step 1: Stating Hypothesis:

- Null Hypothesis (H0): Weight does not affect presence or absence of cardiovascular disease. (Weight and Cardio are in-dependent)
- H1: Weight does affect presence or absence of cardiovascular disease.

Step 2: Computing Critical Value

- We will use chi- square test to calculate critical value to see difference in sample and population distribution.

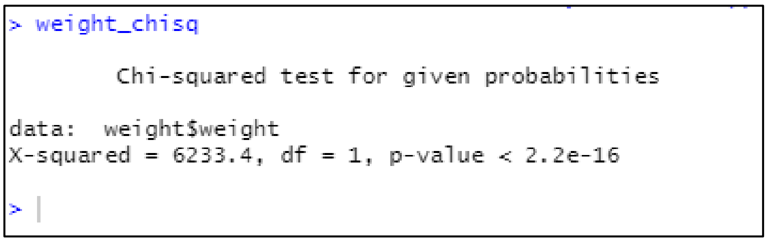


```
> cardio_Critical_Val <- qchisq(p= cardio_LoSig, cardio_DF, lower.tail=TRUE)
> cat("Critical value: ",cardio_Critical_Val)
Critical value:  3.841459
```

- The computed critical value for the given data is 3.84

Step 3: Calculating Test Value

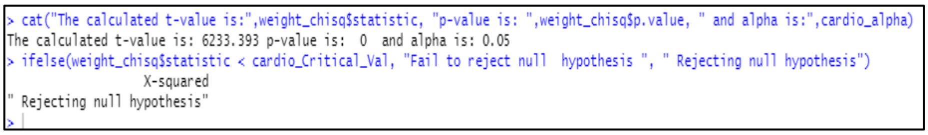
- Using chisq.test function we will compare weight and cardio.



```
> weight_chisq

        Chi-squared test for given probabilities

data:  weight$weight
X-squared = 6233.4, df = 1, p-value < 2.2e-16

> 
```

- From the Test:
  - X-squared i.e., test statistics is 6233.4,
  - Degree of Freedom = 1,
  - P-value of 2.2e-16

Step 4: Making a Decision



```
> cat("The calculated t-value is:",weight_chisq$statistic, "p-value is: ",weight_chisq$p.value, " and alpha is:",cardio_alpha)
The calculated t-value is: 6233.393 p-value is:  0  and alpha is: 0.05
> ifelse(weight_chisq$statistic < cardio_Critical_Val, "Fail to reject null  hypothesis ", " Rejecting null hypothesis")
                X-squared
" Rejecting null hypothesis"
> 
```

- The Chi-square test result we calculated p-value of 2.2e-16, at alpha 0.05, which states that p-value < alpha. Hence, we will **reject null hypothesis**.

Step 5: Summarize Results

We will be rejecting null hypothesis as we have enough evidence to accept alternate hypothesis which states that weight does affect presence or absence of cardiovascular disease at alpha 0.05.

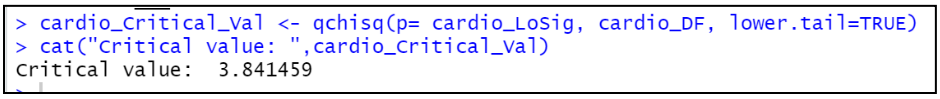
BMI

Step 1: Stating Hypothesis:

- Null Hypothesis (H0): BMI does not affect presence or absence of cardiovascular disease. (BMI and Cardio are in-dependent)
- H1: BMI does affect presence or absence of cardiovascular disease.

Step 2: Computing Critical Value

- We will use chi- square test to calculate critical value to see difference in sample and population distribution.

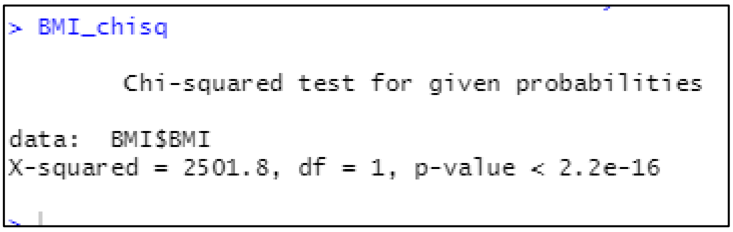


```
> cardio_Critical_Val <- qchisq(p= cardio_LoSig, cardio_DF, lower.tail=TRUE)
> cat("Critical value: ",cardio_Critical_Val)
Critical value:  3.841459
```

- The computed critical value for the given data is 3.84

Step 3: Calculating Test Value

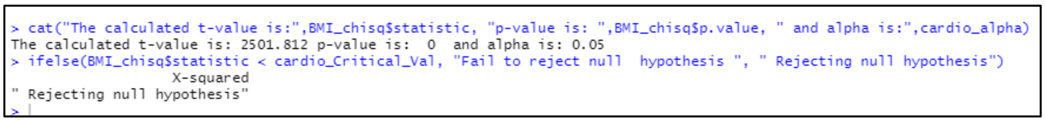
- Using chisq.test function we will compare Age and cardio.



```
> BMI_chisq

	Chi-squared test for given probabilities

data:  BMI$BMI
X-squared = 2501.8, df = 1, p-value < 2.2e-16

>
```

- From the Test:
    - X-squared i.e., test statistics is 2501.8,
    - Degree of Freedom = 1,
    - P-value of 2.2e-16

Step 4: Making a Decision



```
> cat("The calculated t-value is:",BMI_chisq$statistic, "p-value is: ",BMI_chisq$p.value, " and alpha is:",cardio_alpha)
The calculated t-value is: 2501.812 p-value is:  0  and alpha is: 0.05
> ifelse(BMI_chisq$statistic < cardio_Critical_Val, "Fail to reject null  hypothesis ", " Rejecting null hypothesis")
                 X-squared
" Rejecting null hypothesis"
>
```

- The Chi-square test result we calculated p-value of 2.2e-16, at alpha 0.05, which states that p-value < alpha. Hence, we will **reject null hypothesis**.

Step 5: Summarize Results

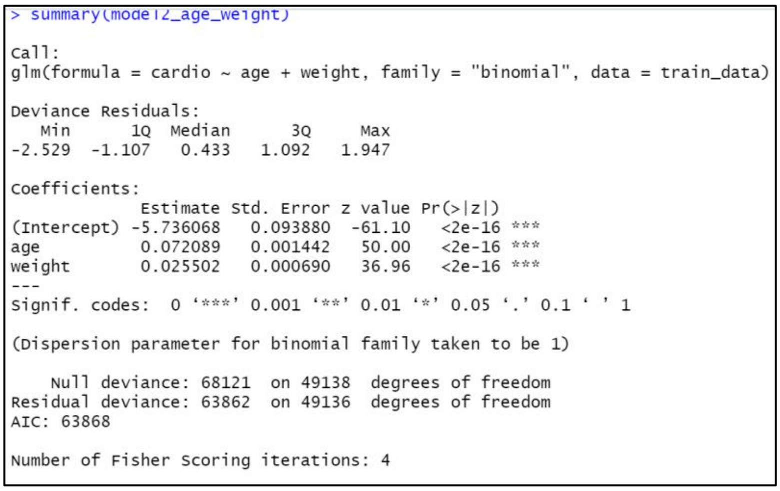
We will be rejecting null hypothesis as we have enough evidence to accept alternate hypothesis which states that BMI does affect presence or absence of cardiovascular disease at alpha 0.05.

Business Questions

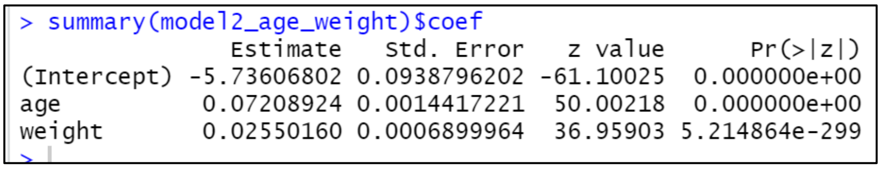
1. Is there a correlation between Age and Weight with respect to presence or absence of cardiovascular disease?

For predicting cardio with using age, weight

Model 2: GLM



```
> summary(model2_age_weight)

Call:
glm(formula = cardio ~ age + weight, family = "binomial", data = train_data)

Deviance Residuals: 
   Min      1Q  Median      3Q     Max  
-2.529  -1.107   0.433   1.092   1.947  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -5.736068   0.093880  -61.10   <2e-16 ***
age          0.072089   0.001442   50.00   <2e-16 ***
weight       0.025502   0.000690   36.96   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68121  on 49138  degrees of freedom
Residual deviance: 63862  on 49136  degrees of freedom
AIC: 63868

Number of Fisher Scoring iterations: 4
```



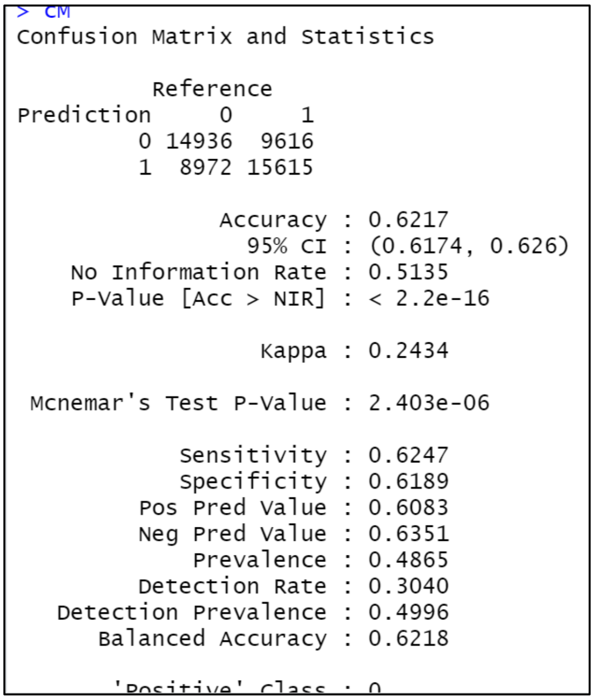
```
> summary(model2_age_weight)$coef
              Estimate    Std. Error   z value      Pr(>|z|)
(Intercept) -5.73606802 0.0938796202 -61.10025  0.000000e+00
age          0.07208924 0.0014417221  50.00218  0.000000e+00
weight       0.02550160 0.0006899964  36.95903 5.214864e-299
> 
```

We are looking at Cardio vs Age and Weight. After fitting model with desired variables we get Weight, and Weight with negative effects. Coefficient of weight is non- significant (p > 0.05) whereas coefficient of age is significant. Our Null Deviance value is 68121 on 49138 degrees of freedom. After including independent variables, our deviance is decrease to 63862 points on 49136 degrees of freedom, which is very less reduction. Residual Deviance has reduced by 4259 with loss of 2 degrees of freedom. For this model, four iterations were performed to fit by Fisher’s Scoring Algorithm.

Confusion Matrix on Data

For evaluating performance of our classification model, we use N * N matrix called as Confusion Matrix. Here, N is total number of targeted classes where we compare actual targeted value with our predicted value. Here, diagonal value states the correct model output whereas off- diagonal values represent incorrect ones

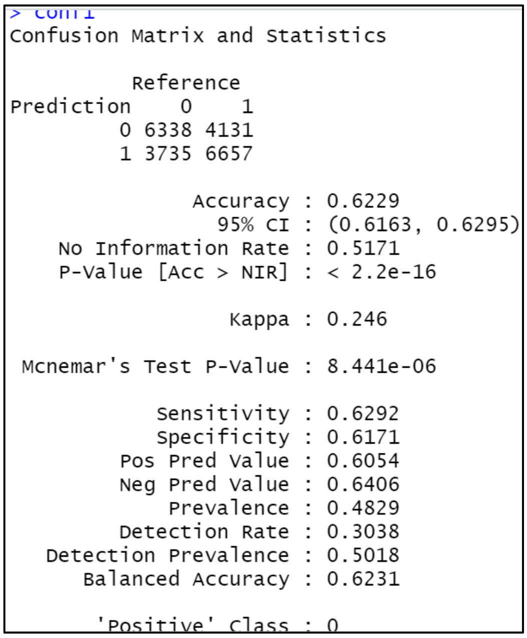
Confusion matrix for train dataset



```
> cm
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 14936  9616
         1  8972 15615

               Accuracy : 0.6217
                 95% CI : (0.6174, 0.626)
    No Information Rate : 0.5135
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2434

 Mcnemar's Test P-Value : 2.403e-06

            Sensitivity : 0.6247
            Specificity : 0.6189
         Pos Pred Value : 0.6083
         Neg Pred Value : 0.6351
             Prevalence : 0.4865
         Detection Rate : 0.3040
   Detection Prevalence : 0.4996
      Balanced Accuracy : 0.6218

       'Positive' Class : 0
```

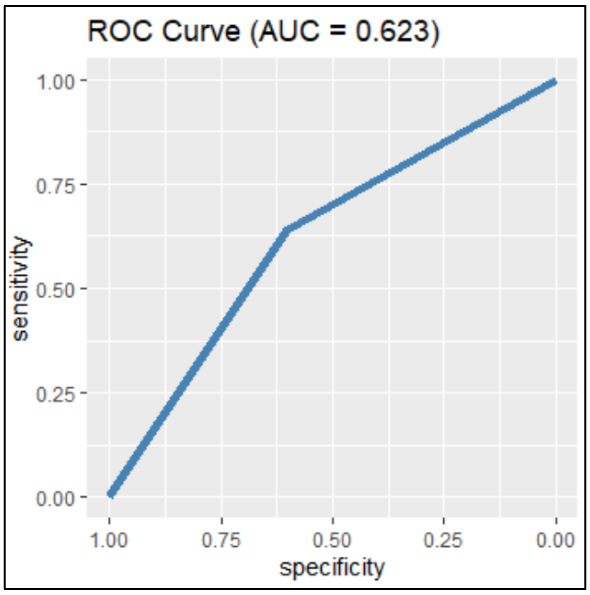
From the above matrix with accuracy value 62%, prediction model does not seem to be working very well and not a desirable model. Value of false- positive and false- negative are very high. Sensitivity is 0.62 and Specificity is 0.62. This model is not good for implementation.

Confusion matrix for test dataset



```
> conf1
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6338 4131
         1 3735 6657

               Accuracy : 0.6229
                 95% CI : (0.6163, 0.6295)
    No Information Rate : 0.5171
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.246

 Mcnemar's Test P-Value : 8.441e-06

            Sensitivity : 0.6292
            Specificity : 0.6171
         Pos Pred Value : 0.6054
         Neg Pred Value : 0.6406
             Prevalence : 0.4829
         Detection Rate : 0.3038
   Detection Prevalence : 0.5018
      Balanced Accuracy : 0.6231

       'Positive' Class : 0
```

From the above matrix with accuracy value 62%, prediction model does not seem to be working very well and not a desirable model. Value of false- positive and false- negative are very high. Sensitivity is 0.62 and Specificity is 0.62. This model is also not good for implementation.



Above plots gives us value of area under Receiver Operating Characteristics curve. AUROC values for training dataset as 0.622 showing us that the model is not a good fit.

Model 3: LM

Age, Weight and Cardio are 2 predictor X variables which are continuous. To predict y, we express it as:

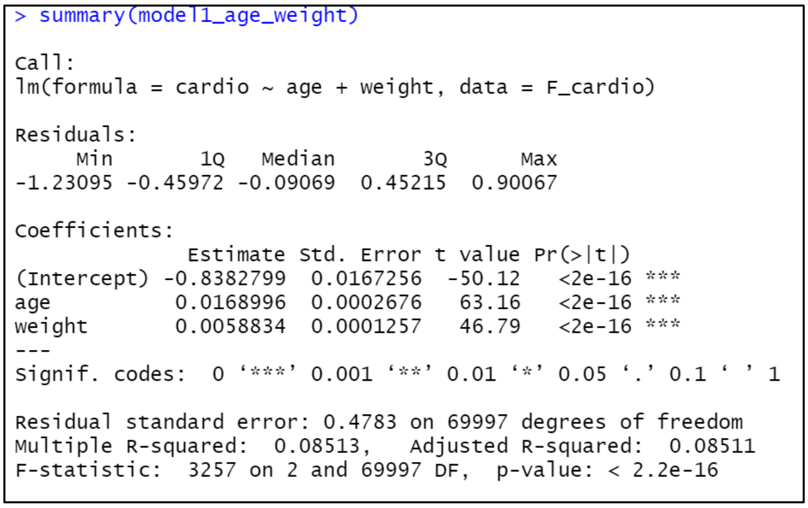
Y = b0 + b1 * x1 + b2 * x2

Where, y – Cardio,

x1 - Age,

x2 - Weight.

Let us interpret and observe each model coefficient for this given problem.

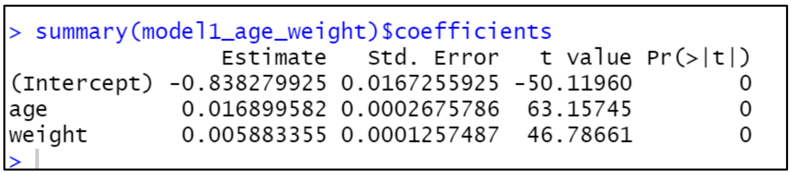


```
> summary(model1_age_weight)

Call:
lm(formula = cardio ~ age + weight, data = F_cardio)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.23095 -0.45972 -0.09069  0.45215  0.90067 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.8382799  0.0167256  -50.12   <2e-16 ***
age          0.0168996  0.0002676   63.16   <2e-16 ***
weight       0.0058834  0.0001257   46.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4783 on 69997 degrees of freedom
Multiple R-squared:  0.08513,	Adjusted R-squared:  0.08511 
F-statistic:  3257 on 2 and 69997 DF,  p-value: < 2.2e-16
```

- An (adjusted) R2 that is close to indicates that a large proportion of the variability in the outcome has been explained by the regression model.
- A number near 0 indicates that the regression model did not explain much of the variability in the outcome.

In our case, value or R square is 0.085 which is high.

A large F-statistic will correspond to a statistically significant p-value ($p < 0.05$). In our example, the F-statistic equal 3257 producing a p-value of 2.2e-16, which is highly significant.



```
> summary(model1_age_weight)$coefficients
               Estimate    Std. Error   t value Pr(>|t|)
(Intercept) -0.838279925 0.0167255925 -50.11960        0
age          0.016899582 0.0002675786  63.15745        0
weight       0.005883355 0.0001257487  46.78661        0
> 
```

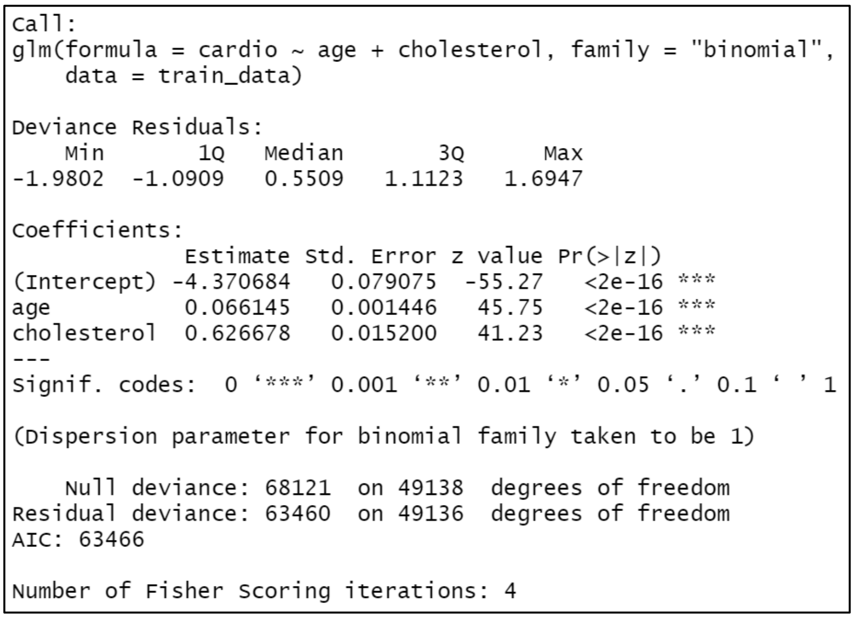
So, above equation for model after substituting value:

Cardio = -0.84 + (0.017) age + (0.0059) weight

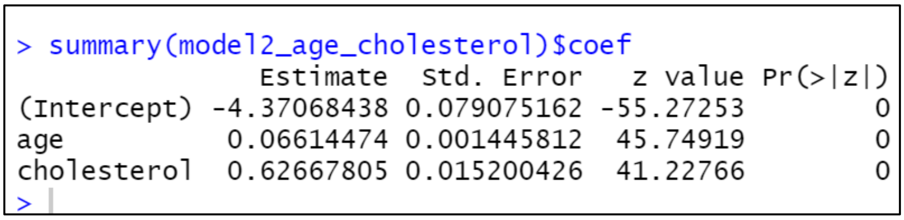
2. Is there a correlation between Age group and Gender with presence or absence of cardiovascular disease among observed patients?

For predicting cardio using age and cholesterol

Model 2: GLM



```
Call:
glm(formula = cardio ~ age + cholesterol, family = "binomial", 
    data = train_data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.9802  -1.0909   0.5509   1.1123   1.6947  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -4.370684   0.079075  -55.27   <2e-16 ***
age          0.066145   0.001446   45.75   <2e-16 ***
cholesterol  0.626678   0.015200   41.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68121  on 49138  degrees of freedom
Residual deviance: 63460  on 49136  degrees of freedom
AIC: 63466

Number of Fisher Scoring iterations: 4
```



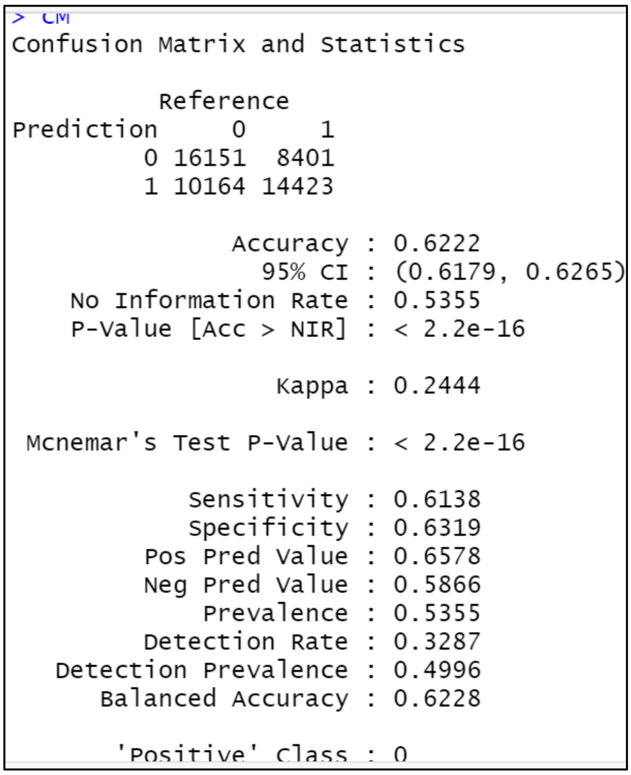
```
> summary(model2_age_cholesterol)$coef
              Estimate  Std. Error   z value Pr(>|z|)
(Intercept) -4.37068438 0.079075162 -55.27253        0
age          0.06614474 0.001445812  45.74919        0
cholesterol  0.62667805 0.015200426  41.22766        0
> 
```

We are looking at Cardio vs age and cholesterol. After fitting model with desired variable we get age and cholesterol with positive effects. Coefficients of age and cholesterol are significant ($p < 0.05$). Our Null Deviance value is 68121 on 49138 degrees of freedom. After including independent variables, our deviance is decrease to 63460 points on 49136 degrees of freedom, which show significant reduction. Residual Deviance has reduced by 4661 with loss of 2 degrees of freedom. For this model, four iterations were performed to fit by Fisher's Scoring Algorithm.

Confusion Matrix on Data

For evaluating performance of our classification model, we use N * N matrix called as Confusion Matrix. Here, N is total number of targeted classes where we compare actual targeted value with our predicted value. Here, diagonal value states the correct model output whereas off- diagonal values represent incorrect ones.

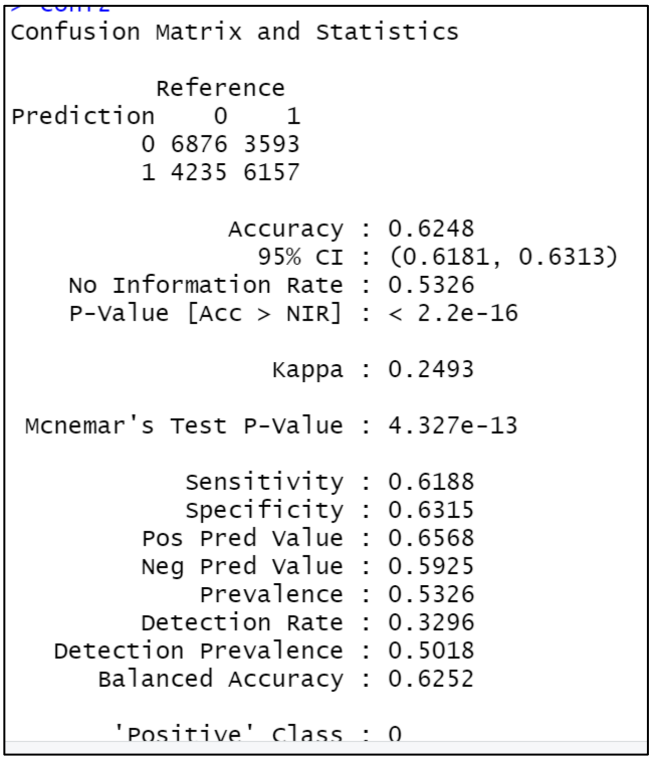
Confusion matrix for train dataset



```
> CM
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 16151  8401
         1 10164 14423

               Accuracy : 0.6222
                 95% CI : (0.6179, 0.6265)
    No Information Rate : 0.5355
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2444

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6138
            Specificity : 0.6319
         Pos Pred Value : 0.6578
         Neg Pred Value : 0.5866
             Prevalence : 0.5355
         Detection Rate : 0.3287
   Detection Prevalence : 0.4996
      Balanced Accuracy : 0.6228

       'Positive' Class : 0
```

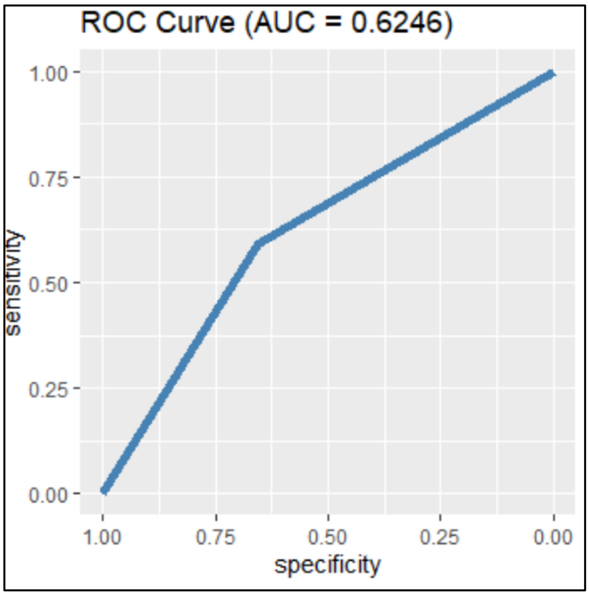
From the above matrix with accuracy value 62%, prediction model does not seem to be working very well and not a desirable model. Value of false- positive and false- negative are very high. Sensitivity is 0.61 and Specificity is 0.63. This model is also not good for implementation

Confusion matrix for test dataset



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6876 3593
         1 4235 6157

               Accuracy : 0.6248
                 95% CI : (0.6181, 0.6313)
    No Information Rate : 0.5326
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2493

 Mcnemar's Test P-Value : 4.327e-13

            Sensitivity : 0.6188
            Specificity : 0.6315
         Pos Pred Value : 0.6568
         Neg Pred Value : 0.5925
             Prevalence : 0.5326
         Detection Rate : 0.3296
   Detection Prevalence : 0.5018
      Balanced Accuracy : 0.6252

       'Positive' Class : 0
```

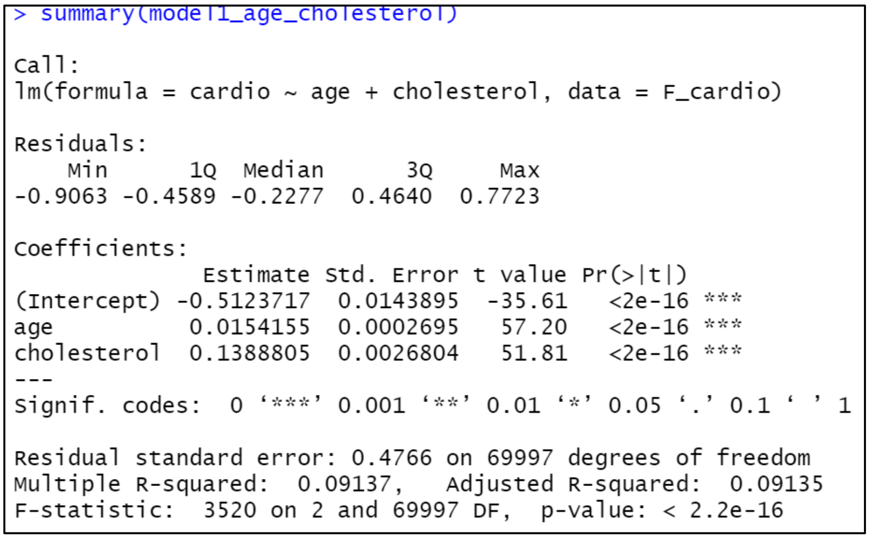
From the above matrix with accuracy value 63%, prediction model does not seem to be working very well and not a desirable model. Value of false- positive and false- negative are very high. Sensitivity is 0.62 and Specificity is 0.63. This model is also not good for implementation



Above plots gives us value of area under Receiver Operating Characteristics curve. AUROC values for training dataset as 0.625 showing us that the model is not a good fit.

Model 3: LM

Let us interpret and observe each model coefficient for this given problem.

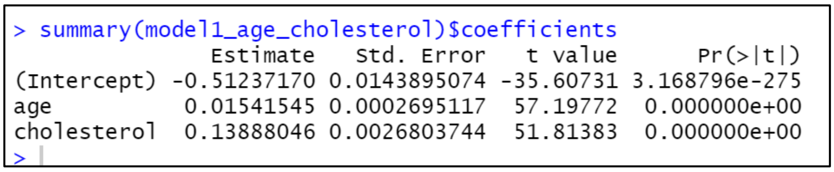


```
> summary(model1_age_cholesterol)

Call:
lm(formula = cardio ~ age + cholesterol, data = F_cardio)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9063 -0.4589 -0.2277  0.4640  0.7723 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.5123717  0.0143895  -35.61   <2e-16 ***
age          0.0154155  0.0002695   57.20   <2e-16 ***
cholesterol  0.1388805  0.0026804   51.81   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4766 on 69997 degrees of freedom
Multiple R-squared:  0.09137,   Adjusted R-squared:  0.09135 
F-statistic:  3520 on 2 and 69997 DF,  p-value: < 2.2e-16
```

- An (adjusted) R2 that is close to indicates that a large proportion of the variability in the outcome has been explained by the regression model.
- A number near 0 indicates that the regression model did not explain much of the variability in the outcome.

In our case, value or R square is 0.091 which is high.

A large F-statistic will correspond to a statistically significant p-value ($p < 0.05$). In our example, the F-statistic equal 3250 producing a p-value of 2.2e-16, which is highly significant.



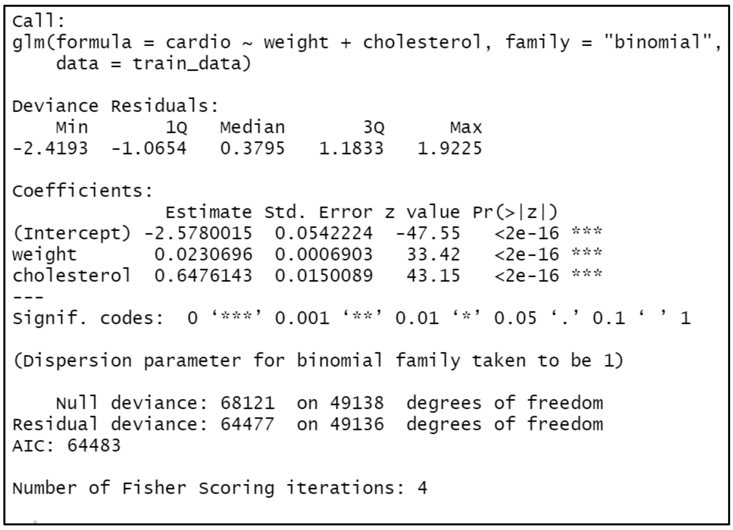
```
> summary(model1_age_cholesterol)$coefficients
               Estimate   Std. Error   t value      Pr(>|t|)
(Intercept) -0.51237170 0.0143895074 -35.60731 3.168796e-275
age          0.01541545 0.0002695117  57.19772  0.000000e+00
cholesterol  0.13888046 0.0026803744  51.81383  0.000000e+00
> 
```

So, above equation for model after substituting value:

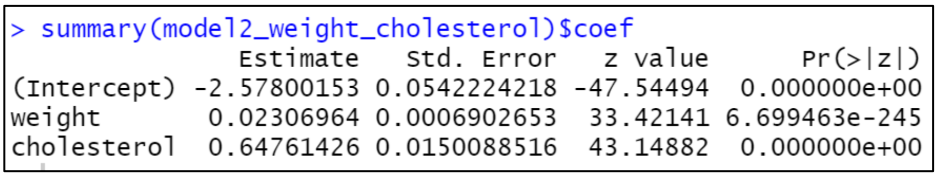
Cardio = -0.512 + (0.015) age + (0.139) cholesterol

3. Is there a correlation between Weight and Cholesterol with respect to presence or absence of cardiovascular disease?

Model 2: GLM



```
Call:
glm(formula = cardio ~ weight + cholesterol, family = "binomial", 
    data = train_data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.4193  -1.0654   0.3795   1.1833   1.9225  

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.5780015  0.0542224  -47.55   <2e-16 ***
weight       0.0230696  0.0006903   33.42   <2e-16 ***
cholesterol  0.6476143  0.0150089   43.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68121  on 49138  degrees of freedom
Residual deviance: 64477  on 49136  degrees of freedom
AIC: 64483

Number of Fisher Scoring iterations: 4
```



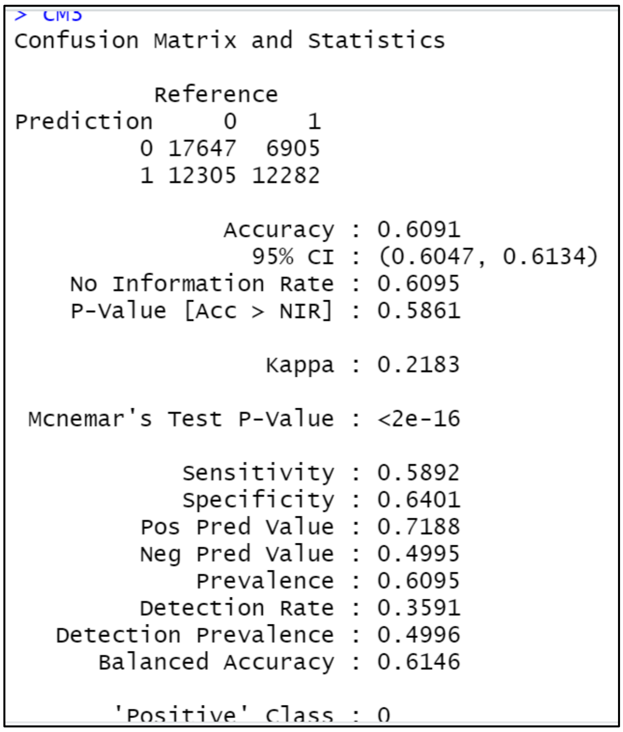
```
> summary(model2_weight_cholesterol)$coef
               Estimate   Std. Error   z value      Pr(>|z|)
(Intercept) -2.57800153 0.0542224218 -47.54494  0.000000e+00
weight       0.02306964 0.0006902653  33.42141 6.699463e-245
cholesterol  0.64761426 0.0150088516  43.14882  0.000000e+00
```

We are looking at Cardio vs weight and cholesterol. After fitting model with desired variables, we get weight and cholesterol with positive effect. Coefficients of weight is non- significant ($p > 0.05$) and significant for cholesterol. Our Null Deviance value is 68121 on 49138 degrees of freedom. After including independent variables, our deviance is decrease to 64477 points on 49136 degrees of freedom, which show significant reduction. Residual Deviance has reduced by 3644 with loss of 2 degrees of freedom. For this model, four iterations were performed to fit by Fisher’s Scoring Algorithm.

Confusion Matrix

For evaluating performance of our classification model, we use N * N matrix called as Confusion Matrix. Here, N is total number of targeted classes where we compare actual targeted value with our predicted value. Here, diagonal value states the correct model output whereas off- diagonal values represent incorrect ones.

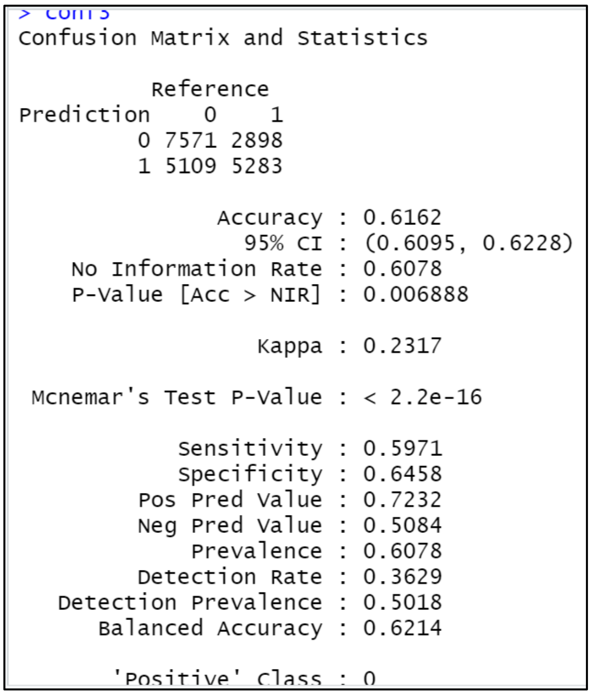
Confusion matrix for train dataset



```
> cm5
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 17647  6905
         1 12305 12282

               Accuracy : 0.6091
                 95% CI : (0.6047, 0.6134)
    No Information Rate : 0.6095
    P-Value [Acc > NIR] : 0.5861

                  Kappa : 0.2183

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.5892
            Specificity : 0.6401
         Pos Pred Value : 0.7188
         Neg Pred Value : 0.4995
             Prevalence : 0.6095
         Detection Rate : 0.3591
   Detection Prevalence : 0.4996
      Balanced Accuracy : 0.6146

       'Positive' Class : 0
```

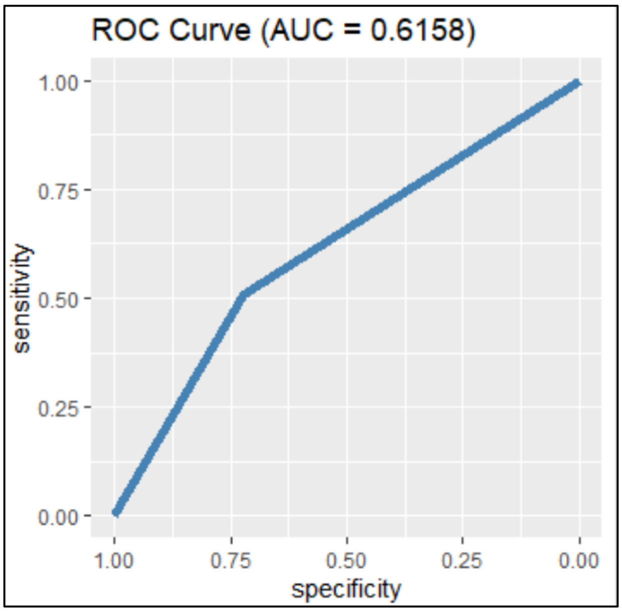
From the above matrix with accuracy value 61%, prediction model does not seem to be working very well and not a desirable model. Value of false- positive and false- negative are very high. Sensitivity is 0.59 and Specificity is 0.64. This model is not good for implementation.

Confusion matrix for test dataset



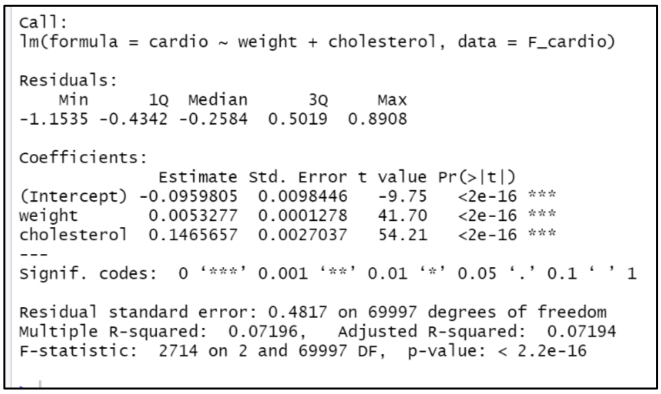
```
> conf5
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7571 2898
         1 5109 5283

               Accuracy : 0.6162
                 95% CI : (0.6095, 0.6228)
    No Information Rate : 0.6078
    P-Value [Acc > NIR] : 0.006888

                  Kappa : 0.2317

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.5971
            Specificity : 0.6458
         Pos Pred Value : 0.7232
         Neg Pred Value : 0.5084
             Prevalence : 0.6078
         Detection Rate : 0.3629
   Detection Prevalence : 0.5018
      Balanced Accuracy : 0.6214

       'Positive' Class : 0
```

From the above matrix with accuracy value 62%, prediction model does not seem to be working very well and not a desirable model. Value of false- positive and false- negative are very high. Sensitivity is 0.59 and Specificity is 0.64. This model is also not good for implementation



Above plots gives us value of area under Receiver Operating Characteristics curve. AUROC values for training dataset as 0.62 showing us that the model is not a good fit.

Model 3: LM

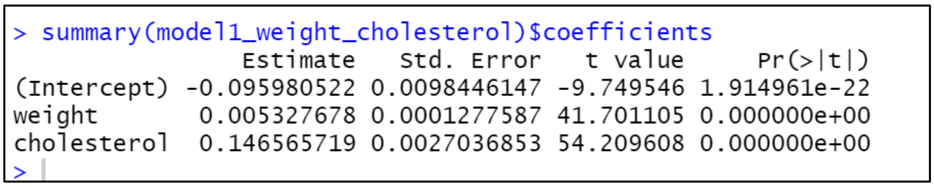


```
Call:
lm(formula = cardio ~ weight + cholesterol, data = F_cardio)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.1535 -0.4342 -0.2584  0.5019  0.8908 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.0959805  0.0098446   -9.75   <2e-16 ***
weight       0.0053277  0.0001278   41.70   <2e-16 ***
cholesterol  0.1465657  0.0027037   54.21   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4817 on 69997 degrees of freedom
Multiple R-squared:  0.07196,	Adjusted R-squared:  0.07194 
F-statistic:  2714 on 2 and 69997 DF,  p-value: < 2.2e-16
```

- An (adjusted) R2 that is close to indicates that a large proportion of the variability in the outcome has been explained by the regression model.
- A number near 0 indicates that the regression model did not explain much of the variability in the outcome.

In our case, value or R square is 0.072 which is high.

A large F-statistic will correspond to a statistically significant p-value ($p < 0.05$). In our example, the F-statistic equal 2714 producing a p-value of 2.2e-16, which is highly significant.



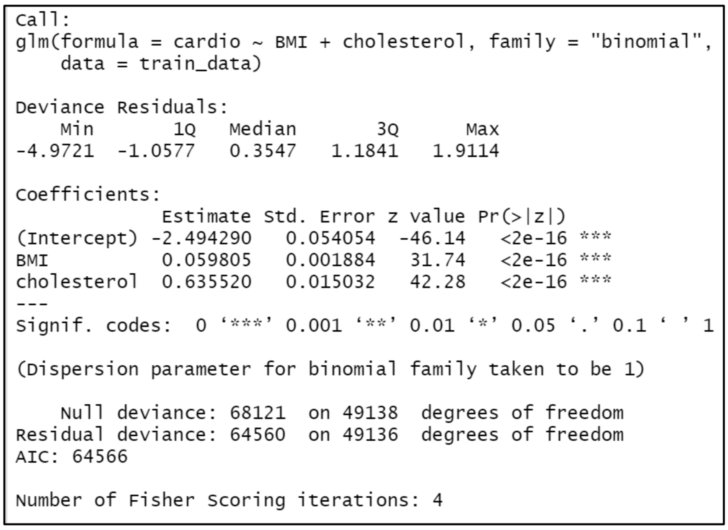
```
> summary(model1_weight_cholesterol)$coefficients
               Estimate   Std. Error   t value     Pr(>|t|)
(Intercept) -0.095980522 0.0098446147 -9.749546 1.914961e-22
weight       0.005327678 0.0001277587 41.701105 0.000000e+00
cholesterol  0.146565719 0.0027036853 54.209608 0.000000e+00
> 
```

So, above equation for model after substituting value:

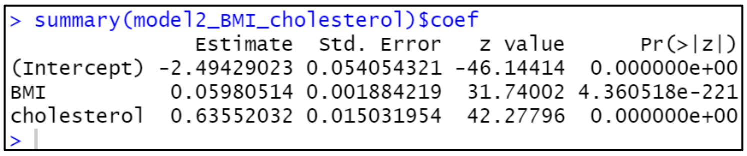
Cardio = -0.0959 + (0.005) weight + (0.146) cholesterol

4. Is there a correlation between BMI and Cholesterol with respect to presence or absence of cardiovascular disease?

Model 2: GLM



```
Call:
glm(formula = cardio ~ BMI + cholesterol, family = "binomial",
    data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.9721  -1.0577   0.3547   1.1841   1.9114

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.494290   0.054054  -46.14   <2e-16 ***
BMI          0.059805   0.001884   31.74   <2e-16 ***
cholesterol  0.635520   0.015032   42.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68121  on 49138  degrees of freedom
Residual deviance: 64560  on 49136  degrees of freedom
AIC: 64566

Number of Fisher Scoring iterations: 4
```



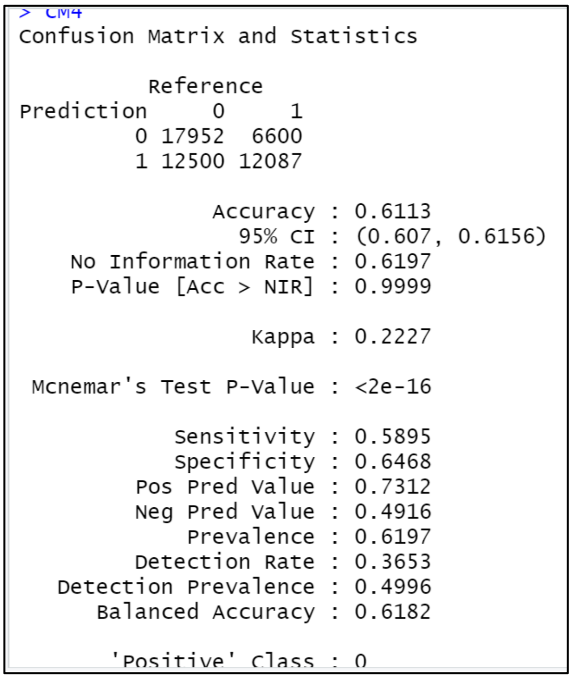
```
> summary(model2_BMI_cholesterol)$coef
              Estimate  Std. Error   z value      Pr(>|z|)
(Intercept) -2.49429023 0.054054321 -46.14414  0.000000e+00
BMI          0.05980514 0.001884219  31.74002 4.360518e-221
cholesterol  0.63552032 0.015031954  42.27796  0.000000e+00
>
```

We are looking at Cardio vs BMI and cholesterol. After fitting model with desired variables, we get weight and cholesterol with positive effect. Coefficients of BMI is non- significant ($p > 0.05$) and significant for cholesterol. Our Null Deviance value is 68121 on 49138 degrees of freedom. After including independent variables, our deviance is decrease to 64560 points on 49136 degrees of freedom, which show significant reduction. Residual Deviance has reduced by 3561 with loss of 2 degrees of freedom. For this model, four iterations were performed to fit by Fisher’s Scoring Algorithm.

Confusion Matrix

For evaluating performance of our classification model, we use N * N matrix called as Confusion Matrix. Here, N is total number of targeted classes where we compare actual targeted value with our predicted value. Here, diagonal value states the correct model output whereas off- diagonal values represent incorrect ones.

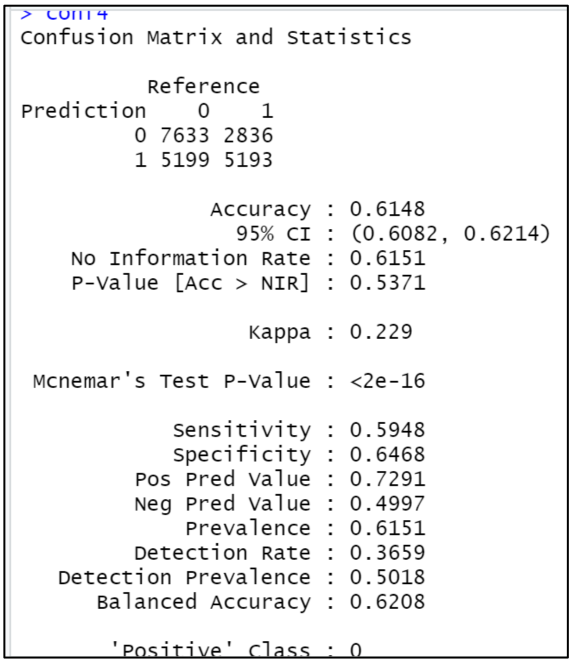
Confusion matrix for train dataset



```
> CM4
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 17952  6600
         1 12500 12087

               Accuracy : 0.6113
                 95% CI : (0.607, 0.6156)
    No Information Rate : 0.6197
    P-Value [Acc > NIR] : 0.9999

                  Kappa : 0.2227

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.5895
            Specificity : 0.6468
         Pos Pred Value : 0.7312
         Neg Pred Value : 0.4916
             Prevalence : 0.6197
         Detection Rate : 0.3653
   Detection Prevalence : 0.4996
      Balanced Accuracy : 0.6182

       'Positive' Class : 0
```

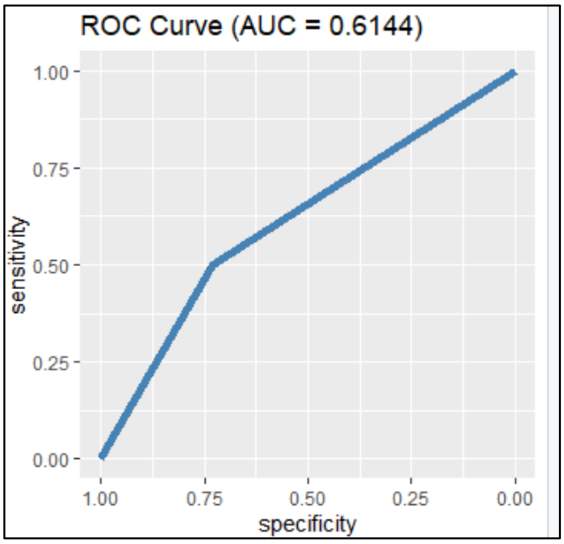
From the above matrix with accuracy value 61%, prediction model does not seem to be working very well and not a desirable model. Value of false- positive and false- negative are very high. Sensitivity is 0.59 and Specificity is 0.65. This model is not good for implementation.

Confusion matrix for test dataset



```
> conf4
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7633 2836
         1 5199 5193

               Accuracy : 0.6148
                 95% CI : (0.6082, 0.6214)
    No Information Rate : 0.6151
    P-Value [Acc > NIR] : 0.5371

                  Kappa : 0.229

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.5948
            Specificity : 0.6468
         Pos Pred Value : 0.7291
         Neg Pred Value : 0.4997
             Prevalence : 0.6151
         Detection Rate : 0.3659
   Detection Prevalence : 0.5018
      Balanced Accuracy : 0.6208

       'Positive' Class : 0
```

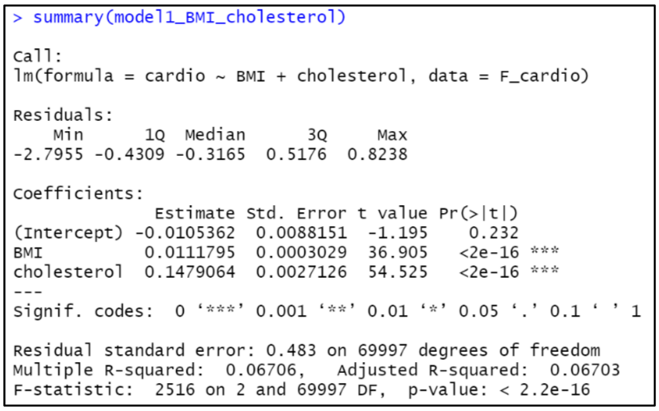
From the above matrix with accuracy value 62%, prediction model does not seem to be working very well and not a desirable model. Value of false- positive and false- negative are very high. Sensitivity is 0.59 and Specificity is 0.65. This model is also not good for implementation



Above plots gives us value of area under Receiver Operating Characteristics curve. AUROC values for training dataset as 0.61 showing us that the model is not a good fit.

Model 3: LM

Let us interpret and observe each model coefficient for this given problem.

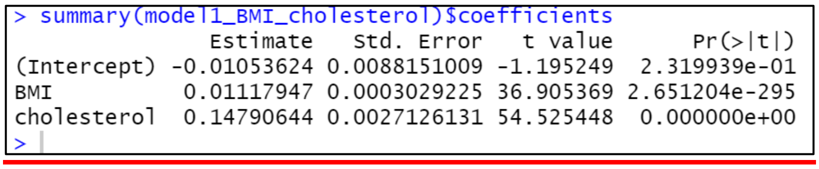


```
> summary(model1_BMI_cholesterol)

Call:
lm(formula = cardio ~ BMI + cholesterol, data = F_cardio)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7955 -0.4309 -0.3165  0.5176  0.8238 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.0105362  0.0088151  -1.195    0.232    
BMI          0.0111795  0.0003029  36.905   <2e-16 ***
cholesterol  0.1479064  0.0027126  54.525   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.483 on 69997 degrees of freedom
Multiple R-squared:  0.06706,	Adjusted R-squared:  0.06703 
F-statistic:  2516 on 2 and 69997 DF,  p-value: < 2.2e-16
```

- An (adjusted) R2 that is close to indicates that a large proportion of the variability in the outcome has been explained by the regression model.
- A number near 0 indicates that the regression model did not explain much of the variability in the outcome.

In our case, value or R square is 0.067 which is high.

A large F-statistic will correspond to a statistically significant p-value ($p < 0.05$). In our example, the F-statistic equal 2516 producing a p-value of 2.2e-16, which is highly significant.



```
> summary(model1_BMI_cholesterol)$coefficients
               Estimate    Std. Error   t value      Pr(>|t|)
(Intercept) -0.01053624 0.0088151009 -1.195249  2.319939e-01
BMI          0.01117947 0.0003029225 36.905369 2.651204e-295
cholesterol  0.14790644 0.0027126131 54.525448  0.000000e+00
> 
```

So, above equation for model after substituting value:

Cardio = -0.011 + (0.011) BMI + (0.147) cholesterol

Conclusion:

To conclude this report, we have completed the preliminary (Exploratory Data Analysis - EDA) analysis. From correlation matrix, we identified highly correlated variables which are age, weight and cholesterol. We know weight and height together gives BMI which is indicator of body mass. We calculated BMI. We also identified methods to answer the question and justify those methods. Interpretation of our dataset is much more precise now than before. Though, it was difficult to conclude the appropriate methods as I think we may need to perform extra pre procedures to get the data ready for modeling. We also need to know how we will be handling missing or unknown values in the whole dataset creating the model. Will it affect the performance of the model or not? We performed hypothesis test to test if presence or absence of cardiovascular disease is dependent or independent of age, weight, cholesterol and BMI and it can be concluded that yes, it is dependent. We also performed GLM model where accuracy was maximum 63% for all variable considered which states it is not good model to implement. Later we performed Linear regression. We can see its comparison in table below.

LM factor with respect to Cardio	R- square value	Fit or Unfit
Age and Weight	0.085	Fit
Age and Cholesterol	0.091	Fit
Weight and Cholesterol	0.072	Fit
BMI and Cholesterol	0.07	Fit

Looking at result of above analysis, it can be concluded that LM model fits best. Age, weight, cholesterol, and BMI does affect presence or absence of cardiovascular disease.

But after careful analysis, we believe variables are highly skewed which means model have most of rows with same value for above considered columns which makes it hard to capture the pattern. Hence, data needs to be more randomized than this and more information should have been included to increase rate of accuracy.

Reference:

1. The Analysis Factor. 2021. Generalized Linear Models in R, Part 2: Understanding Model Fit in Logistic Regression Output - The Analysis Factor. Retrieved July 3, 2021, from https://www.theanalysisfactor.com/r-glm-model-fit/

2. The Analysis Factor. 2021. Generalized Linear Models in R, Part 2: Understanding Model Fit in Logistic Regression Output - The Analysis Factor. Retrieved July 3, 2021, from https://www.theanalysisfactor.com/r-glm-model-fit/

3. Statinfer | Data Science starts here. 2021. 203.4.2 Calculating Sensitivity and Specificity in R Statinfer. Retrieved July 3, 2021, from https://statinfer.com/203-4-2-calculating-sensitivity-and-specificity-in-r

4. Rpubs.com. 2021. RPubs - How do I get P-values and critical values from R?. Retrieved July 3, 2021, from https://rpubs.com/mdlama/spring2017-lab6supp1

5. Support.google.com. 2021. CHISQ.DIST - Docs Editors Help. Retrieved July 3, 2021, from https://support.google.com/docs/answer/7003347?hl=en

6. Quora.com. 2021. *How to recode R-studio so I can find BMI - Quora*. Retrieved July 3, 2021, from https://www.quora.com/How-do-I-recode-R-studio-so-I-can-find-BMI

7. Porras, E., 2018. *Linear Regression in R*. Datacamp. Retrieved July 3, 2021, from https://www.datacamp.com/community/tutorials/linear-regression-R

Appendix:

```
library(reshape2)
library(gginference)
library(RColorBrewer)
library(GGally)
library(lattice)
library(olsrr)
library(performance)
library(Ecdat)
library(leaps)
library(lmtest)
library(visdat)
library(inspectdf)
library(skimr)
library(ggcorrplot)
library(gridExtra)
library(e1071)
library(lattice)
library(caret)
library(ISLR)
library(pROC)
library(glmnet)
library(Metrics)
library(dplyr)
library(psych)
library(ggplot2)
library(ggpubr)
library(tidyverse)
library(hrbrthemes)
library(viridis)
library(gridExtra)
library(corrplot)
library(scales)
library(lmSubsets)

# Formatting
F_cardio <- as.data.frame(read.csv(file.choose( ) , sep=";",header = TRUE,stringsAsFactors = FALSE) )
head(F_cardio)

#view(F_cardio)

describe(F_cardio)
summary(F_cardio)
glimpse(F_cardio)

#Excluding id
F_cardio <- select(F_cardio, -c(id))
View(F_cardio)

#Correlation Matrix:
cardioExplor <- F_cardio
correlation = cor(cardioExplor[,1:12])
cols<- colorRampPalette(c("red", "blue"))(20)
corrplot(correlation,  method ="number",col=cols,type="upper",
        title = "\n\n Correlation Plot Of Cardio train")



#Correlation plot
ggplot(melt(correlation_matrix), aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low="blue", mid="white", high="red") +
  coord_equal()
```

```
# #Changing variable to factor
# cols <- c("cholesterol","gluc", "smoke", "alco", "active", "cardio")
# F_cardio[,cols] <- lapply(F_cardio[,cols], factor)
# str(F_cardio)
```

#changing the male and female values into 0's and 1's

```
F_cardio$gender <- factor(F_cardio$gender, levels=c(1,2), labels=c(0,1))
head(F_cardio$gender)
```

#Cleaning data

#Checking for missing values - NA's

```
any(is.na(F_cardio))

F_cardio[F_cardio == "?"] <- NA
any(is.na(F_cardio))
```

#Outliers

#For Systolic Blood Pressure Range

```
#boxplot(F_cardio$ap_hi ~ F_cardio$cardio, main="Systolic blood pressure  by cardio", ylab = "Systolic blood pressure", xlab = "cardio")

ggplot(F_cardio, aes(x = ap_hi, y= cardio)) +
  geom_boxplot(fill="gray")+
  labs(title="Systolic blood pressure  by cardio",x="cardio", y = "Systolic blood pressure")+
  theme_classic()

#F_cardio <- F_cardio[!(F_cardio$ap_hi>370),]
#F_cardio <- F_cardio[!(F_cardio$ap_lo>360),]
```

#Replacing value with median

```
med_ap_hi <- median(F_cardio$ap_hi)
F_cardio$ap_hi[F_cardio$ap_hi < 90 | F_cardio$ap_hi > 120 ] = med_ap_hi
```

#Box plot after removing outliers

```
ggplot(F_cardio, aes(x = ap_hi, y= cardio)) +
  geom_boxplot(fill="gray")+
  labs(title="Systolic blood pressure  by cardio",x="cardio", y = "Systolic blood pressure")+
  theme_classic()



#----For Diastolic Blood Pressure Range-----
ggplot(F_cardio, aes(x = ap_lo, y= cardio)) +
  geom_boxplot(fill="gray")+
  labs(title="Diastolic blood pressure  by cardio",x="cardio", y = "Diastolic blood pressure")+
  theme_classic()

#Replacing value with median
med_ap_lo <- median(F_cardio$ap_lo)
F_cardio$ap_lo[F_cardio$ap_lo < 60 | F_cardio$ap_lo > 80 ] = med_ap_lo

#Box plot after removing outliers
ggplot(F_cardio, aes(x = ap_lo, y= cardio)) +
  geom_boxplot(fill="gray")+
  labs(title="Diastolic blood pressure by cardio",x="cardio", y = "Diastolic blood pressure")+
  theme_classic()
#-----------------------
#Scaling
#F_cardio$age <- gsub("(^\\d{2}).*", "\\1", F_cardio$age)

F_cardio$age <- F_cardio$age/365
F_cardio$age<-round(F_cardio$age,digits = 0)

#plotting age and cardio
a <- ggplot(F_cardio, aes(x = weight))
a + geom_histogram(aes(color = as.factor(cardio)), fill = "white",
                  position = "dodge") +
  scale_color_manual(values = c("#800000", "#E7B800"))

#calculating BMI
BMI = function(height,weight){(weight/(height/100)^2)}
F_cardio$BMI = BMI(F_cardio$height,F_cardio$weight)
head(F_cardio$BMI)

view(F_cardio)
#-------------------------------------------------
#F_cardio
copy_cardio <- F_cardio
#----------------------------------
#----------------------------------------
# chi- square on age and cardio
age <- copy_cardio %>%
  dplyr::group_by(cardio) %>%
  summarise(age = sum(age)) %>%
  as_tibble()

cardio_alpha   = 0.05
cardio_LoSig   = 1- cardio_alpha
cardio_k       = nrow(age)  ## No. of rows
cardio_DF      = cardio_k-1



age$Expected <- 1/cardio_k  ### Lets assume that expected frequencies are equal- 1/6th
age

cardio_Critical_Val <- qchisq(p= cardio_LoSig, cardio_DF, lower.tail=TRUE)

cat("Critical value: ",cardio_Critical_Val)
```

```
age_chisq = chisq.test(age$age,
                        p = age$Expected,  ## Values in Probability
                        correct = FALSE) # not to apply continuity correction

age_chisq

cat("The calculated t-value is:",age_chisq$statistic, "p-value is: ",age_chisq$p.value, " and alpha is:",cardio_alpha)

ifelse(wght_chisq$statistic < cardio_Critical_Val, "Fail to reject null  hypothesis ", " Rejecting null hypothesis")
#-----------------------
#-------------------------------
```

# chi- square on cholesterol

```
cholesterol <- copy_cardio %>%
  dplyr::group_by(cardio) %>%
  summarise(cholesterol = sum(cholesterol)) %>%
  as_tibble()

cardio_alpha   = 0.05
cardio_LoSig   = 1- cardio_alpha
cardio_k       = nrow(cholesterol)  ## No. of rows
cardio_DF      = cardio_k-1



cholesterol$Expected <- 1/cardio_k  ### Lets assume that expected frequencies are equal- 1/6th
cholesterol

cardio_Critical_Val <- qchisq(p= cardio_LoSig, cardio_DF, lower.tail=TRUE)

cat("Critical value: ",cardio_Critical_Val)
```

```
cholesterol_chisq = chisq.test(cholesterol$cholesterol,
                               p = cholesterol$Expected,  ## Values in Probability
                               correct = FALSE) # not to apply continuity correction


cholesterol_chisq


cat("The calculated t-value is:",cholesterol_chisq$statistic, "p-value is: ",cholesterol_chisq$p.value, " and alpha is:",cardio_alpha)


ifelse(cholesterol_chisq$statistic < cardio_Critical_Val, "Fail to reject null  hypothesis ", " Rejecting null hypothesis")


#---------------------------------
#-------------------------------
# chi- square on Weight and cardio


weight <- copy_cardio %>%
  dplyr::group_by(cardio) %>%
  summarise(weight = sum(weight)) %>%
  as_tibble()


cardio_alpha   = 0.05
cardio_LoSig   = 1- cardio_alpha
cardio_k       = nrow(weight)  ## No. of rows
cardio_DF      = cardio_k-1


weight$Expected <- 1/cardio_k  ### Lets assume that expected frequencies are equal- 1/6th
weight


cardio_Critical_Val <- qchisq(p= cardio_LoSig, cardio_DF, lower.tail=TRUE)


cat("Critical value: ",cardio_Critical_Val)
```

```
weight_chisq = chisq.test(weight$weight,
                          p = weight$Expected,  ## Values in Probability
                          correct = FALSE) # not to apply continuity correction


weight_chisq


cat("The calculated t-value is:",weight_chisq$statistic, "p-value is: ",weight_chisq$p.value, " and alpha is:",cardio_alpha)


ifelse(weight_chisq$statistic < cardio_Critical_Val, "Fail to reject null  hypothesis ", " Rejecting null hypothesis")
#-------------------------
#---------------------------------------
```

#Chi- square on BMI and cardio

```
BMI <- copy_cardio %>%
  dplyr::group_by(cardio) %>%
  summarise(BMI = sum(BMI)) %>%
  as_tibble()


cardio_alpha   = 0.05
cardio_LoSig   = 1- cardio_alpha
cardio_k       = nrow(BMI)  ## No. of rows
cardio_DF      = cardio_k-1



BMI$Expected <- 1/cardio_k  ### Lets assume that expected frequencies are equal- 1/6th
BMI


cardio_Critical_Val <- qchisq(p= cardio_LoSig, cardio_DF, lower.tail=TRUE)


cat("Critical value: ",cardio_Critical_Val)
```

```
BMI_chisq = chisq.test(BMI$BMI,
                    p = BMI$Expected,  ## Values in Probability
                    correct = FALSE) # not to apply continuity correction


BMI_chisq
cat("The calculated t-value is:",BMI_chisq$statistic, "p-value is: ",BMI_chisq$p.value, " and alpha is:",cardio_alpha)
ifelse(BMI_chisq$statistic < cardio_Critical_Val, "Fail to reject null  hypothesis ", " Rejecting null hypothesis")


#-------------------------------------------
#-----------LM model----------------------
# Imapact of age and weight on cardio
F_cardio$cardio <- as.numeric(F_cardio$cardio)
ls(F_cardio)
model1_age_weight <- lm(cardio ~ age + weight, data = F_cardio)
model1_age_weight
summary(model1_age_weight)
summary(model1_age_weight)$coefficients


# Imapact of age and cholesterol on cardio
model1_age_cholesterol <- lm(cardio ~ age + cholesterol, data = F_cardio)
model1_age_cholesterol
summary(model1_age_cholesterol)
summary(model1_age_cholesterol)$coefficients


#------------------------------
#--------------------------------------
# Imapact of weight and cholesterol on cardio
model1_weight_cholesterol <- lm(cardio ~ weight + cholesterol, data = F_cardio)
model1_weight_cholesterol
summary(model1_weight_cholesterol)
summary(model1_weight_cholesterol)$coefficients
#-------------------
#--------------------------
# Impact of BMI and Cholesterol on cardio

F_cardio$cardio <- as.numeric(F_cardio$cardio
model1_BMI_cholesterol <- lm(cardio ~ BMI + cholesterol, data = F_cardio)
model1_BMI_cholesterol
summary(model1_BMI_cholesterol)
summary(model1_BMI_cholesterol)$coefficients

#-----------------------------
#-----------------------------------
#-----------GLM----------------------

set.seed(123)
trainIndex <- sample(c(TRUE,FALSE), nrow(F_cardio), replace = TRUE, prob = c(0.7,0.3))
train_data <- F_cardio[trainIndex,]
test_data <- F_cardio[!trainIndex,]

F_cardio$cardio <- as.factor(F_cardio$cardio)

## Impact of age and weight on cardio
model2_age_weight <- glm(cardio ~age + weight, family = "binomial", data = train_data)

#disable scientific notation for model summary
options(scipens=999)
summary(model2_age_weight)
summary(model2_age_weight)$coef

# fitting the data
train_data$pred <- predict(model2_age_weight, train_data, type = "response")
```

```
train_data$pred_label <- as.factor(ifelse(train_data$pred >= 0.5, "1", "0"))


train_data$cardio <- as.factor(train_data$cardio)

train_data$cardio

train_data$pred_label

train_data$pred


# Confusion Matrix on Train Data

cM1 <- confusionMatrix(train_data$cardio, train_data$pred_label)

cM1


#Test dataset

test_data$pred <- predict(model2_age_weight, test_data, type = "response")

test_data$pred_label <- as.factor(ifelse(test_data$pred >= 0.5, "1", "0"))

test_data$cardio <- as.factor(test_data$cardio)




#Confusion Matrix on Test Data

conf1<-confusionMatrix(test_data$cardio,test_data$pred_label, )

conf1


#define object to plot and calculate AUC

rocobj <- roc(as.ordered(test_data$cardio), as.ordered(test_data$pred_label), ordered = TRUE)


auc <- round(auc(as.ordered(test_data$cardio), as.ordered(test_data$pred_label)),4)


ggroc(rocobj, colour = 'steelblue', size = 2) +

  ggtitle(paste0('ROC Curve ', '(AUC = ', auc, ')'))


#-----------------------------------------

#---------------------------------------------------
```

```
#Imapact of age and cholesterol on cardio
F_cardio$cholesterol
F_cardio$cardio <- as.factor(F_cardio$cardio)

model2_age_cholesterol <- glm(cardio ~age + cholesterol, family = "binomial", data = train_data)

#disable scientific notation for model summary
options(scipens=999)
summary(model2_age_cholesterol)
summary(model2_age_cholesterol)$coef

# fitting the data
train_data$pred <- predict(model2_age_cholesterol, train_data, type = "response")
train_data$pred_label <- as.factor(ifelse(train_data$pred >= 0.5, "1", "0"))
train_data$cardio <- as.factor(train_data$cardio)
train_data$cardio
train_data$pred_label
train_data$pred

# Confusion Matrix on Train Data
cM2 <- confusionMatrix(train_data$cardio, train_data$pred_label)
cM2

# test dataset
test_data$pred <- predict(model2_age_cholesterol, test_data, type = "response")
test_data$pred_label <- as.factor(ifelse(test_data$pred >= 0.5, "1", "0"))
test_data$cardio <- as.factor(test_data$cardio)

#Confusion Matrix on Test Data
conf2<-confusionMatrix(test_data$cardio,test_data$pred_label, )
conf2
```

```
#define object to plot and calculate AUC
rocobj <- roc(as.ordered(test_data$cardio), as.ordered(test_data$pred_label), ordered = TRUE)
auc <- round(auc(as.ordered(test_data$cardio), as.ordered(test_data$pred_label)),4)
ggroc(rocobj, colour = 'steelblue', size = 2) +
  ggtitle(paste0('ROC Curve ', '(AUC = ', auc, ')'))

#----------------------------------------
#-----------------------------------------
#Imapact of weight and cholesterol on cardio

F_cardio$weight
F_cardio$cardio <- as.factor(F_cardio$cardio)

model2_weight_cholesterol <- glm(cardio ~weight + cholesterol, family = "binomial", data = train_data)

#disable scientific notation for model summary
options(scipens=999)
summary(model2_weight_cholesterol)
summary(model2_weight_cholesterol)$coef

# fitting the data
train_data$pred <- predict(model2_weight_cholesterol, train_data, type = "response")
train_data$pred_label <- as.factor(ifelse(train_data$pred >= 0.5, "1", "0"))

train_data$cardio <- as.factor(train_data$cardio)
train_data$cardio
train_data$pred_label
train_data$pred

# Confusion Matrix on Train Data
cM3 <- confusionMatrix(train_data$cardio, train_data$pred_label)
cM3
```

```
# test dataset
test_data$pred <- predict(model2_weight_cholesterol, test_data, type = "response")
test_data$pred_label <- as.factor(ifelse(test_data$pred >= 0.5, "1", "0"))
test_data$cardio <- as.factor(test_data$cardio)

#Confusion Matrix on Test Data
conf3<-confusionMatrix(test_data$cardio,test_data$pred_label, )
conf3

#define object to plot and calculate AUC
rocobj <- roc(as.ordered(test_data$cardio), as.ordered(test_data$pred_label), ordered = TRUE)
auc <- round(auc(as.ordered(test_data$cardio), as.ordered(test_data$pred_label)),4)
ggroc(rocobj, colour = 'steelblue', size = 2) +
  ggtitle(paste0('ROC Curve ', '(AUC = ', auc, ')'))

#-----------------------------------
#-------------------------------------------
#Impact of BMI and cholesterol on cardio
F_cardio$BMI
F_cardio$cardio <- as.factor(F_cardio$cardio)

model2_BMI_cholesterol <- glm(cardio ~BMI + cholesterol, family = "binomial", data = train_data)

#disable scientific notation for model summary
options(scipens=999)
summary(model2_BMI_cholesterol)
summary(model2_BMI_cholesterol)$coef

# fitting the data
train_data$pred <- predict(model2_BMI_cholesterol, train_data, type = "response")
train_data$pred_label <- as.factor(ifelse(train_data$pred >= 0.5, "1", "0"))
```

```
train_data$cardio <- as.factor(train_data$cardio)
train_data$cardio
train_data$pred_label
train_data$pred

# Confusion Matrix on Train Data
cM4 <- confusionMatrix(train_data$cardio, train_data$pred_label)
cM4

#test dataset
test_data$pred <- predict(model2_BMI_cholesterol, test_data, type = "response")
test_data$pred_label <- as.factor(ifelse(test_data$pred >= 0.5, "1", "0"))
test_data$cardio <- as.factor(test_data$cardio)

#Confusion Matrix on Test Data
conf4<-confusionMatrix(test_data$cardio,test_data$pred_label, )
conf4

#define object to plot and calculate AUC
rocobj <- roc(as.ordered(test_data$cardio), as.ordered(test_data$pred_label), ordered = TRUE)
auc <- round(auc(as.ordered(test_data$cardio), as.ordered(test_data$pred_label)),4)
ggroc(rocobj, colour = 'steelblue', size = 2) +
ggtitle(paste0('ROC Curve ', '(AUC = ', auc, ')'))
```