(1) VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY - HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY

FACULTY OF COMPUTER SCIENCE AND ENGINEERING, CHEMICAL ENGINEERING

REPORT OF PROBABILITY AND STATISTIC’S PROJECT

Lecturer: Professor Nguyễn Tiến Dũng

Class: DTQ1

Group: 04

Major: Computer Science and Engineering, Chemical Engineering

No.	Student ID	Name	Contribution	Note
1	2052971	Nguyễn Anh Hào	20%	Leader
2	2152456	Võ Tấn Cường	20%	
3	2153036	Lê Nguyễn Việt Tiến	20%	
4	2150033	Nguyễn Minh Tiến	20%	
5	2052396	Lê Minh Gia Bảo	20%	

Ho Chi Minh City, August 2023

Teamwork result

No.	**Student ID**	**First name**	**Last name**	**Project's contribution**
1	2052971	Nguyễn Anh	Hào	20%
2	2152456	Võ Tấn	Cường	20%
3	2153036	Lê Nguyễn Việt	Tiến	20%
4	2150033	Nguyễn Minh	Tiến	20%
5	2052396	Lê Minh Gia	Bảo	20%

List of contents

Acknowledgement

We would like to thank Doctor Nguyen Tien Dung for offering our team the chance to work with R studio software in the first place. We also appreciate your thorough understanding of statistics and probability. Now is our chance to run the R studio. We also acknowledge the importance of R Studio as current mathematical curriculum. The program allows us to increase both our expertise and our ideas for new projects.

Achievement

The goal of my team in completing this project is to understand the test methods taught in probability and statistics and be able to apply those methods to the analysis of data in related domains, particularly in this report's focus on predicting coronary heart disease based on the physical health figure. We employed the processes of hypothesis Ordinal Logistic Regression to collect data. We also wish to learn how to use the R Studio program to handle and calculate data.

1 Introduction

1.1 Topic introduction and requirements

The term "heart disease" is used to describe a wide range of disorders that affect the structure and function of the heart. A kind of heart illness called coronary heart disease occurs when the heart's arteries are unable to provide the organ with adequate oxygen-rich blood. In the US, it is the top cause of death. According to the Centers for Disease Control and Prevention, coronary artery disease affects 18.2 million adult Americans, making it the most prevalent form of heart disease in the country**.**

According to the latest WHO data published in 2020 Coronary Heart Disease Deaths in Viet Nam reached 91,939 or 13.41% of total deaths. The age adjusted Death Rate is 86.67 per 100,000 of population ranks Viet Nam #125 in the world. Review other causes of death by clicking the links below or choose the full health profile.

When contributing to the diagnosis, this data gives information regarding, etc. of the patients (those with/without stroke). The prerequisites for the subject are that we be well-educated, possess solid statistical data analysis skills, and be able to apply statistical techniques to foretell the statistical value of CHD as it relates to numerous aspects.

Data we collected stored in the following link:

https://www.kaggle.com/datasets/billbasener/coronary-heart-disease

The main variables to consider are:

- Sbp (continuous variable)
- Tobacco (continuous variable)
- Ldl (continuous variable)
- Adiposity (continuous variable)
- Famhist (Present or Absent)
- Typea (continuous variable)
- Obesity (continuous variable)
- Alcohol (continuous variable)
- Age (continuous variable)
- Chd (1: Yes, 0: No)

1.2 Statistical methods

In the health sciences it is quite common to carry out studies designed to determine the influence of one or more variables upon a given response variable. When this response variable is numerical, simple or multiple regression techniques are used, depending on the case. If the response variable is a qualitative variable (dichotomic or polychotomic), as for example the presence or absence of a disease, linear regression methodology is not applicable, and simple or multinomial logistic regression is used, as applicable.

In this project, we are going to use "**Logistic regression model**" to evaluate the factors affecting the risk of Stroke.

2 Theoretical basis

2.1 Logistic regression model

Binary or binomial logistic regression and multinomial logistic regression are statistical models that assess the connection between a dependent qualitative, dichotomous variable and a variable having more than two values, one or more covariables, whether qualitative or quantitative, that are independent explanatory factors.

To illustrate, we will use the example of a binary output variable Y and the conditional probability P(Y = 1|X = x) that we wish to model as a function of x; any unknown parameters in the function will be approximated using maximum likelihood.

2.2 Why choosing Logistic regression model instead of Linear regresson model?

The simplest solution is to assume that p(x) is a linear function of x. Each increase or decrease in an x component would have a significant impact on the likelihood. Since p must be between 0 and 1, and since linear functions are unbounded, there is a conceptual issue. Furthermore, we frequently observe "diminishing returns" in empirical data; that is, when p is already high (or little), the same change in p takes a larger change in x than when p is near to 1/2. This cannot be done using linear models. The next most apparent solution is to make log p(x) a linear function of x, which multiplies the probability by a constant amount when an input variable is changed. Finally, the easiest modification of log p which has an unbounded range is the logistic (or logit) transformation, $\log\frac{p}{1-p}$. We can make this a linear function of x without fear of nonsensical results.

2.3 Effects logistic regression

For estimation and prediction purposes, the probabilities are severely limited. First, they are bound to the range 0 to 1. This implies that if the real effect of variable X on the outcome of variable Y exceeds 1, interpretation may be problematic. The second limit, the probability cannot be negative. Assuming that the effect of an independent variable on variable Y is negative, the logistic regression coefficient interpretation is meaningless. One problem is that the regression coefficient should only be positive. To solve the above two problems, we have a two-step approach through performing two transformations. First, we convert the probabilities in Odds (O) to:

$$0 = \frac{p}{1-p} = \frac{\text{Probability of the event happening}}{\text{Probability of the event not happening}}$$

$$p = \frac{0}{1+0}$$

O is Odd; P is the probability

That is, the odds that an event will occur is the ratio of the number of times it is expected that the event will happen to the number of times it is expected that the event will not happen. This is a direct relationship between Odds (Y=1) and the probability Y=1. Thus, given that Odds can have infinity, the probability with Odds now allows the regression coefficient to have any value. The next step is to solve the second problem. The relationship between Odds and probability, extending a little algebraically, we can restate the above Odds (O) formula in terms of the logarithm of Odds (Y=1):

$loge[Odd(Yi = 1)] = \beta 0 + \beta 1X1i + \beta 2X2i + ... + \beta kXki$

To calculate the logarithm for a random case in the population for the value of an independent variable or a covariate. Add to the dependent sea Y the value 1 (for example, 1 (vote for Obama in 2008), 0 (vote for McCain in 2008, in the US election). Assume that the probability of voting for Obama P(Y=1) is 0.218 ; and so 1-P = 0.782. We calculate Odds as: Odds=0.218/0.782=0.279. This value just shows us the resulting Odds, now they we have to continue to assume that the logistic regression coefficients involved are in the correct direction, so we need to use the logarithmic formula of Odds. Accordingly, the natural logarithm ($loge$, symbol ln) of Odds (eg $ln0, 279 = -1, 276$).Therefore, the logarithm of the probability of voting for Obama is -1.276'. Thus, if we just stop at probabilistic predictions, we can arrive at false results (a positive number).

Second, the true effect of the covariates involved is underrated (underestimated). The main advantage of logarithmic Odds is that the coefficients are constrained, and that they can be negative as well as positive, ranging from negative infinity to positive infinity. Stated this way, logistic regression looks exactly like multiple regression on the right side of the logarithmic Odds equation. The left side of the equation is not the score of Y. It is the logarithm of Odds (Y=1). This means that each unit of X has the effect of β on the logarithm of Odds of Y.

2.4 Estimation of logistic regression model with Maximum Likelihood

Since categorical variables are used in logistic regression, the ordinary least squares, the (OLS) approach is useless (it assumes a normally distributed dependent variable). To find a satisfactory fit of the parameters, a more broad estimator is therefore used. The greatest likelihood estimation refers to this. An interactive estimating method called maximum likelihood is used to choose parameter estimates that increase the possibility of observing a sample dataset. For a given set of X values, maximum reasonably finds coefficient estimates in logistic regression that maximize the logarithm of the probability of observing a specific set of dependent variable values in the sample.

Since logistic regression employs the greatest likelihood approach, it's possible that the coefficient of determination (R-) cannot be calculated directly. As a result, there are two problems with how to interpret logistic regression: As a broad null hypothesis, how can we first measure the goodness of fit? Second, how do we calculate each variable

- The Akaike information criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. The lower is the AIC value, the better is the model.

The formula for calculating the AIC value:$AIC = - 2 \times log(Likelihood) + 2 \times k = 2[k - log(Likelihood)]$

2.5 Statistical inference and null hypothesis

First question, how can we also measure the goodness of fit – a general null hypothesis?The statistical inferences, together with the null hypothesis, are interpreted according tothe following steps:

- The first step in the regression interpretation is to evaluate the global null hypothesis that the independent seas do not have any relationship with Y. In the OLS regression method, we perform This is equal to testing whether R2 must be 0 in the population using an F-test. While logistic regression uses the method of maximum likelihood (non-OLS): The null hypothesis H0 is β0 = β1 = β2 = 0. We measure the size of the residuals from this model with a statistical logarithm. likelihood statistic.

- We then estimate the model again, assuming that the null hypothesis is false, that we findthe maximum reasonable value of the coefficients β in the sample. Again, we measure the size of the residuals from this model with a statistical logarithm of reasonableness.

- Finally, we compare the two statistics by computing a test statistic: $-2(lnLnull - lnLmodel)$

This statistic tells us how much residual (or prediction error) can be reduced using X variables. The null hypothesis suggests that the reduction is 0 ; if the statistic is large enough (in a chi-square test with df = number of independent variables), we reject the null hypothesis. Here, we conclude that at least one independent variable has a logarithmic Odds effect.

SPSS also runs R^2 statistics to help evaluate the strength of associations. But it as a pseudo R^2, should not be interpreted because logistic regression does not use R2 like linear regression.

Second question, how do we estimate the partial effect of each variable X? When the general null hypothesis is rejected, we evaluate the partial effects of the predictors.

As in multiple linear regression, in logistic regression this implies that the null hypothesis for each independent variable included in the equation. The null hypothesis is that each regression coefficient is zero, or it has no effect on the logarithm of Odds.

Each coefficient estimator B has a standard error – the extent to which, on average, we would expect B to vary from one sample to another by chance. To check the significance of B, a test statistic (not a t-test, but a Wald Chi-squared) is calculated, with 1df – degrees of freedom.

It should be remembered that the coefficient B expresses the effects of a unit change of X on logarithmic Odds.

In education, the effect is positive, as education increases, the logarithm of Odds also increases. The Exp(B) value of an independent variable X is used to predict the probability of an event occurring based on the change in one unit change in an independent variable when all other independent variables are held constant. It indicates that when it is increased by one, the Odds for the "yes" event is multiplied by one value of the value Exp(B) (this is a function e to the power B, say 1.05, which is an increase of 5%).

3 Read the data:

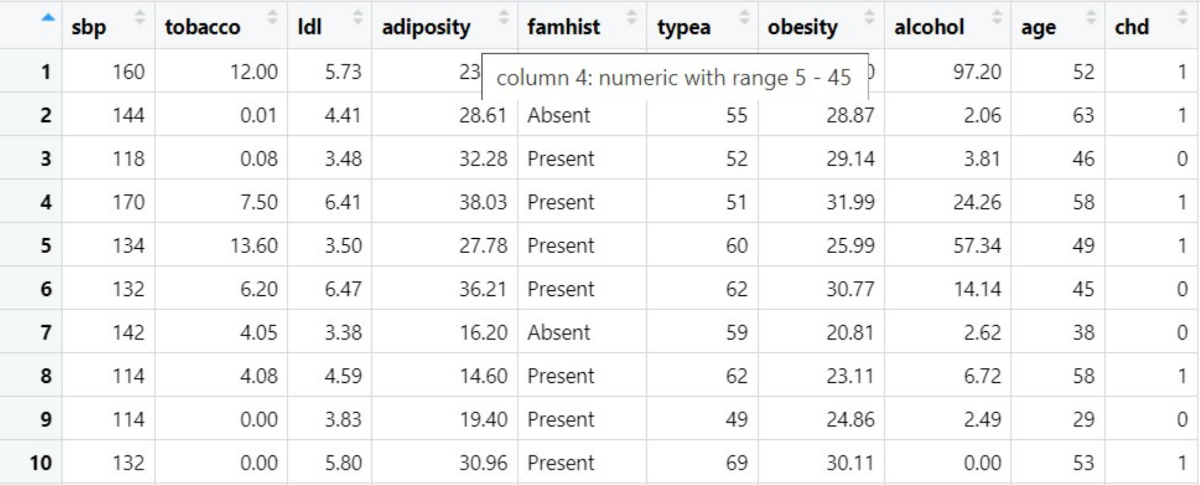
3.1 Loading library:

```
library(tidyverse)
library(rsample)
library(recipes)
library(parsnip)
library(yardstick)
```

3.2 Import data:

- Read the data with new column names.
- We read the data by using **read.csv()** function in order to read the file **“CHDdata”** and save it in **heart_disease_dataset** variable.

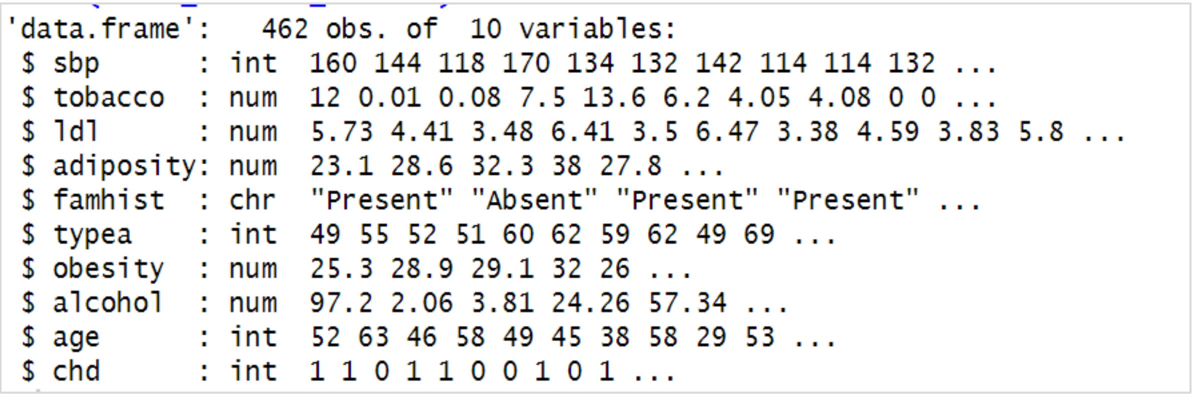
```
heart_data <- read.csv(file = "CHDdata.csv", header = FALSE)
colnames(heart_data) <- c("sbp", "tobacco", "ldl", "adiposity",
"famhist", "typea", "obesity", "alcohol", "age", “chd”)
```



	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
1	160	12.00	5.73	23	column 4: numeric with range 5 - 45		)	97.20	52	1
2	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	1
3	118	0.08	3.48	32.28	Present	52	29.14	3.81	46	0
4	170	7.50	6.41	38.03	Present	51	31.99	24.26	58	1
5	134	13.60	3.50	27.78	Present	60	25.99	57.34	49	1
6	132	6.20	6.47	36.21	Present	62	30.77	14.14	45	0
7	142	4.05	3.38	16.20	Absent	59	20.81	2.62	38	0
8	114	4.08	4.59	14.60	Present	62	23.11	6.72	58	1
9	114	0.00	3.83	19.40	Present	49	24.86	2.49	29	0
10	132	0.00	5.80	30.96	Present	69	30.11	0.00	53	1

Commented: Heart_data contains 462 observations and 10 variables.

```
Str(heart_data)
```



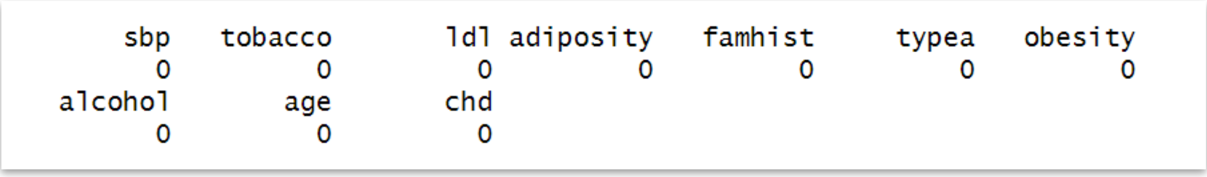
```
'data.frame':   462 obs. of  10 variables:
 $ sbp      : int  160 144 118 170 134 132 142 114 114 132 ...
 $ tobacco  : num  12 0.01 0.08 7.5 13.6 6.2 4.05 4.08 0 0 ...
 $ ldl      : num  5.73 4.41 3.48 6.41 3.5 6.47 3.38 4.59 3.83 5.8 ...
 $ adiposity: num  23.1 28.6 32.3 38 27.8 ...
 $ famhist  : chr  "Present" "Absent" "Present" "Present" ...
 $ typea    : int  49 55 52 51 60 62 59 62 49 69 ...
 $ obesity  : num  25.3 28.9 29.1 32 26 ...
 $ alcohol  : num  97.2 2.06 3.81 24.26 57.34 ...
 $ age      : int  52 63 46 58 49 45 38 58 29 53 ...
 $ chd      : int  1 1 0 1 1 0 0 1 0 1 ...
```

Commented: The dataset contains 462 observations of 10 variables.

3.3 Data cleaning:

Due to the dataset have many NA value, we must replace missing data to NA status then couting number of missing value

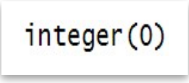
```
heart_data [heart_data =="N/A"]<-NA
apply(is.na(heart_data),2,sum)
```



```
      sbp   tobacco       ldl adiposity   famhist     typea   obesity 
        0         0         0         0         0         0         0 
  alcohol       age       chd 
        0         0         0 
```

There is no missing values.

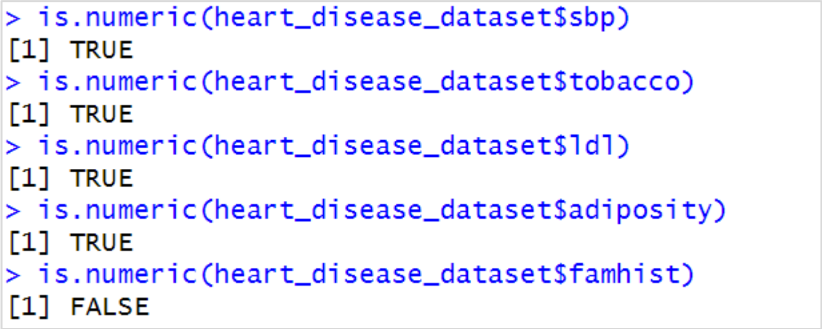
```
heart_data <-na.omit(heart_data)
apply(is.na(heart_data),2,which)
```



```
integer(0)
```

Then, we clarify data and checking variables whether they are numeric values or not

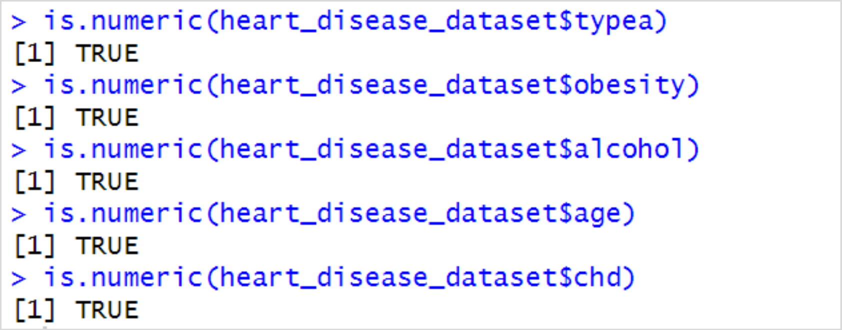
```
is.numeric(heart_data$sbp)

is.numeric(heart_data$tobacco)

is.numeric(heart_data$ldl)

is.numeric(heart_data$adiposity)

is.numeric(heart_data$famhist)
```



```
> is.numeric(heart_disease_dataset$sbp)
[1] TRUE
> is.numeric(heart_disease_dataset$tobacco)
[1] TRUE
> is.numeric(heart_disease_dataset$ldl)
[1] TRUE
> is.numeric(heart_disease_dataset$adiposity)
[1] TRUE
> is.numeric(heart_disease_dataset$famhist)
[1] FALSE
```

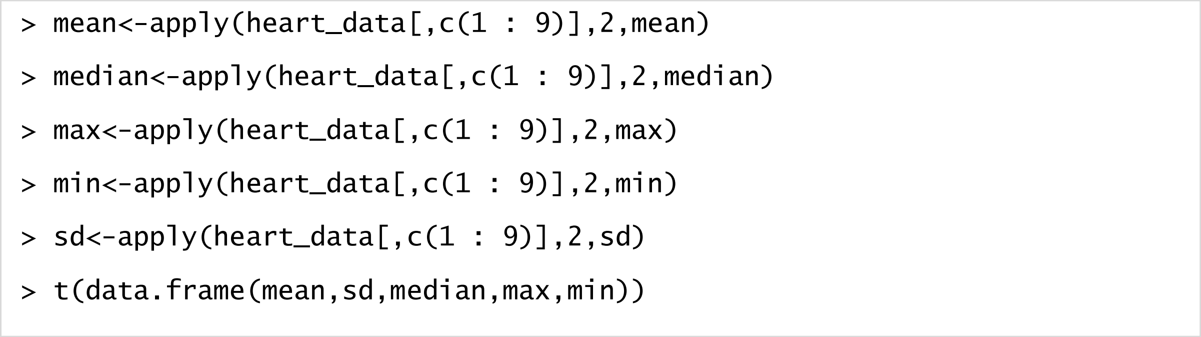
```
heart_data <- heart_data [, -5]
```

```
is.numeric(heart_data$typea)

is.numeric(heart_data$obesity)

is.numeric(heart_data$alcohol)

is.numeric(heart_data$age)

is.numeric(heart_data$chd)
```

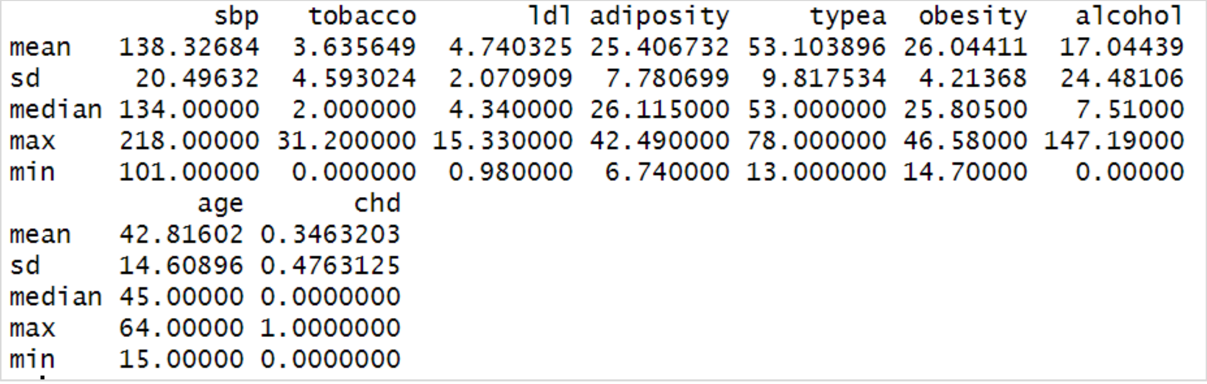


```
> is.numeric(heart_disease_dataset$typea)
[1] TRUE
> is.numeric(heart_disease_dataset$obesity)
[1] TRUE
> is.numeric(heart_disease_dataset$alcohol)
[1] TRUE
> is.numeric(heart_disease_dataset$age)
[1] TRUE
> is.numeric(heart_disease_dataset$chd)
[1] TRUE
```

Finally, we calculate descriptive statistics for variables.



```
> mean<-apply(heart_data[,c(1 : 9)],2,mean)

> median<-apply(heart_data[,c(1 : 9)],2,median)

> max<-apply(heart_data[,c(1 : 9)],2,max)

> min<-apply(heart_data[,c(1 : 9)],2,min)

> sd<-apply(heart_data[,c(1 : 9)],2,sd)

> t(data.frame(mean,sd,median,max,min))
```

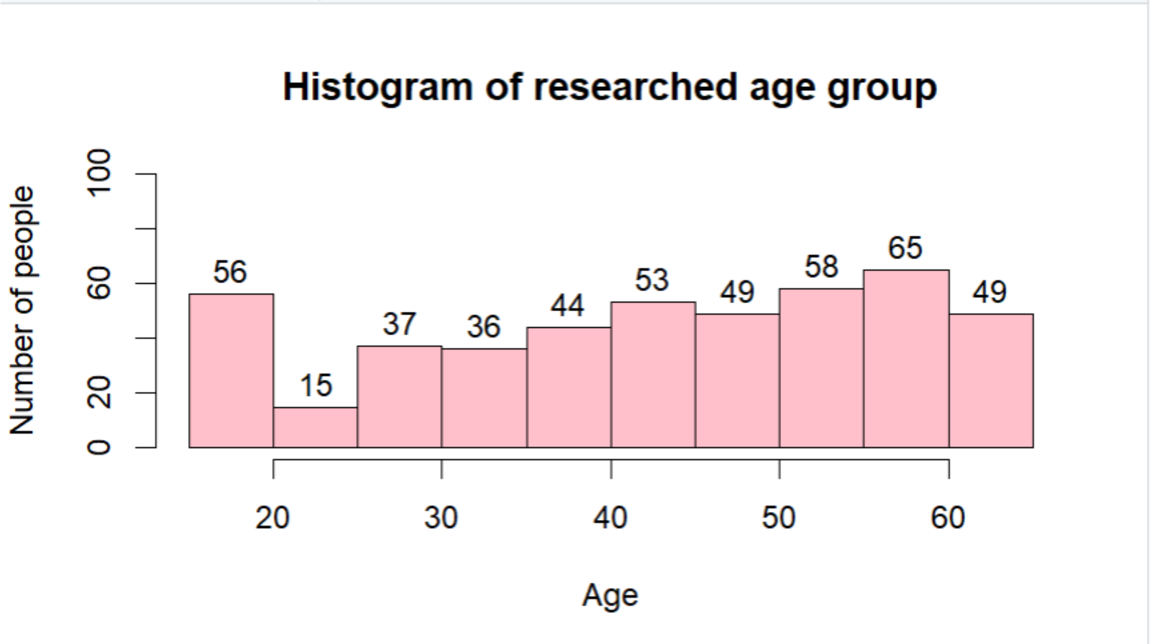


```
             sbp   tobacco       ldl adiposity     typea  obesity   alcohol
mean   138.32684  3.635649  4.740325 25.406732 53.103896 26.04411  17.04439
sd      20.49632  4.593024  2.070909  7.780699  9.817534  4.21368  24.48106
median 134.00000  2.000000  4.340000 26.115000 53.000000 25.80500   7.51000
max    218.00000 31.200000 15.330000 42.490000 78.000000 46.58000 147.19000
min    101.00000  0.000000  0.980000  6.740000 13.000000 14.70000   0.00000
            age       chd
mean   42.81602 0.3463203
sd     14.60896 0.4763125
median 45.00000 0.0000000
max    64.00000 1.0000000
min    15.00000 0.0000000
```

4 *Data visualization:*

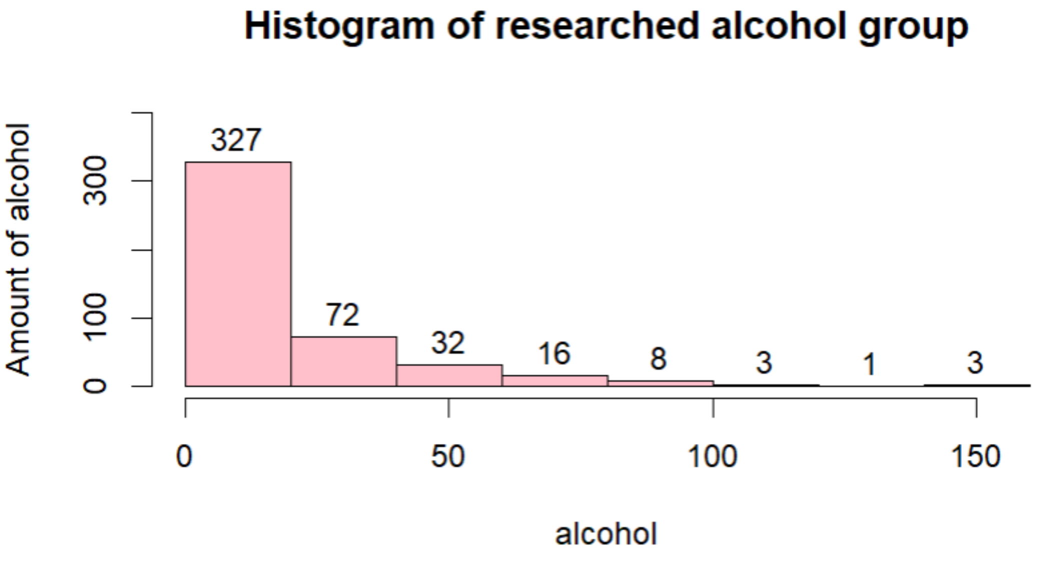
4.1. Draw a histogram showing the distribution of quantitative variables, plot a bar plot or pie chart of quantitative statistics for each classifier.

```
hist(heart_data$age,xlab="Age",ylab="Number of people",main="Histo
gram of researched age group",ylim=c(0,100),labels=T,col="pink")
```



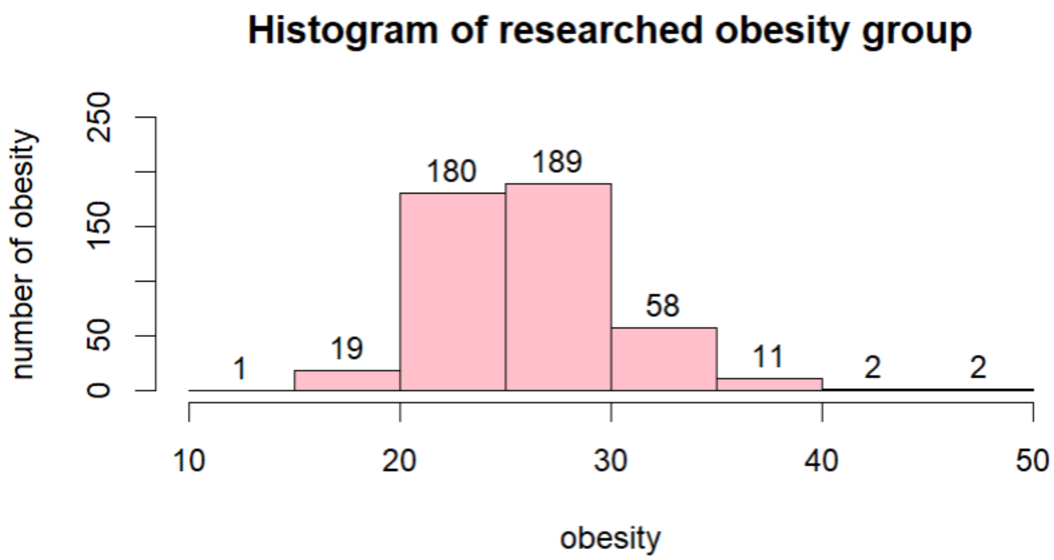
Comment: • Based on the variable "Age" graph, it is found that the age of the patients is concentrated mostly at 40-60 years old, the highest is at 50-60 years old (418 patients) and the lowest is in over 60+-year-old age group.

```
hist(heart_data$alcohol,xlab="alcohol",ylab="Amount of alcohol",main="Histogram of researched alcohol group",ylim=c(0,400),labels=T,col="pink")
```



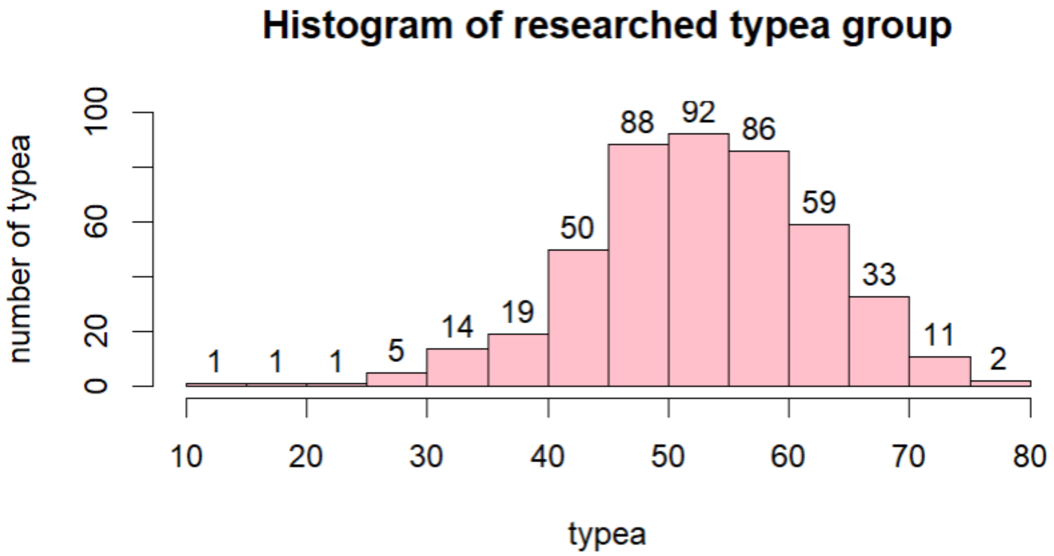
Comment: Heavy drinking, on the other hand, is linked to a number of poor health outcomes, including heart conditions. Excessive alcohol intake can lead to high blood pressure, heart failure or stroke. Excessive drinking can also contribute to cardiomyopathy, a disorder that affects the heart muscle.

```
hist(heart_data$alcohol,xlab="alcohol",ylab="Amount of alcohol",main="Histogram of researched alcohol group",ylim=c(0,400),labels=T,col="pink")
```

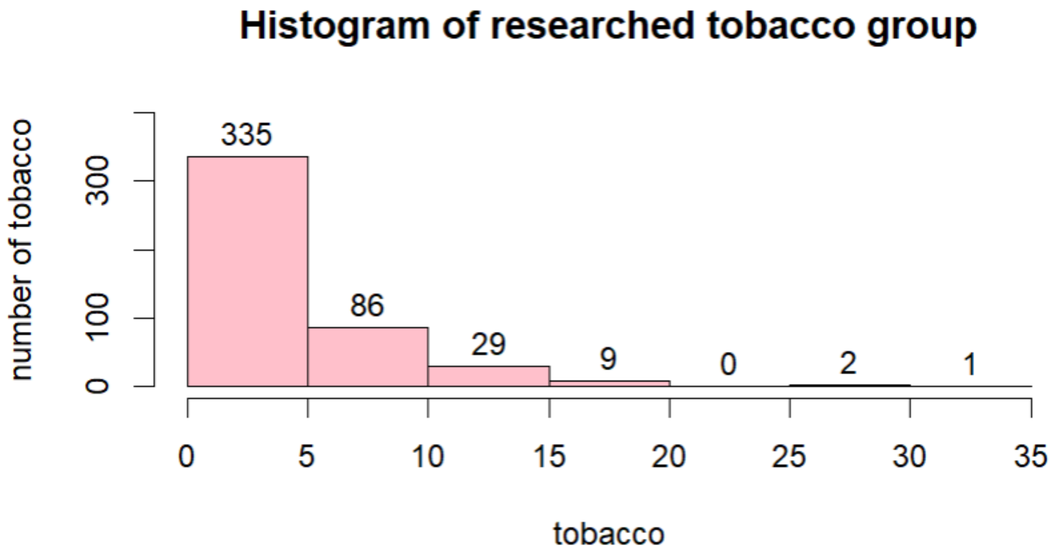


```
hist(heart_data$typea,xlab="typea",ylab="number of typea",main="Histogra
m of researched typea group",ylim=c(0,100),labels=T,col="pink")
```

Comment: Obesity are defined as abnormal or excessive fat accumulation that presents a risk to health. A body mass index (BMI) over 25 is considered overweight, and over 30 is obese. Obesity increases your risk of developing many other risk factors for heart disease. It also triggers inflammatory processes that can harm your cardiovascular system, and it can lead to structural or functional changes in the heart itself.

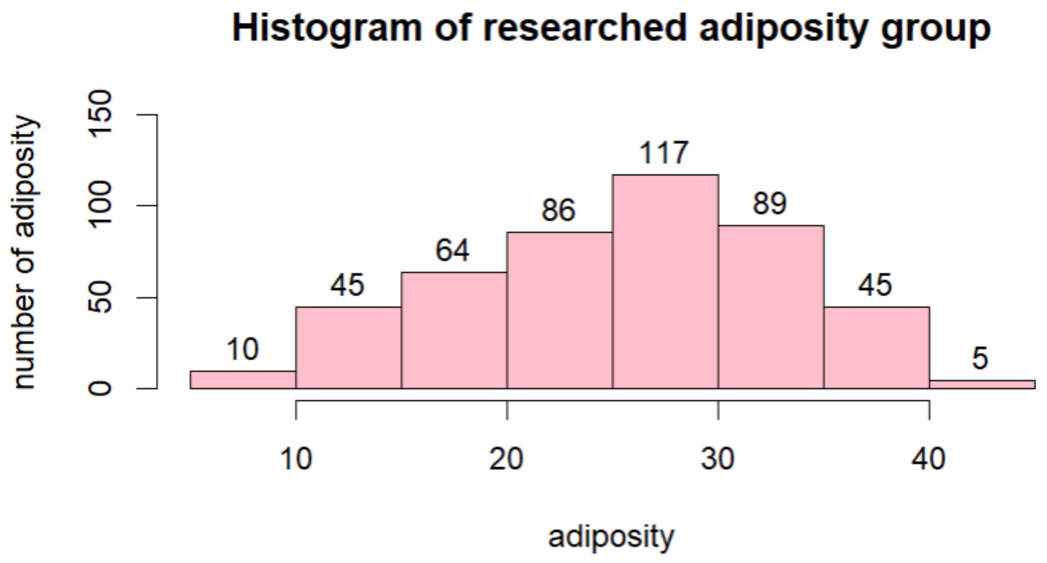


```
hist(heart_data$tobacco,xlab="tobacco",ylab="number of tobacco",main="Hi
stogram of researched tobacco group",ylim=c(0,400),labels=T,col="pink")
```



```
hist(heart_data$adiposity,xlab="adiposity",ylab="number of adiposity",ma
in="Histogram of researched adiposity group",ylim=c(0,150),labels=T,col=
```

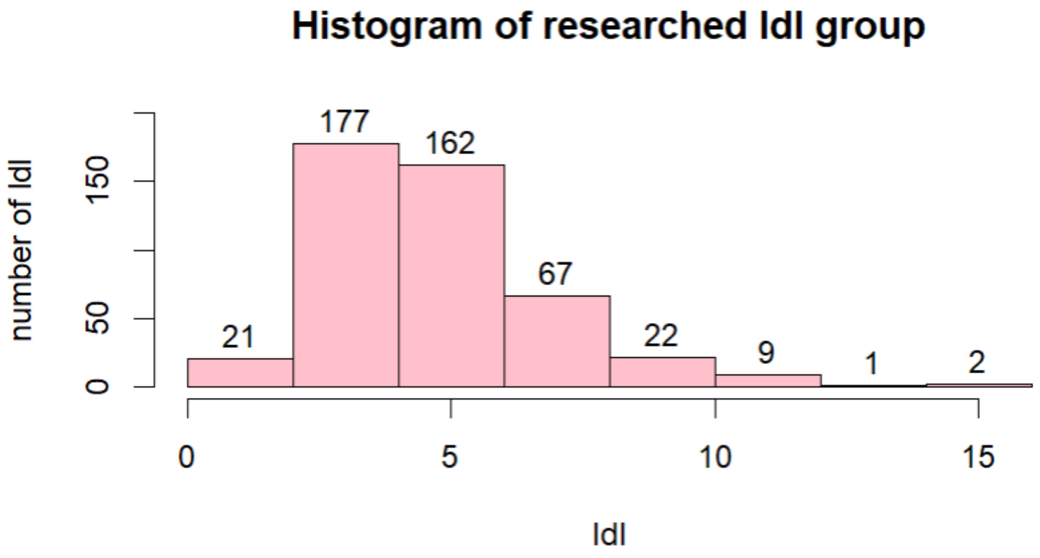
Comment: Tobacco increases the formation of plaque in blood vessels. Coronary Heart Disease occurs when arteries that carry blood to the heart muscle are narrowed by plaque or blocked by clots. Chemicals in cigarette smoke cause the blood to thicken and form clots inside veins and arteries.



```
hist(heart_data$ldl,xlab="ldl",ylab="number of ldl",main="Histogram of r
esearched ldl group",ylim=c(0,200),labels=T,col="pink")
```

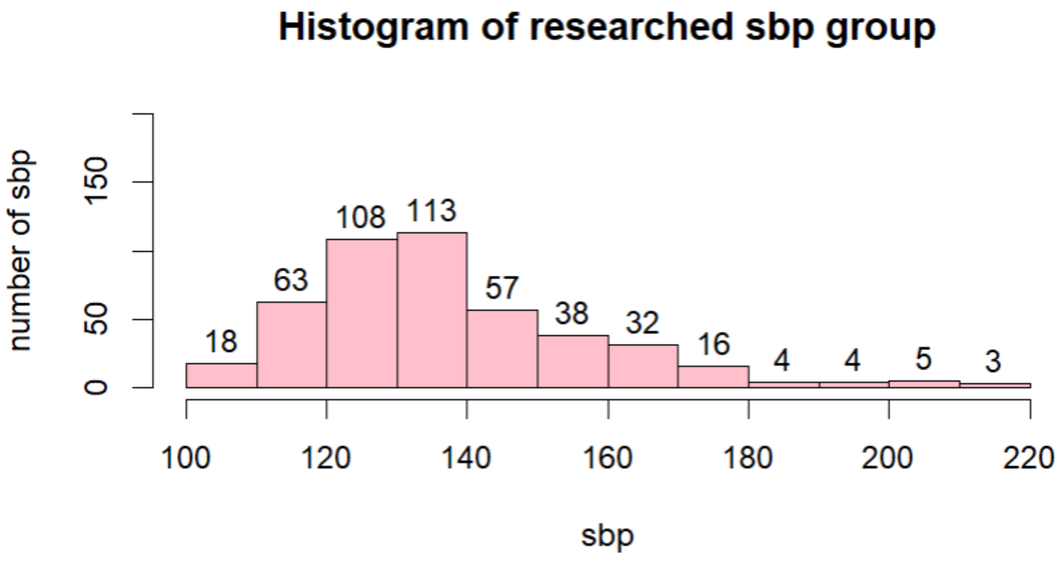
Comment: Adipose tissue, otherwise known as body fat, is a connective tissue that extends throughout your body. It's found under your skin (subcutaneous fat), between your internal organs (visceral fat) and even in the inner cavities of bones (bone marrow adipose tissue). Adipose tissue is crucial for health.

However, having too much — or too little — can cause its regulatory systems to malfunction. Healthy levels vary by age and sex, ranging between 10% and 35%.



```
hist(heart_data$sbp,xlab="sbp",ylab="number of sbp",main="Histogram of r
esearched sbp group",ylim=c(0,200),labels=T,col="pink")
```

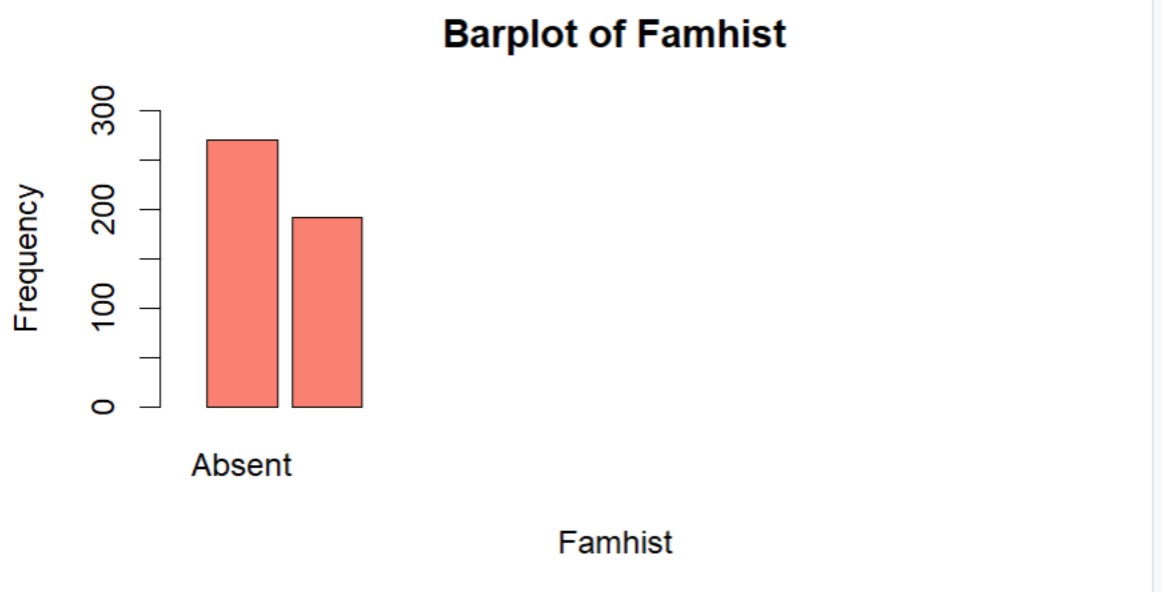
Comment: Low-density lipoprotein cholesterol is an important causal risk factor for atherosclerotic cardiovascular disease . However, a sizable proportion of middle-aged individuals with elevated LDL-C level have not developed coronary atherosclerosis as assessed by coronary artery calcification.



Comment: This is the pressure of the blood flowing through your arteries as your heart contracts and pumps blood around the body. It is measured in millimetres of mercury (mmHg) and it is a more frequent cardiovascular risk factor and has a greater impact on blood pressure staging, though this can vary with age, sex and country.

4.2 Barplot:

```
barplot(table(heart_data$famhist),xlab="Famhist",ylab="Frequency",main="
Barplot of Famhist",col="salmon",ylim=c(0,300),xlim=c(0,12))
```



4.3 Pie Chart:

```
Famhist_P<-heart_data[heart_data$famhist=="Present",]
table(Famhist_P$chd)
```

```
 0  1
96 95
```

```
quantity<-c(96,95)
presentchd<-c("No","Yes")
percentage<-round(quantity/sum(quantity)*100,2)
presentchd<-paste(presentchd,percentage)
presentchd<-paste(presentchd,"%")
```

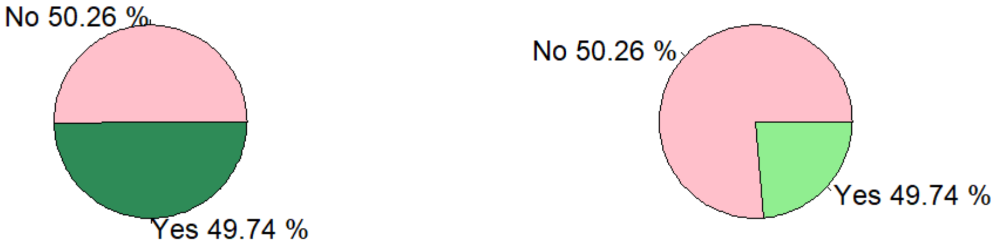
```
Famhist_A<-heart_data[heart_data$famhist=="Absent",]
table(Famhist_A$chd)
```

```
  0   1
206  64
```

```
Quantity1<-c(206,64)
absentchd<-c("No","Yes")
percentage<-round(quantity/sum(quantity)*100,2)
absentchd<-paste(absentchd,percentage)
absentchd<-paste(absentchd,"%")
```

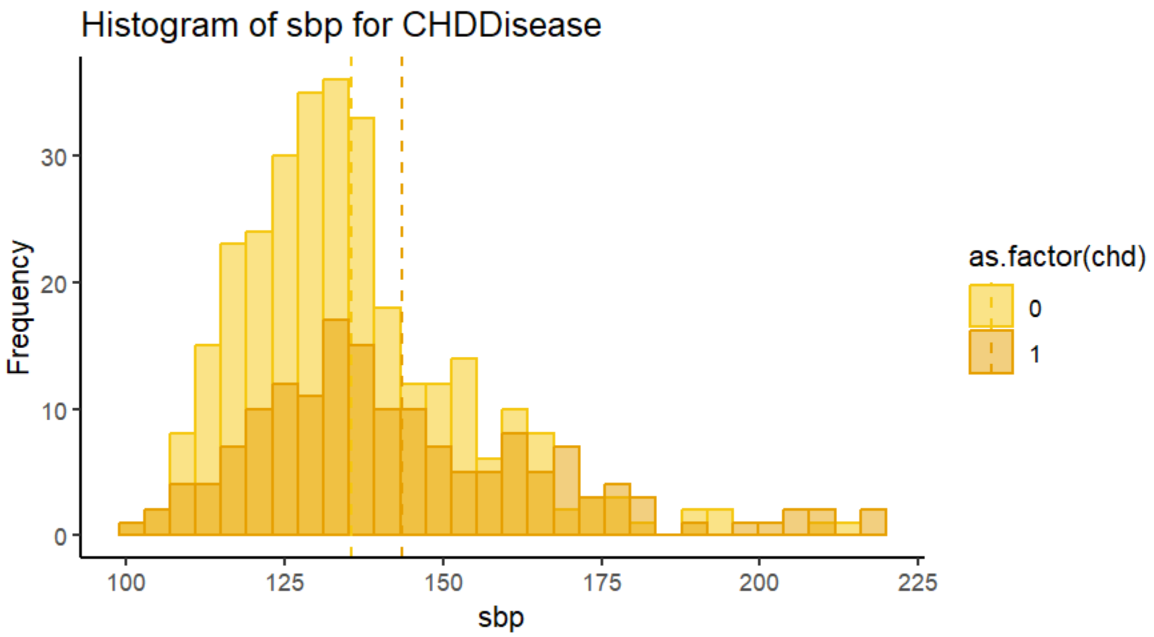
After the descriptive data is converted to percentage, pie charts should be drawn

```
par(mfrow=c(1,2))
pie(quantity,labels = presentchd,main="Famhist present chd percentage",c
ol=c("pink","seagreen"))
pie(quantity1,labels = absentchd,main="Famhist absent chd percentage",co
l=c("pink","lightgreen"))
```



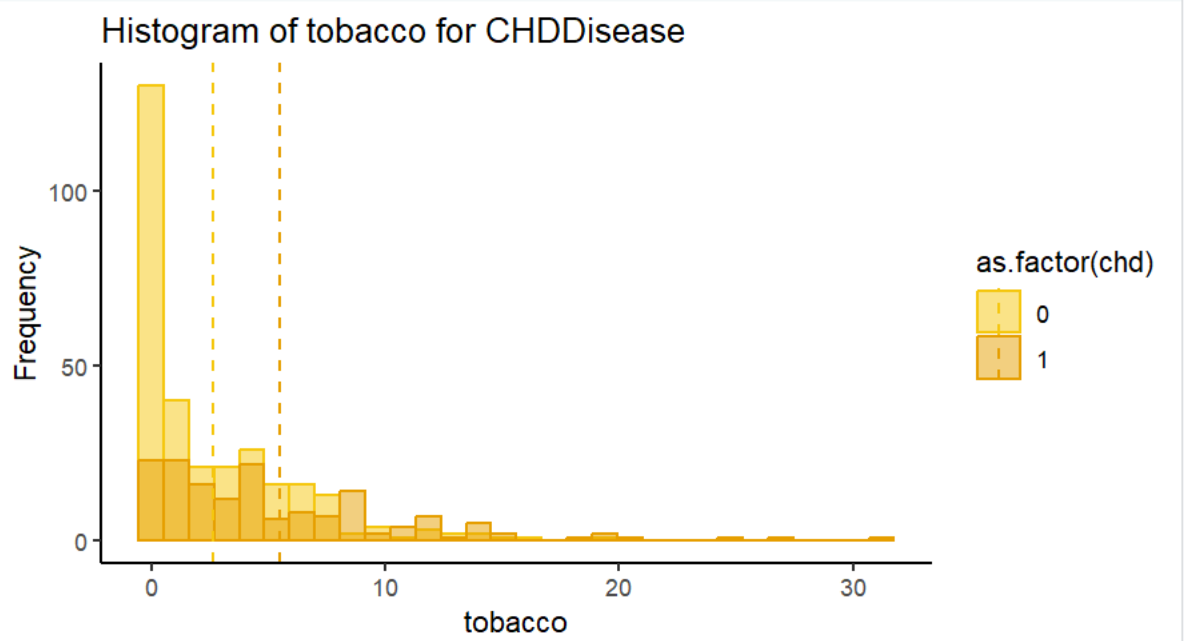
4.4 Histogram:

```
library(ggplot2)
> library(plyr)
> mu_sbp<-ddply(heart_data,"chd",summarise,grp.mean=mean(sbp))
> ggplot(heart_data,aes(x=sbp,color=as.factor(chd),fill=as.factor(chd)))
+
+     geom_histogram(position="identity",alpha=0.5)+
+     geom_vline(data=mu_sbp,aes(xintercept=grp.mean,color=as.factor(chd
)),linetype="dashed")+
+     scale_color_manual(values=c("999999","#E69F00","#56B4E9"))+
+     scale_fill_manual(values=c("999999","#E69F00","#56B4E9"))+
+     labs(title="Histogram of sbp for CHDDisease",x="sbp",y="Frequency"
)+
+     theme_classic()
```



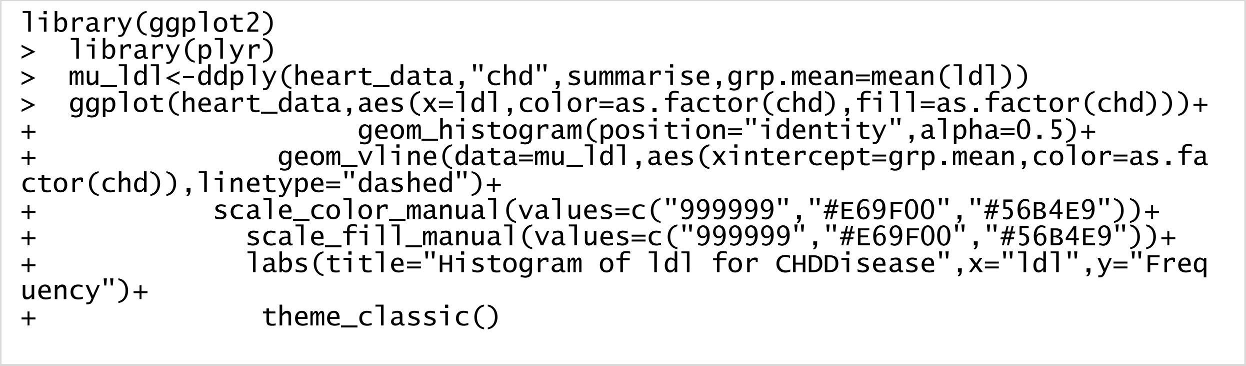
Comment: Through the graph, it can be concluded that survey participants with an sbp index between 125 and 150 have a higher rate of CHD than those with an sbp index outside the upper range.

```
library(ggplot2)
>   library(plyr)
>   mu_tobacco<-ddply(heart_data,"chd",summarise,grp.mean=mean(tobacco))
>  ggplot(heart_data,aes(x=tobacco,color=as.factor(chd),fill=as.factor(c
hd)))+
+               geom_histogram(position="identity",alpha=0.5)+
+          geom_vline(data=mu_tobacco,aes(xintercept=grp.mean,color=as.f
actor(chd)),linetype="dashed")+
+       scale_color_manual(values=c("999999","#E69F00","#56B4E9"))+
+         scale_fill_manual(values=c("999999","#E69F00","#56B4E9"))+
+         labs(title="Histogram of tobacco for CHDDisease",x="tobacco",y
="Frequency")+
+          theme_classic()
```

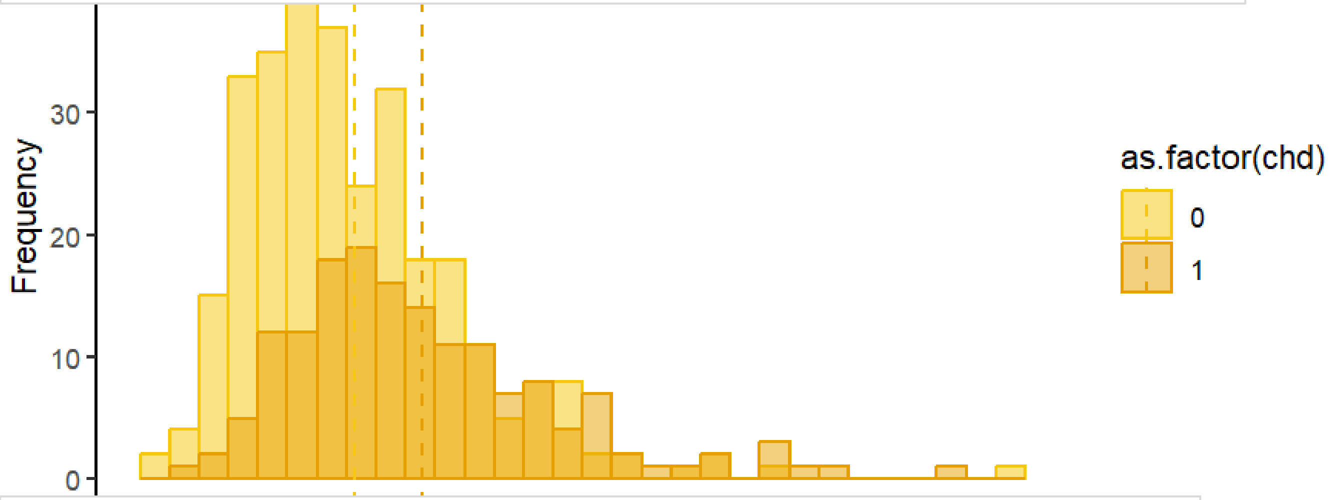


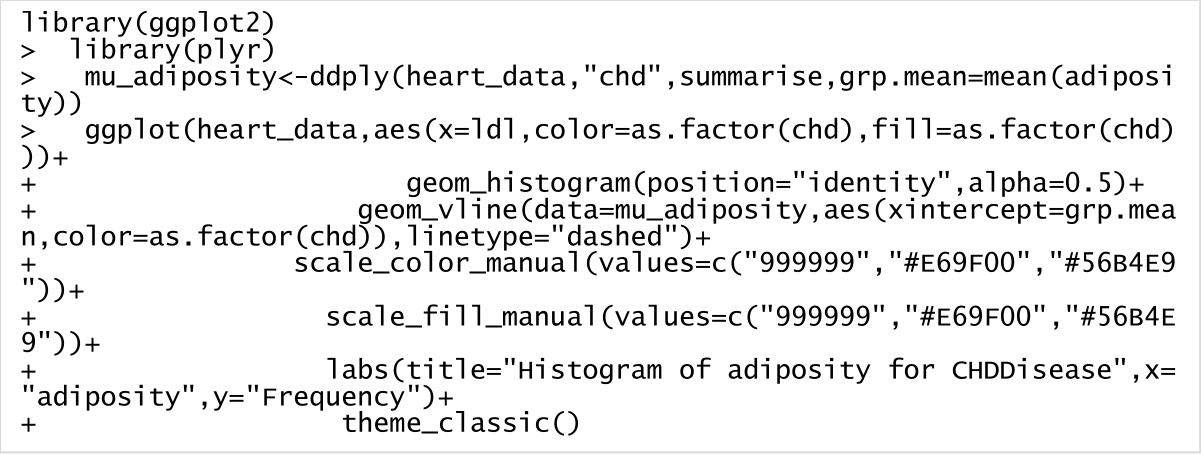
```
library(ggplot2)
>  library(plyr)
>  mu_ldl<-ddply(heart_data,"chd",summarise,grp.mean=mean(ldl))
>  ggplot(heart_data,aes(x=ldl,color=as.factor(chd),fill=as.factor(chd))
)+
+                    geom_histogram(position="identity",alpha=0.5)+
+               geom_vline(data=mu_ldl,aes(xintercept=grp.mean,color=as.
factor(chd)),linetype="dashed")+
+           scale_color_manual(values=c("999999","#E69F00","#56B4E9"))+
+             scale_fill_manual(values=c("999999","#E69F00","#56B4E9"))+
+             labs(title="Histogram of ldl for CHDDisease",x="ldl",y="Fr
equency")+
+              theme_classic()
```

Comment: Based on cigarette usage, the graph above depicts the proportion of survey respondents that have CHD syndrome. It is evident that non-users who use less frequently have a lower risk of developing cardiovascular disease, despite the fact that there are continuously more projections for numbers above 10.



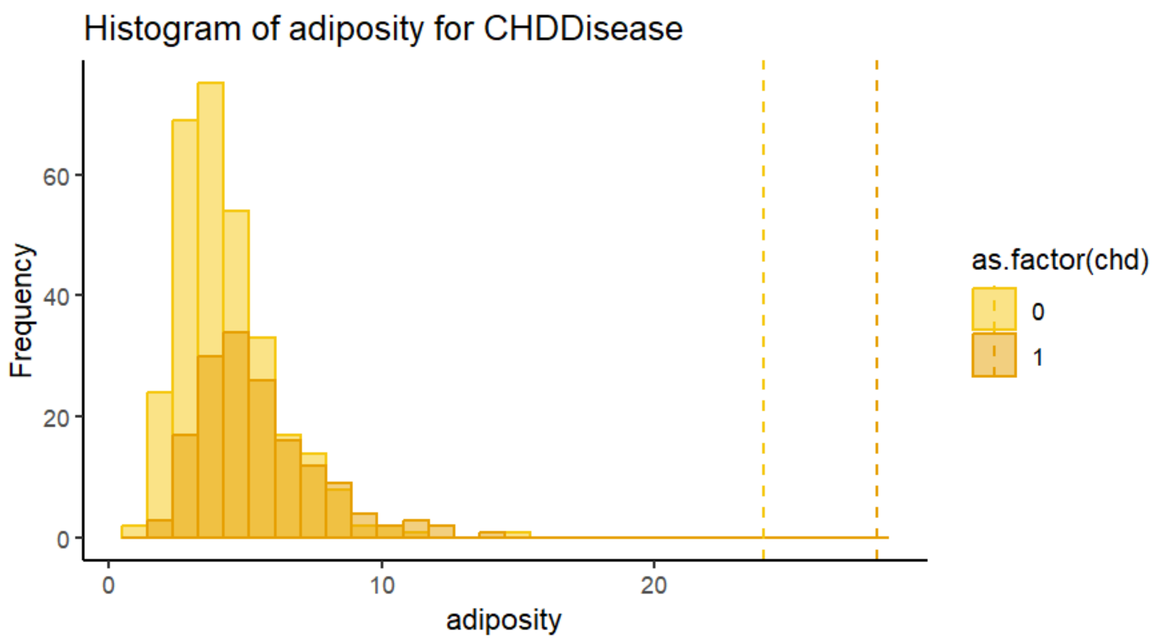
```
library(ggplot2)
>  library(plyr)
>  mu_ldl<-ddply(heart_data,"chd",summarise,grp.mean=mean(ldl))
>  ggplot(heart_data,aes(x=ldl,color=as.factor(chd),fill=as.factor(chd)))+
+                     geom_histogram(position="identity",alpha=0.5)+
+                geom_vline(data=mu_ldl,aes(xintercept=grp.mean,color=as.fa
ctor(chd)),linetype="dashed")+
+           scale_color_manual(values=c("999999","#E69F00","#56B4E9"))+
+             scale_fill_manual(values=c("999999","#E69F00","#56B4E9"))+
+             labs(title="Histogram of ldl for CHDDisease",x="ldl",y="Freq
uency")+
+              theme_classic()
```





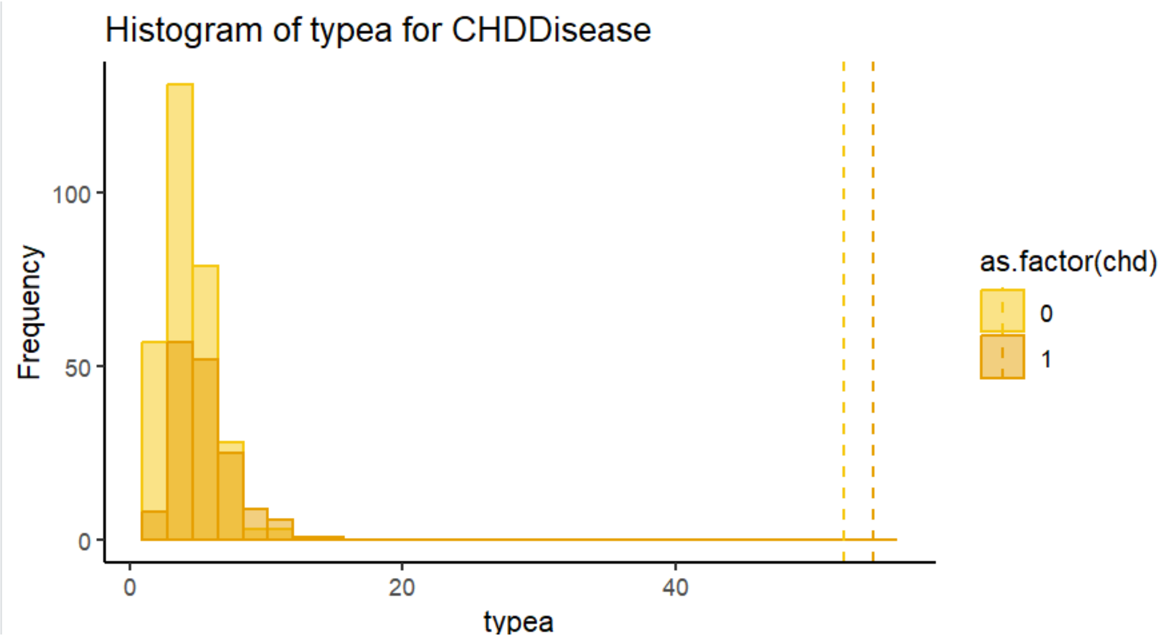
```
library(ggplot2)
>  library(plyr)
>   mu_adiposity<-ddply(heart_data,"chd",summarise,grp.mean=mean(adiposi
ty))
>   ggplot(heart_data,aes(x=ldl,color=as.factor(chd),fill=as.factor(chd)
))+
+                        geom_histogram(position="identity",alpha=0.5)+
+                     geom_vline(data=mu_adiposity,aes(xintercept=grp.mea
n,color=as.factor(chd)),linetype="dashed")+
+                 scale_color_manual(values=c("999999","#E69F00","#56B4E9
"))+
+                   scale_fill_manual(values=c("999999","#E69F00","#56B4E
9"))+
+                   labs(title="Histogram of adiposity for CHDDisease",x=
"adiposity",y="Frequency")+
+                    theme_classic()
```

Comment: The average ldl level of the patients with the disease was much higher than that of the patients without the disease



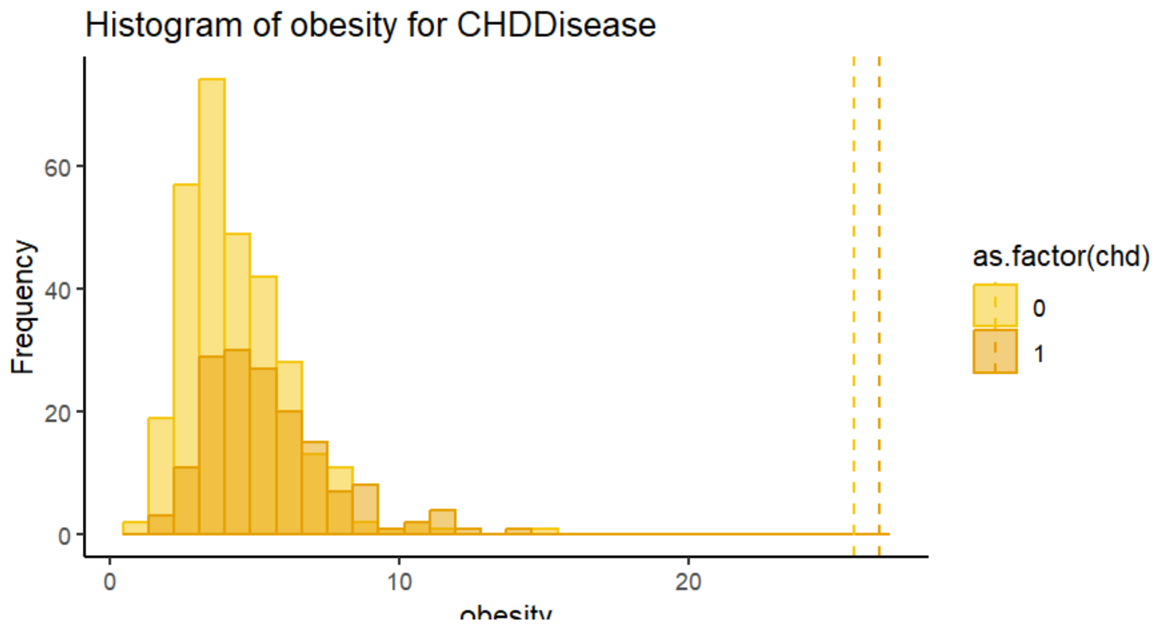
```
library(ggplot2)
>  library(plyr)
>    mu_typea<-ddply(heart_data,"chd",summarise,grp.mean=mean(typea))
>    ggplot(heart_data,aes(x=ldl,color=as.factor(chd),fill=as.factor(chd
)))+
+                          geom_histogram(position="identity",alpha=0.5)
+
+                         geom_vline(data=mu_typea,aes(xintercept=grp.me
an,color=as.factor(chd)),linetype="dashed")+
+                     scale_color_manual(values=c("999999","#E69F00","#5
6B4E9"))+
+                       scale_fill_manual(values=c("999999","#E69F00","#
56B4E9"))+
+                       labs(title="Histogram of typea for CHDDisease",x
="typea",y="Frequency")+
+                        theme_classic()
```

Comment: From the chart, we can clearly see that the surveyed person who have CHD disease occur the upper level adiposity in their body than the one who don't have this disease.



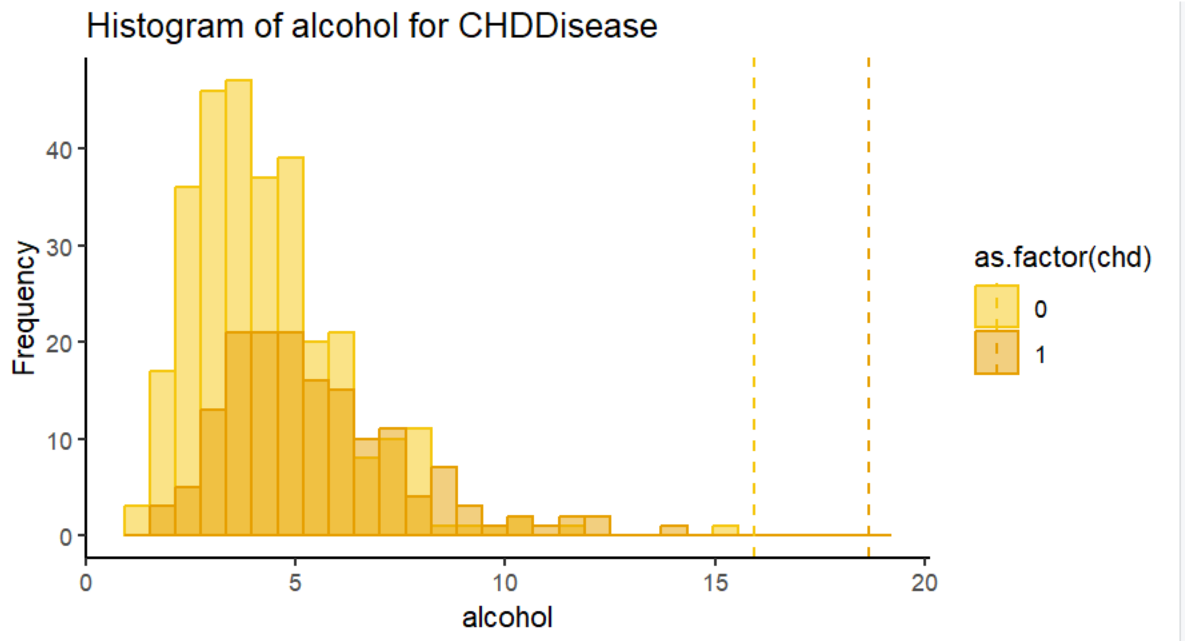
```
library(ggplot2)
>  library(plyr)
>      mu_obesity<-ddply(heart_data,"chd",summarise,grp.mean=mean(obesit
y))
>      ggplot(heart_data,aes(x=ldl,color=as.factor(chd),fill=as.factor(c
hd)))+
+                                   geom_histogram(position="identity",a
lpha=0.5)+
+                                   geom_vline(data=mu_obesity,aes(xinte
rcept=grp.mean,color=as.factor(chd)),linetype="dashed")+
+                               scale_color_manual(values=c("999999","#E
69F00","#56B4E9"))+
+                                 scale_fill_manual(values=c("999999","#
E69F00","#56B4E9"))+
+                              labs(title="Histogram of obesity for CHDD
isease",x="obesity",y="Frequency")+
+                                  theme_classic()
```

Comment:



```
library(ggplot2)
library(plyr)
>       mu_alcohol<-ddply(heart_data,"chd",summarise,grp.mean=mean(alcoh
ol))
>       ggplot(heart_data,aes(x=ldl,color=as.factor(chd),fill=as.factor(
chd)))+
+                                 geom_histogram(position="identi
ty",alpha=0.5)+
+                                 geom_vline(data=mu_alcohol,aes(
xintercept=grp.mean,color=as.factor(chd)),linetype="dashed")+
+                             scale_color_manual(values=c("999999
","#E69F00","#56B4E9"))+
+                               scale_fill_manual(values=c("99999
9","#E69F00","#56B4E9"))+
+                            labs(title="Histogram of alcohol for
CHDDisease",x="alcohol",y="Frequency")+
+                                theme_classic()
```

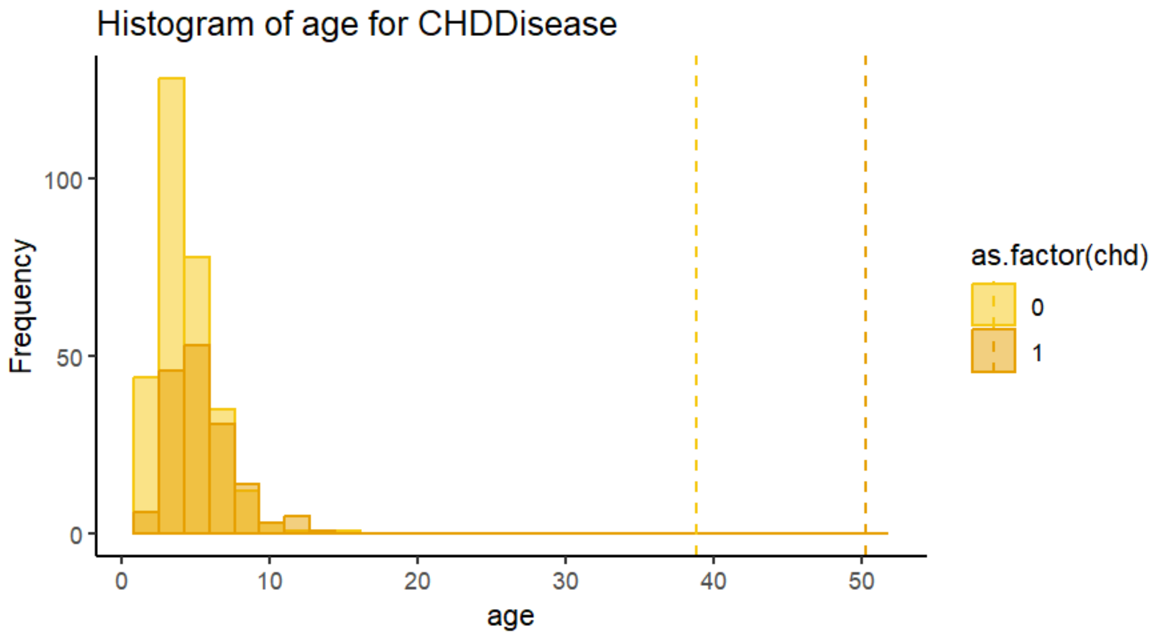
Comment: According to the graph, patients with CHD disease were more likely to be obese individuals than the remaining group. Furthermore, compared to other symptoms, the likelihood of having CHD illness is quite low.



```
library(ggplot2)
> library(plyr)
> mu_age<-ddply(heart_data,"chd",summarise,grp.mean=mean(age))
> ggplot(heart_data,aes(x=ldl,color=as.factor(chd),fill=as.factor(chd))
+
+                                              geom_histogram(position="i
dentity",alpha=0.5)+
+                                             geom_vline(data=mu_age,aes(
xintercept=grp.mean,color=as.factor(chd)),linetype="dashed")+
+                                          scale_color_manual(values=c("9
99999","#E69F00","#56B4E9"))+
+                                            scale_fill_manual(values=c("
999999","#E69F00","#56B4E9"))+
+                                           labs(title="Histogram of age fo
r CHDDisease",x="age",y="Frequency")+
+                                             theme_classic()
```

Comment: Based on the bar chat, we can clearly know that the patients who get CHD disease use alcolhol more usualy than the people hardly ever or sometimes use alcolhol.

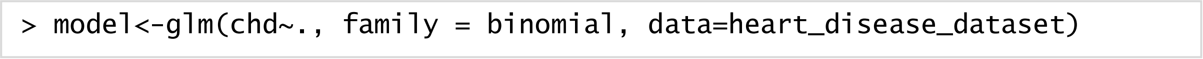
PROBABILITY AND STATTISTIC'S PROJECT



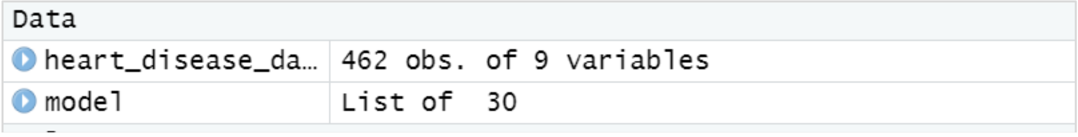
Comment:

The average age of the patients with the CHD disease higher than that of the patients without the disease. It can be concluded that the elder have a higher risk of CHD.

5.Building a multi-variable Logistic regression model

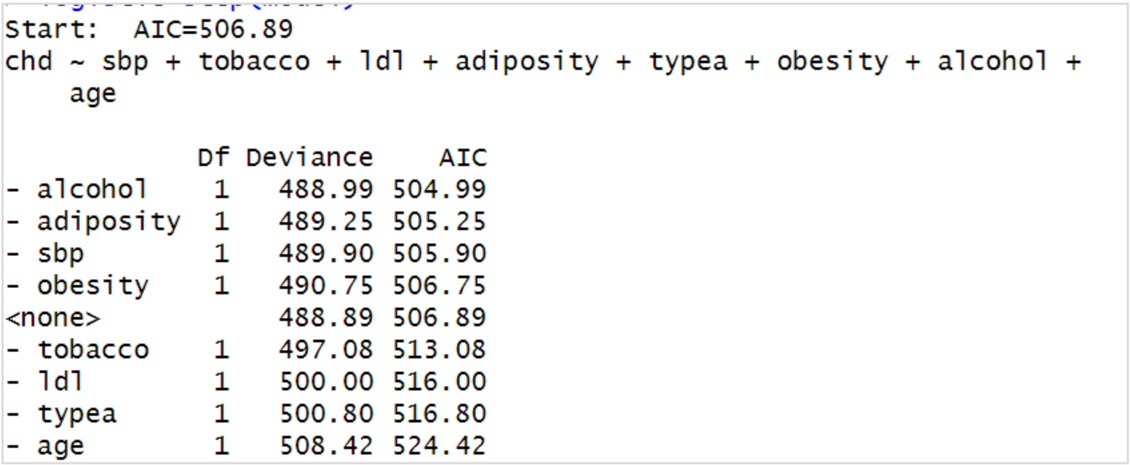


```
> model<-glm(chd~., family = binomial, data=heart_disease_dataset)
```

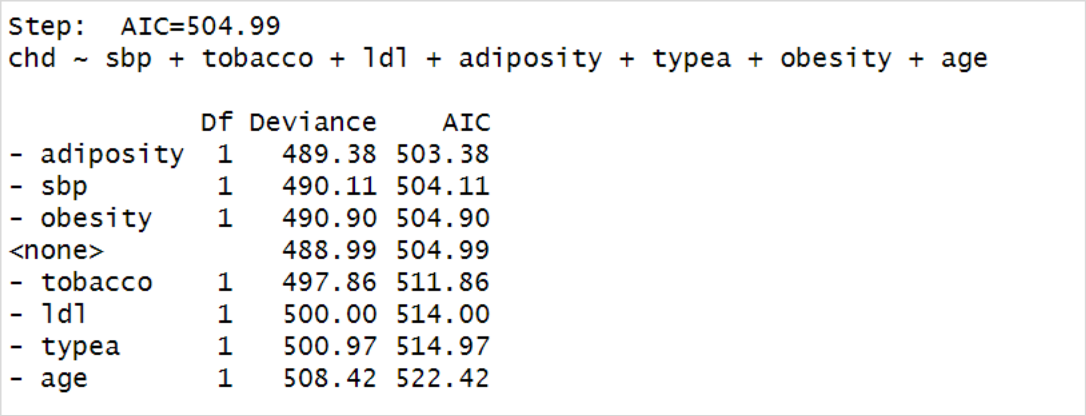


Data	
heart_disease_da…	462 obs. of 9 variables
model	List of 30

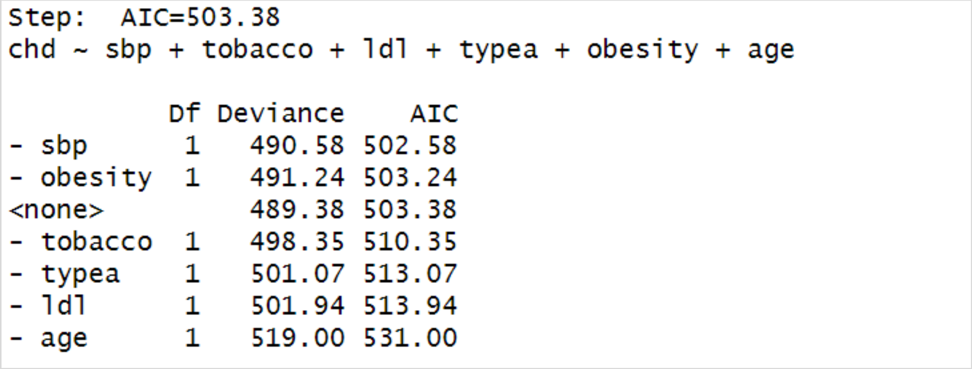
```
Logistic = step(model)
```



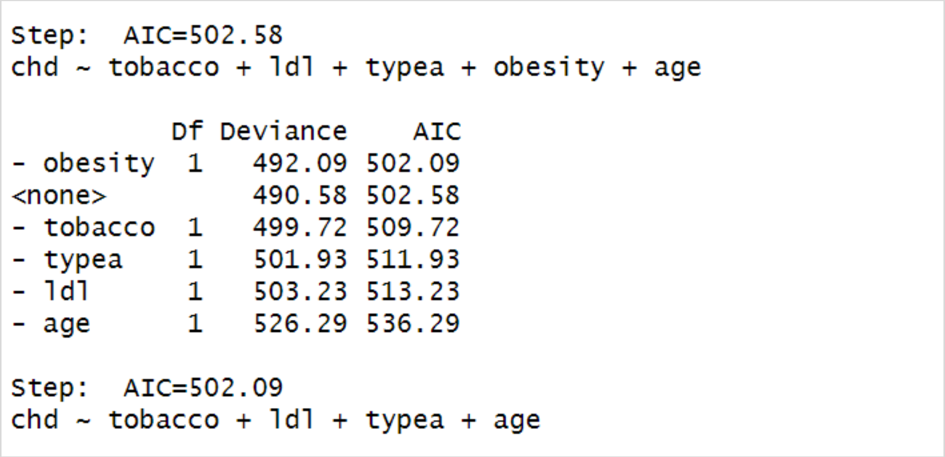
```
Start:  AIC=506.89
chd ~ sbp + tobacco + ldl + adiposity + typea + obesity + alcohol + 
    age

            Df Deviance    AIC
- alcohol    1   488.99 504.99
- adiposity  1   489.25 505.25
- sbp        1   489.90 505.90
- obesity    1   490.75 506.75
<none>           488.89 506.89
- tobacco    1   497.08 513.08
- ldl        1   500.00 516.00
- typea      1   500.80 516.80
- age        1   508.42 524.42
```



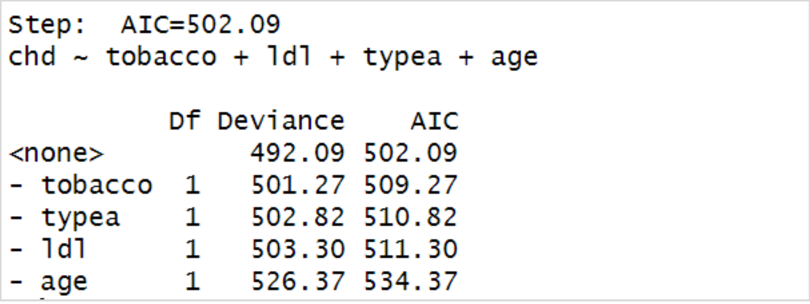
```
Step:  AIC=504.99
chd ~ sbp + tobacco + ldl + adiposity + typea + obesity + age

            Df Deviance    AIC
- adiposity  1   489.38 503.38
- sbp        1   490.11 504.11
- obesity    1   490.90 504.90
<none>           488.99 504.99
- tobacco    1   497.86 511.86
- ldl        1   500.00 514.00
- typea      1   500.97 514.97
- age        1   508.42 522.42
```



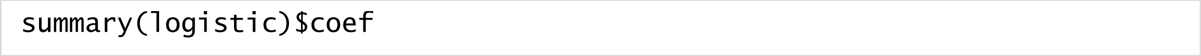
```
Step:  AIC=503.38
chd ~ sbp + tobacco + ldl + typea + obesity + age

          Df Deviance    AIC
- sbp      1   490.58 502.58
- obesity  1   491.24 503.24
<none>         489.38 503.38
- tobacco  1   498.35 510.35
- typea    1   501.07 513.07
- ldl      1   501.94 513.94
- age      1   519.00 531.00
```



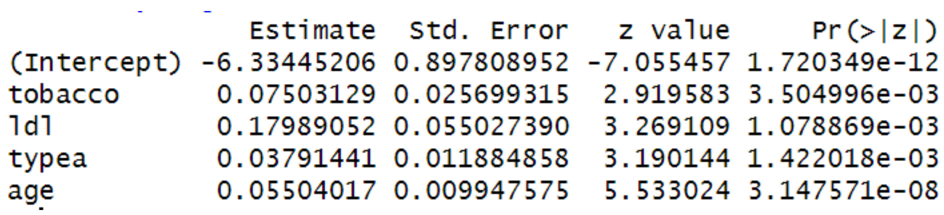
```
Step:  AIC=502.58
chd ~ tobacco + ldl + typea + obesity + age

          Df Deviance    AIC
- obesity  1   492.09 502.09
<none>         490.58 502.58
- tobacco  1   499.72 509.72
- typea    1   501.93 511.93
- ldl      1   503.23 513.23
- age      1   526.29 536.29

Step:  AIC=502.09
chd ~ tobacco + ldl + typea + age
```



```
Step:  AIC=502.09
chd ~ tobacco + ldl + typea + age

          Df Deviance    AIC
<none>         492.09 502.09
- tobacco  1   501.27 509.27
- typea    1   502.82 510.82
- ldl      1   503.30 511.30
- age      1   526.37 534.37
```

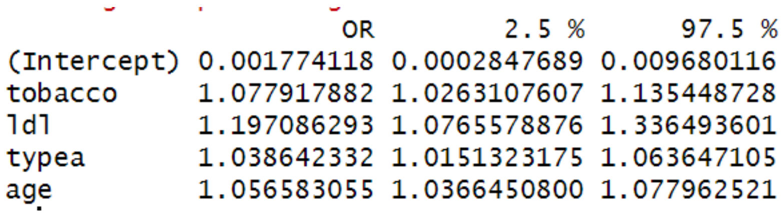


```
summary(logistic)$coef
```



```
               Estimate  Std. Error   z value     Pr(>|z|)
(Intercept) -6.33445206 0.897808952 -7.055457 1.720349e-12
tobacco      0.07503129 0.025699315  2.919583 3.504996e-03
ldl          0.17989052 0.055027390  3.269109 1.078869e-03
typea        0.03791441 0.011884858  3.190144 1.422018e-03
age          0.05504017 0.009947575  5.533024 3.147571e-08
```

```
exp(cbind(OR=coef(logistic),confint(logistic)))
```



```
                     OR        2.5 %      97.5 %
(Intercept) 0.001774118 0.0002847689 0.009680116
tobacco     1.077917882 1.0263107607 1.135448728
ldl         1.197086293 1.0765578876 1.336493601
typea       1.038642332 1.0151323175 1.063647105
age         1.056583055 1.0366450800 1.077962521
```

Comment: As a consequence, every step needed to create an optimal model is demonstrated. First, the model with 10 variables (AIC=506.69) is used. Second, we exclude one variable from the original model, resulting in AIC = 504.99 and subsequent values. And after doing five steps, we come to the model with the four variables that has the lowest AIC value.

Finally, we create a logistic regression model for four variables: age, type A, ldl, and cigarette use.

This model's fitting resembles fitting a straightforward linear regression pretty closely. We use glm() in place of lm(). The usage of family = "binomial" to denote a two-class categorical answer is the only other variation. The standard linear regression would be carried out by calling glm() with family = "gaussian". As a result, the ideal logistic regression model has the following structure:

$$\ln \langle\frac{p}{1-p}\rangle = \beta_0 + \beta_1 . heart\_disease + \beta_2 . hyper\_tension + \beta_3 . avg\_glucose\_level + \beta_4 . age + \varepsilon$$

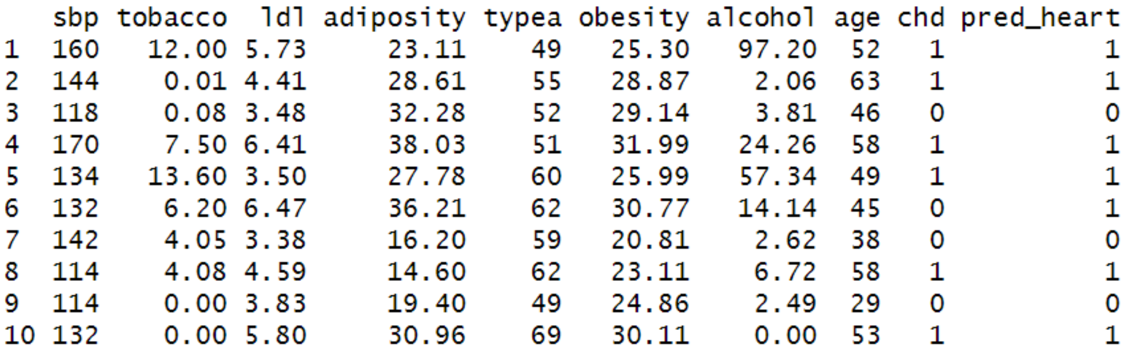
Analyze how different factors affect CHD:

When the estimated coefficient of the variable tobacco has a positive value, it suggests that smoking more is linked to an increased risk of developing coronary heart disease.

Furthermore, the age regression coefficient is 0.05504017. This demonstrates that when age rises by one unit, the risk of coronary heart disease rises by an exponential factor of exp(0.05504017) = 1.056583055 times

6. Forecasting:

```
pred<-predict(logistic,heart_data,type="response")

heart_data$pred_heart<-round(pred,digits=0)

head(heart_data,10)
```



```
   sbp tobacco  ldl adiposity typea obesity alcohol age chd pred_heart
1  160   12.00 5.73     23.11    49   25.30   97.20  52   1          1
2  144    0.01 4.41     28.61    55   28.87    2.06  63   1          1
3  118    0.08 3.48     32.28    52   29.14    3.81  46   0          0
4  170    7.50 6.41     38.03    51   31.99   24.26  58   1          1
5  134   13.60 3.50     27.78    60   25.99   57.34  49   1          1
6  132    6.20 6.47     36.21    62   30.77   14.14  45   0          1
7  142    4.05 3.38     16.20    59   20.81    2.62  38   0          0
8  114    4.08 4.59     14.60    62   23.11    6.72  58   1          1
9  114    0.00 3.83     19.40    49   24.86    2.49  29   0          0
10 132    0.00 5.80     30.96    69   30.11    0.00  53   1          1
```

5 Comment:Looking at the findings, we can see that in both observation and prediction, the proportion of persons having CHD syndrome is about the same. This demonstrates that the logistic regression approach, which we used to predict who was likely to develop CHD, was successful and absolutely viable, further demonstrating the usefulness of statistics in regular life.

R codes:

```
library(tidyverse)
library(rsample)
library(recipes)
library(parsnip)
library(yardstick)

heart_data <- read.csv(file = "CHDdata.csv", header = FALSE)
colnames(heart_data) <-
 c("sbp", "tobacco", "ldl", "adiposity", "famhist", "typea", "obesity", "alcohol", "age", “chd”)
Str(heart_data)
heart_data [heart_data =="N/A"]<-NA
apply(is.na(heart_data),2,sum)
heart_data <-na.omit(heart_data)
apply(is.na(heart_data),2,which)
is.numeric(heart_data$sbp)

is.numeric(heart_data$tobacco)

is.numeric(heart_data$ldl)

is.numeric(heart_data$adiposity)

is.numeric(heart_data$famhist)
heart_data <- heart_data [, -5]

is.numeric(heart_data$typea)

is.numeric(heart_data$obesity)

is.numeric(heart_data$alcohol)

is.numeric(heart_data$age)

is.numeric(heart_data$chd)
> mean<-apply(heart_data[,c(1 : 9)],2,mean)

> median<-apply(heart_data[,c(1 : 9)],2,median)

> max<-apply(heart_data[,c(1 : 9)],2,max)

> min<-apply(heart_data[,c(1 : 9)],2,min)

> sd<-apply(heart_data[,c(1 : 9)],2,sd)

> t(data.frame(mean,sd,median,max,min))

hist(heart_data$age,xlab="Age",ylab="Number of people",main="Histogram of researched age group",ylim=c(0,100),labels=T,col="pink")
hist(heart_data$alcohol,xlab="alcohol",ylab="Amount of alcohol",main="Histogram of researched alcohol group",ylim=c(0,400),labels=T,col="pink")
hist(heart_data$alcohol,xlab="alcohol",ylab="Amount of alcohol",main="Histogram of researched alcohol group",ylim=c(0,400),labels=T,col="pink")

hist(heart_data$typea,xlab="typea",ylab="number of typea",main="Histogram of researched typea group",ylim=c(0,100),labels=T,col="pink")
hist(heart_data$tobacco,xlab="tobacco",ylab="number of tobacco",main="Histogram of researched tobacco group",ylim=c(0,400),labels=T,col="pink")
hist(heart_data$adiposity,xlab="adiposity",ylab="number of adiposity",main="Histogram of researched adiposity group",ylim=c(0,150),labels=T,col="pink")
hist(heart_data$ldl,xlab="ldl",ylab="number of ldl",main="Histogram of researched ldl group",ylim=c(0,200),labels=T,col="pink")
hist(heart_data$sbp,xlab="sbp",ylab="number of sbp",main="Histogram of researched sbp group",ylim=c(0,200),labels=T,col="pink")
barplot(table(heart_data$famhist),xlab="Famhist",ylab="Frequency",main="Barplot of Famhist",col="salmon",ylim=c(0,300),xlim=c(0,12))
Famhist_P<-heart_data[heart_data$famhist=="Present",]
table(Famhist_P$chd)
quantity<-c(96,95)
presentchd<-c("No","Yes")
percentage<-round(quantity/sum(quantity)*100,2)
presentchd<-paste(presentchd,percentage)
presentchd<-paste(presentchd,"%")
Famhist_A<-heart_data[heart_data$famhist=="Absent",]
table(Famhist_A$chd)
Quantity1<-c(206,64)
absentchd<-c("No","Yes")
percentage<-round(quantity/sum(quantity)*100,2)
absentchd<-paste(absentchd,percentage)
absentchd<-paste(absentchd,"%")
par(mfrow=c(1,2))
pie(quantity,labels = presentchd,main="Famhist present chd percentage",col=c("pink","seagreen"))
pie(quantity1,labels = absentchd,main="Famhist absent chd percentage",col=c("pink","lightgreen"))
library(ggplot2)
> library(plyr)
> mu_sbp<-ddply(heart_data,"chd",summarise,grp.mean=mean(sbp))
> ggplot(heart_data,aes(x=sbp,color=as.factor(chd),fill=as.factor(chd)))+
+     geom_histogram(position="identity",alpha=0.5)+
+     geom_vline(data=mu_sbp,aes(xintercept=grp.mean,color=as.factor(chd)),linetype="dashed")+
+     scale_color_manual(values=c("999999","#E69F00","#56B4E9"))+
+     scale_fill_manual(values=c("999999","#E69F00","#56B4E9"))+
+     labs(title="Histogram of sbp for CHDDisease",x="sbp",y="Frequency")+
+     theme_classic()
library(ggplot2)
>   library(plyr)
>   mu_tobacco<-ddply(heart_data,"chd",summarise,grp.mean=mean(tobacco))
>  ggplot(heart_data,aes(x=tobacco,color=as.factor(chd),fill=as.factor(chd)))+
+                geom_histogram(position="identity",alpha=0.5)+
+           geom_vline(data=mu_tobacco,aes(xintercept=grp.mean,color=as.factor(chd)),linetype="dashed")+
+        scale_color_manual(values=c("999999","#E69F00","#56B4E9"))+
+          scale_fill_manual(values=c("999999","#E69F00","#56B4E9"))+
+          labs(title="Histogram of tobacco for CHDDisease",x="tobacco",y="Frequency")+
+           theme_classic()
library(ggplot2)
>  library(plyr)
>   mu_adiposity<-ddply(heart_data,"chd",summarise,grp.mean=mean(adiposity))
>   ggplot(heart_data,aes(x=ldl,color=as.factor(chd),fill=as.factor(chd)))+
+                        geom_histogram(position="identity",alpha=0.5)+
+                     geom_vline(data=mu_adiposity,aes(xintercept=grp.mean,color=as.factor(chd)),linetype="dashed")+
+                 scale_color_manual(values=c("999999","#E69F00","#56B4E9"))+

+                   scale_fill_manual(values=c("999999","#E69F00","#56B4E9"))+
+                   labs(title="Histogram of adiposity for CHDDisease",x="adiposity",y="Frequency")+
+                    theme_classic()
library(ggplot2)
>  library(plyr)
>    mu_typea<-ddply(heart_data,"chd",summarise,grp.mean=mean(typea))
>    ggplot(heart_data,aes(x=ldl,color=as.factor(chd),fill=as.factor(chd)))+

+                          geom_histogram(position="identity",alpha=0.5)+
+                         geom_vline(data=mu_typea,aes(xintercept=grp.mean,color=as.factor(chd)),linetype="dashed")+
+                      scale_color_manual(values=c("999999","#E69F00","#56B4E9"))+
+                        scale_fill_manual(values=c("999999","#E69F00","#56B4E9"))+
+                        labs(title="Histogram of typea for CHDDisease",x="typea",y="Frequency")+
+                          theme_classic()
library(ggplot2)
>  library(plyr)
>      mu_obesity<-ddply(heart_data,"chd",summarise,grp.mean=mean(obesity))
>      ggplot(heart_data,aes(x=ldl,color=as.factor(chd),fill=as.factor(chd)))+
+                                  geom_histogram(position="identity",alpha=0.5)+
+                                  geom_vline(data=mu_obesity,aes(xintercept=grp.mean,color=as.factor(chd)),linetype="dashed")+
+                              scale_color_manual(values=c("999999","#E69F00","#56B4E9"))+
+                                scale_fill_manual(values=c("999999","#E69F00","#56B4E9"))+
+                             labs(title="Histogram of obesity for CHDDisease",x="obesity",y="Frequency")+
+                                 theme_classic()

library(ggplot2)
library(plyr)
>       mu_alcohol<-ddply(heart_data,"chd",summarise,grp.mean=mean(alcohol))
>       ggplot(heart_data,aes(x=ldl,color=as.factor(chd),fill=as.factor(chd)))+
+                                       geom_histogram(position="identity",alpha=0.5)+
+                                       geom_vline(data=mu_alcohol,aes(xintercept=grp.mean,color=as.factor(chd)),linetype="dashed")+
+                                   scale_color_manual(values=c("999999","#E69F00","#56B4E9"))+
+ scale_fill_manual(values=c("999999","#E69F00","#56B4E9"))+
+ labs(title="Histogram of alcohol for CHD Disease",x="alcohol",y="Frequency")+
+ theme_classic()
library(ggplot2)
> library(plyr)
> mu_age<-ddply(heart_data,"chd",summarise,grp.mean=mean(age))
> ggplot(heart_data,aes(x=ldl,color=as.factor(chd),fill=as.factor(chd)))+
+ geom_histogram(position="identity",alpha=0.5)+
+ geom_vline(data=mu_age,aes(xintercept=grp.mean,color=as.factor(chd)),linetype="dashed")+
+ scale_color_manual(values=c("999999","#E69F00","#56B4E9"))+
+ scale_fill_manual(values=c("999999","#E69F00","#56B4E9"))+
+ labs(title="Histogram of age for CHDDisease",x="age",y="Frequency")+
+ theme_classic()
> model<-glm(chd~., family = binomial, data=heart_disease_dataset)
Logistic = step(model)
pred<-predict(logistic,heart_data,type="response")

heart_data$pred_heart<-round(pred,digits=0)

head(heart_data,10)
```

References

1. Nguyễn Tiến Dũng (chủ biên), Nguyễn Đình Huy, (2019), *Xác suất - Thống kê & Phân tích số liệu*

2. J Jambers - D.Hand - W.Hardle, *Introductory Statistic with R*

3. Ph.D Nguyen Van Hanh, *Binomal logistic regression*, website: https://nghiencuugiaoduc.com.vn/hoi-quy-logistic-nhi-thuc-binomial-logisticregression/