Heart disease analysis

Sanjana Gowda

2023-07-15

### Introduction

This dataset contains detailed information on the risk factors for cardiovascular disease. It includes information on age, gender, height, weight, blood pressure values, cholesterol levels, glucose levels, smoking habits and alcohol consumption of over 70 thousand individuals. Additionally it outlines if the person is active or not and if he or she has any cardiovascular diseases.

Source: <https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-diseas>

The dataset consists of:

* Numeric values for age, height, weight, systolic blood pressure (ap\_hi) and diastolic blood pressure (ap\_lo).
* Binary values for gender, alcohol consumption (alco), smoking habits (smoke), active person (active), cardiovascular diseases (cardio). In the gender variable it’s not defined which value correspond to which gender. In the rest of the cases 0 = No and 1 = Yes.
* Levels in the cholesterol and glucose (gluc) variables. In this cases I will consider that 1 = normal, 2 = above normal, 3 = well above normal.

### Hypothesis to be tested

* The lifestyle of the people impact on the risk of having cardiovascular diseases.
* There isn’t a correlation between gender and having cardiovascular diseases
* High levels of glucose and cholesterol contributes to developing cardiovascular diseases.
* Older people have more risk to have cardiovascular diseases.
* People with higher body mass index have more risks to develop cardiovascular diseases.
* People with higher blood pressure have more risks to develop cardiovascular diseases.

## Analysis

### Loading libraries and dataset

heart\_data <- read.csv("C:/Users/Sanjana/Downloads/heart\_data.csv")  
  
library(tidyverse)  
library(skimr)  
library(ggpmisc)  
library(cowplot)  
library(caret)  
library(stringr)  
library(dplyr)

### Preview of the dataset

head(heart\_data)

## index id age gender height weight ap\_hi ap\_lo cholesterol gluc smoke alco  
## 1 0 0 18393 2 168 62 110 80 1 1 0 0  
## 2 1 1 20228 1 156 85 140 90 3 1 0 0  
## 3 2 2 18857 1 165 64 130 70 3 1 0 0  
## 4 3 3 17623 2 169 82 150 100 1 1 0 0  
## 5 4 4 17474 1 156 56 100 60 1 1 0 0  
## 6 5 8 21914 1 151 67 120 80 2 2 0 0  
## active cardio  
## 1 1 0  
## 2 1 1  
## 3 0 1  
## 4 1 1  
## 5 0 0  
## 6 0 0

### Complete summary of the dataset

skim\_without\_charts(heart\_data)

Data summary

|  |  |
| --- | --- |
| Name | heart\_data |
| Number of rows | 70000 |
| Number of columns | 14 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 14 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| index | 0 | 1 | 34999.50 | 20207.40 | 0 | 17499.75 | 34999.5 | 52499.25 | 69999 |
| id | 0 | 1 | 49972.42 | 28851.30 | 0 | 25006.75 | 50001.5 | 74889.25 | 99999 |
| age | 0 | 1 | 19468.87 | 2467.25 | 10798 | 17664.00 | 19703.0 | 21327.00 | 23713 |
| gender | 0 | 1 | 1.35 | 0.48 | 1 | 1.00 | 1.0 | 2.00 | 2 |
| height | 0 | 1 | 164.36 | 8.21 | 55 | 159.00 | 165.0 | 170.00 | 250 |
| weight | 0 | 1 | 74.21 | 14.40 | 10 | 65.00 | 72.0 | 82.00 | 200 |
| ap\_hi | 0 | 1 | 128.82 | 154.01 | -150 | 120.00 | 120.0 | 140.00 | 16020 |
| ap\_lo | 0 | 1 | 96.63 | 188.47 | -70 | 80.00 | 80.0 | 90.00 | 11000 |
| cholesterol | 0 | 1 | 1.37 | 0.68 | 1 | 1.00 | 1.0 | 2.00 | 3 |
| gluc | 0 | 1 | 1.23 | 0.57 | 1 | 1.00 | 1.0 | 1.00 | 3 |
| smoke | 0 | 1 | 0.09 | 0.28 | 0 | 0.00 | 0.0 | 0.00 | 1 |
| alco | 0 | 1 | 0.05 | 0.23 | 0 | 0.00 | 0.0 | 0.00 | 1 |
| active | 0 | 1 | 0.80 | 0.40 | 0 | 1.00 | 1.0 | 1.00 | 1 |
| cardio | 0 | 1 | 0.50 | 0.50 | 0 | 0.00 | 0.0 | 1.00 | 1 |

### Looking for duplicates

length(unique(heart\_data$id))

## [1] 70000

There are 14 columns and 70000 rows, 0 NAs, and 0 duplicates in the dataset

It can be seen a lot of inconsistency in the data:

* The ap\_hi and ap\_lo columns have values that are biologically impossible.
* In the weight and height column there are also weird values.

In these measured variables there may be errors when recording them. So when analyzing those variables, i will remove the outliers applying the IQR method (1). In that range, the wrong data will be deleted, and the analyzed data will be large enough to be representative of the entire population.

str(heart\_data)

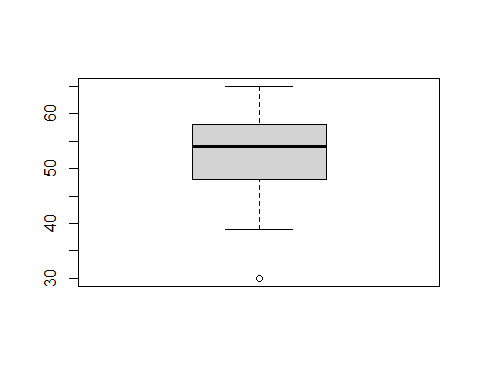
## 'data.frame': 70000 obs. of 14 variables:  
## $ index : int 0 1 2 3 4 5 6 7 8 9 ...  
## $ id : int 0 1 2 3 4 8 9 12 13 14 ...  
## $ age : int 18393 20228 18857 17623 17474 21914 22113 22584 17668 19834 ...  
## $ gender : int 2 1 1 2 1 1 1 2 1 1 ...  
## $ height : int 168 156 165 169 156 151 157 178 158 164 ...  
## $ 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: int 1 3 3 1 1 2 3 3 1 1 ...  
## $ gluc : int 1 1 1 1 1 2 1 3 1 1 ...  
## $ smoke : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ alco : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ active : int 1 1 0 1 0 0 1 1 1 0 ...  
## $ cardio : int 0 1 1 1 0 0 0 1 0 0 ...

### Data processing

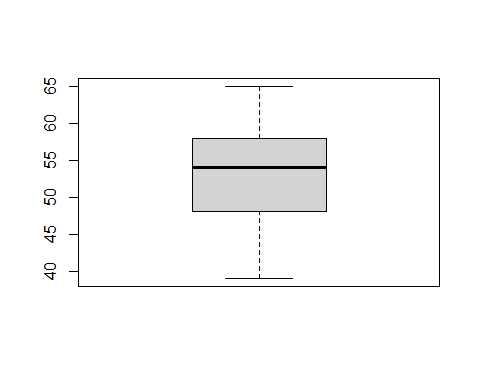
I’ve made some processing in the data:

1. Turn age from days to years.
2. Remove outliers from the age variable since there are only four values corresponding to the age of 30, the rest are in the range of 39-65 years old.
3. Change 0, 1 code in smoke, alco, active, and cardio. E.g “Smoker”, “No smoker”.
4. Change cholesterol, gluc and gender columns to factor.
5. Calculate body mass index (BMI). Formula = weight(kg)/height(m)^2). And classificate those bmi values according to the World Health Organization classification (2):
   * <18.5 underweight (uw)
   * 18.5-24.9 normal weight (normal)
   * 25.0 - 29.9 pre-obesity (pre-ob)
   * 30.0 - 34.9 obesity class 1 (ob 1)
   * 35.0 - 39.9 obesity class 2 (ob 2)
   * ≥ 40 obesity class 3 (ob 3)
6. Calculate mean arterial pressure (MAP). Formula = (ap\_hi+ap\_lo\*2)/3. And classificate those map values according to the American Hearth Association (2020) (3):
   * <90 Normal (normal)
   * 90 to 91.99 Elevated blood pressure (high-bp)
   * 92 to 95.99 Hypertension stage 1 (hyp1)
   * ≥ 96 Hypertension stage 2 (hyp2)
7. Selecting desired columns

#Converting the units from the age column from days to years  
heart\_data$age\_years <- round(heart\_data$age/365)  
#Analyzing the range of ages from the participants  
box<-boxplot(heart\_data$age\_years) #Here we can see that exists a clearly outlier in the age of 30. There are only 4 rows with that value, so they have to be removed.



heart\_data<-filter(heart\_data, age\_years != 30)  
  
box2<-boxplot(heart\_data$age\_years)



#All adults from 39 to 65 years, with a mean of 53.34   
  
  
#Changing values in columns to be more descriptive  
heart\_data["cardio"][heart\_data["cardio"] == 0] <- 'No disease'  
heart\_data["cardio"][heart\_data["cardio"] == 1] <- 'Disease'  
heart\_data["smoke"][heart\_data["smoke"] == 0] <- 'No smoker'  
heart\_data["smoke"][heart\_data["smoke"] == 1] <- 'Smoker'  
heart\_data["alco"][heart\_data["alco"] == 0] <- 'No alco'  
heart\_data["alco"][heart\_data["alco"] == 1] <- 'Alco'  
heart\_data["active"][heart\_data["active"] == 0] <- 'No active'  
heart\_data["active"][heart\_data["active"] == 1] <- 'Active'  
  
  
  
  
#Changing some columns from int to factor format  
heart\_data$cholesterol<-as.factor(heart\_data$cholesterol)  
heart\_data$gluc<-as.factor(heart\_data$gluc)  
heart\_data$gender <- as.factor(heart\_data$gender)  
  
  
  
  
#Calculating the body mass index (bmi); Formula = weight(kg)/height(m)^2  
heart\_data$bmi <- heart\_data$weight/(heart\_data$height/100)^2  
  
#Classifying bmi accord to the World Health Organization classification:  
  
heart\_data$bmi\_class <- cut(heart\_data$bmi,   
 breaks = c(-Inf, 18.49, 24.9, 29.9,  
 34.9, 39.9, Inf),  
 labels = c("uw", "normal", "pre-ob",   
 "ob 1", "ob 2", "ob 3"))  
  
  
#Calculating the mean arterial pressure (map) ; Formula = (ap\_hi+ap\_lo\*2)/3  
heart\_data$map <- (heart\_data$ap\_hi+heart\_data$ap\_lo\*2)/3  
  
#Classificating map according to the American Hearth Association (2020)  
  
  
heart\_data$map\_class<- cut(heart\_data$map,   
 breaks = c(-Inf, 89.99, 91.99, 95.99, Inf),  
 labels = c("normal", "high-bp", "hyp1", "hyp2"))  
  
  
#Keeping a dataset deleting unnecesary columns  
heart\_data<-select(heart\_data, age\_years, gender, bmi, bmi\_class, map,  
 map\_class,gluc, cholesterol, smoke, alco, active, cardio)  
  
# Capitalizing the first letter of each column name  
heart\_data <- heart\_data %>%  
 rename\_with(~str\_to\_title(.))

And this are the first ten rows of the dataset I will work with:

head(heart\_data)

## Age\_years Gender Bmi Bmi\_class Map Map\_class Gluc Cholesterol  
## 1 50 2 21.96712 normal 90.00000 high-bp 1 1  
## 2 55 1 34.92768 ob 2 106.66667 hyp2 1 3  
## 3 52 1 23.50781 normal 90.00000 high-bp 1 3  
## 4 48 2 28.71048 pre-ob 116.66667 hyp2 1 1  
## 5 48 1 23.01118 normal 73.33333 normal 1 1  
## 6 60 1 29.38468 pre-ob 93.33333 hyp1 2 2  
## Smoke Alco Active Cardio  
## 1 No smoker No alco Active No disease  
## 2 No smoker No alco Active Disease  
## 3 No smoker No alco No active Disease  
## 4 No smoker No alco Active Disease  
## 5 No smoker No alco No active No disease  
## 6 No smoker No alco No active No disease

# Identify integer variables  
integer\_vars <- sapply(heart\_data, is.integer)  
  
# Convert integer variables to numeric  
heart\_data[integer\_vars] <- lapply(heart\_data[integer\_vars], as.numeric)  
  
str(heart\_data)

## 'data.frame': 69996 obs. of 12 variables:  
## $ Age\_years : num 50 55 52 48 48 60 61 62 48 54 ...  
## $ Gender : Factor w/ 2 levels "1","2": 2 1 1 2 1 1 1 2 1 1 ...  
## $ Bmi : num 22 34.9 23.5 28.7 23 ...  
## $ Bmi\_class : Factor w/ 6 levels "uw","normal",..: 2 5 2 3 2 3 5 4 3 3 ...  
## $ Map : num 90 106.7 90 116.7 73.3 ...  
## $ Map\_class : Factor w/ 4 levels "normal","high-bp",..: 2 4 2 4 1 3 4 4 1 1 ...  
## $ Gluc : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 2 1 3 1 1 ...  
## $ Cholesterol: Factor w/ 3 levels "1","2","3": 1 3 3 1 1 2 3 3 1 1 ...  
## $ Smoke : chr "No smoker" "No smoker" "No smoker" "No smoker" ...  
## $ Alco : chr "No alco" "No alco" "No alco" "No alco" ...  
## $ Active : chr "Active" "Active" "No active" "Active" ...  
## $ Cardio : chr "No disease" "Disease" "Disease" "Disease" ...

So it will be analyzed the risk factors for developing cardiovascular diseases in adult people from 39 to 65 years old.

## Analysis

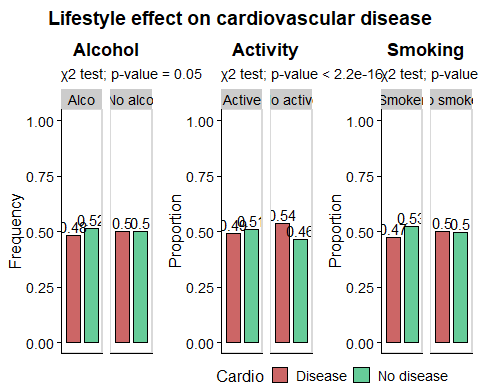
### Lifestyle analisis: smoking, drinking and activity.

First I’m gonna analyze the relationship of lifestyles (smoking, drinking alcohol and being active) and the development of cardiovascular diseases.

* In each case a Chisquare test was made to determine if there are statistical differences between conditions (e.g smoking or not smoking).
* In all cases H0 = no differences between conditions; H1 = differences between conditions.

#Function to make a relative frequencies table  
freq.table <- function (x, y, z){substitute(x %>%  
 group\_by(y, z) %>%  
 summarise(n = n ()) %>%  
 mutate(freq = n / sum (n)))%>%  
 eval}

#Alcohol drinking and cardiovascular disease  
  
alco <- freq.table(heart\_data, Alco, Cardio)  
  
alcochi<-chisq.test(heart\_data$Cardio, heart\_data$Alco, correct=FALSE)  
  
alco.plot<- ggplot(data = alco, aes(x = Cardio, y = freq, fill = Cardio)) +  
 geom\_bar(stat='identity',   
 colour = 'black',  
 width=.75)+   
 geom\_text(aes(label=round(freq,2)),   
 position=position\_dodge(width=0.9),   
 vjust=-0.25)+  
 ylim (0, 1)+  
 facet\_grid(~Alco)+ #Split plot by lifestyle  
 scale\_fill\_manual(values=c( "#CC6666", "#66CC99"))+  
 theme\_half\_open(12)+  
 panel\_border()+  
 ggtitle("Alcohol")+ #Add title  
 theme(plot.title = element\_text(hjust = 0.5), #Center the title  
 axis.title.x=element\_blank(), #Remove X label  
 axis.text.x=element\_blank(),  
 axis.ticks.x=element\_blank())+  
 labs(y = "Frequency",  
 subtitle = "χ2 test; p-value = 0.05") #subtitle with p-value  
  
  
  
  
#Activity of the people and cardiovascular disease  
  
active<-freq.table(heart\_data, Active, Cardio)  
  
activechi<-chisq.test(heart\_data$Cardio, heart\_data$Active, correct=FALSE)  
  
active.plot<-ggplot(data = active, aes(x = Cardio, y = freq, fill = Cardio)) +  
 geom\_bar(stat='identity', colour = 'black', width=0.75)+   
 geom\_text(aes(label=round(freq,2)), position=position\_dodge(width=0.9), vjust=-0.25)+  
 ylim (0, 1)+  
 facet\_grid(~Active)+ #Split plot by lifestyle  
 scale\_fill\_manual(values=c("#CC6666", "#66CC99"))+  
 theme\_half\_open(12)+  
 panel\_border()+  
 ggtitle("Activity")+ #Add title  
 theme(plot.title = element\_text(hjust = 0.5), #Center the title  
 axis.title.x=element\_blank(), #Remove X label  
 axis.text.x=element\_blank(),  
 axis.ticks.x=element\_blank())+   
 labs(y = "Proportion",  
 subtitle = "χ2 test; p-value < 2.2e-16") #subtitle with p-value  
  
  
  
#Now if smokes contributes to the cardiovascular disease  
smoke <- freq.table(heart\_data, Smoke, Cardio)  
  
smokechi<-chisq.test(heart\_data$Cardio, heart\_data$Smoke, correct=FALSE)  
  
smoke.plot <- ggplot(data = smoke, aes(x = Cardio, y = freq, fill = Cardio)) +  
 geom\_bar(stat='identity', colour = 'black', width=0.75)+   
 geom\_text(aes(label=round(freq,2)), position=position\_dodge(width=0.9), vjust=-0.25)+  
 ylim (0, 1)+  
 facet\_grid(~factor(Smoke, levels=c('Smoker', 'No smoker')))+ #Split plot by lifestyle  
 scale\_fill\_manual(values=c("#CC6666", "#66CC99"))+  
 theme\_half\_open(12)+  
 panel\_border()+  
 ggtitle("Smoking")+ #Add title  
 theme(plot.title = element\_text(hjust = 0.5), #Center the title  
 axis.title.x=element\_blank(), #Remove X label  
 axis.text.x=element\_blank(),  
 axis.ticks.x=element\_blank())+   
 labs( y = "Proportion",  
 subtitle = "χ2 test; p-value = 4.1e-05")#subtitle with p-value  
  
  
#putting all three plots in one  
  
  
lifestyle.plot<-plot\_grid(alco.plot+ theme(legend.position="none"),  
 active.plot+ theme(legend.position="none"),  
 smoke.plot+theme(legend.position="none"),  
 ncol=3)  
  
title <- ggdraw() + draw\_label("Lifestyle effect on cardiovascular disease",  
 fontface='bold') #Creating title  
  
lifestyle.plot<- plot\_grid(title,  
 lifestyle.plot,  
 ncol = 1,  
 rel\_heights=c(0.1, 1))  
  
  
legend <- get\_legend(smoke.plot +theme(legend.position = c(.45,.5),  
 legend.direction="horizontal")) #Creating legend  
  
  
lifestyle.plot<-plot\_grid(lifestyle.plot,  
 legend,  
 ncol=1,  
 rel\_heights = c(1, .05))  
  
lifestyle.plot



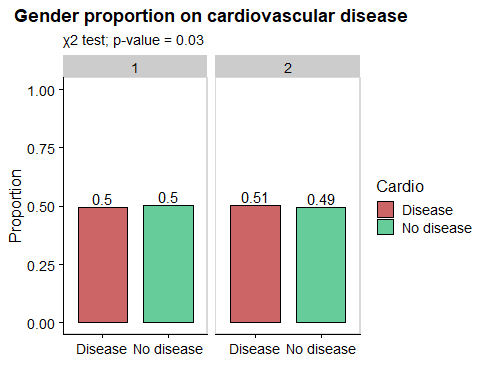
From the previous graphs it can be made some deductions:

* The lifestyle of the adult people between 39 and 65 years, contributes to developing or not a cardiovascular disease.
* Being an active person it’s fundamental to minimize the risk of developing a cardiovascular disease.
* In the people who drink alcohol, there aren’t significant differences.
* Analyzing the smoking people, statically speaking, there are significant differences (p-value << 0.05), and in the group who smoke there are less people with cardiovascular disease. But this is erroneous and it’s related to a small sample size. A lot of bibliography says the opposite (4).
* More detail in the information is needed to be more precisely in the impact of lifestyles in the development of a cardiovascular disease, e.g: the amount of alcohol drinking per day, amount of cigarettes per day, etc.

Next I’m gonna analyze if there is a relationship between the personal characteristics of the people (age, gender, cholesterol, gluc, map ,bmi) and cardiovascular diseases

### Analyzing gender

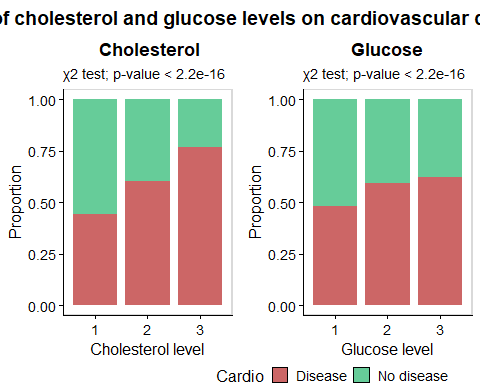
gender <- freq.table(heart\_data, Gender, Cardio)  
  
genderchi<-chisq.test(heart\_data$Gender, heart\_data$Cardio)  
  
gender.plot<-ggplot(gender, aes(x = Cardio, y = freq, fill = Cardio))+  
 geom\_bar(stat='identity', colour = 'black', width=0.75)+   
 ylim (0, 1)+  
 geom\_text(aes(label=round(freq,2)), position=position\_dodge(width=0.9), vjust=-0.25)+  
 facet\_grid(~Gender)+ #dividing by gender  
 scale\_fill\_manual(values=c("#CC6666", "#66CC99"))+   
 theme\_half\_open(12)+  
 panel\_border()+  
 ggtitle("Gender proportion on cardiovascular disease")+ #Add title  
 theme(plot.title = element\_text(hjust = 0.5))+ #Center the title  
 labs(x = "", y = "Proportion",  
 subtitle = "χ2 test; p-value = 0.03") #Subtitle with p-value  
  
  
gender.plot



It seems that there isn’t difference between gender and the develop of cardiovascular disease, the p-value of the chisquare test is near to 0.05 so it cannot reject H0.

### Now analyzing cholesterol and glucose levels

#Cholesterol levels and cardiovascular disease  
  
cholesterol<-freq.table(heart\_data, Cholesterol, Cardio)  
  
  
cholchi<-chisq.test(heart\_data$Cholesterol, heart\_data$Cardio, correct = FALSE)  
  
cholesterol.plot<-ggplot(cholesterol, aes(x = Cholesterol, y = freq, fill = Cardio))+  
 geom\_bar(stat = 'identity', width=0.85, position = position\_fill (reverse = TRUE))+  
 scale\_fill\_manual(values=c("#CC6666", "#66CC99"))+  
 theme\_half\_open(12)+  
 panel\_border()+  
 ggtitle(" Cholesterol")+ #Add title  
 theme(plot.title = element\_text(hjust = 0.5))+ #Center the title  
 labs(x = "Cholesterol level", y = "Proportion",  
 subtitle = "χ2 test; p-value < 2.2e-16")  
  
  
#Checking the same with the gluc  
  
gluc<-freq.table(heart\_data, Gluc, Cardio)  
  
glucchi<-chisq.test(heart\_data$Gluc, heart\_data$Cardio, correct = FALSE)  
  
gluc.plot<-ggplot(gluc, aes(x = Gluc, y = freq, fill = Cardio))+  
 geom\_bar(stat = 'identity', width=0.85, position = position\_fill (reverse = TRUE))+  
 scale\_fill\_manual(values=c("#CC6666", "#66CC99"))+  
 theme\_half\_open(12)+  
 panel\_border()+  
 ggtitle("Glucose")+ #Add title  
 theme(plot.title = element\_text(hjust = 0.5))+ #Center the title  
 labs(x = "Glucose level", y = "Proportion",  
 subtitle = "χ2 test; p-value < 2.2e-16")  
  
  
  
#Putting the two plots together  
  
gluc\_and\_chol<-plot\_grid(cholesterol.plot+ theme(legend.position="none"),  
 gluc.plot+ theme(legend.position="none"),  
 ncol=2,  
 rel\_heights = c(1, .1))  
  
title\_chol <- ggdraw() + draw\_label("Effect of cholesterol and glucose levels on cardiovascular disease", fontface='bold')   
   
gluc\_and\_chol<- plot\_grid(title\_chol,  
 gluc\_and\_chol,  
 ncol = 1,  
 rel\_heights=c(0.1, 1))  
   
gluc\_and\_chol<-plot\_grid(gluc\_and\_chol,  
 legend,  
 ncol=1,  
 rel\_heights = c(1, .05))  
  
gluc\_and\_chol

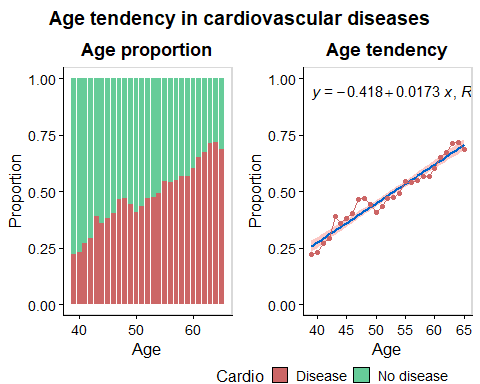


Higher levels of glucose and cholesterol are associated with higher proportion of people with cardiovascular diseases:

* Having “well above normal” levels of cholesterol it’s highly correlated with a high proportion of people with cardiovascular diseases (77%).
* Having “above normal” and “well above normal” levels of glucose have similar proportions of people with cardiovascular diseases (59% and 62% respectively).

### Relationship between age and cardiovascular diseases

age <- freq.table(heart\_data, Age\_years, Cardio)  
age.plot<-ggplot(age, aes(x = Age\_years, y = freq, fill = Cardio))+  
 geom\_bar(stat = 'identity', width=0.85, position = position\_fill (reverse = TRUE))+  
 scale\_fill\_manual(values=c("#CC6666", "#66CC99"))+  
 theme\_half\_open(12)+  
 panel\_border()+  
 ggtitle("Age proportion")+ #Add title  
 theme(plot.title = element\_text(hjust = 0.5))+ #Center the title  
 labs(x = "Age", y = "Proportion" )  
  
  
  
age.1<-subset(age, Cardio == "Disease") #Focusing only in the cases which have cardiovascular disease  
  
corage<-cor(age.1$Age\_years,age.1$freq) #exists a correlation of 0.97  
  
age.tendency.plot<-ggplot(data = age.1, aes(x = Age\_years, y = freq, fill = Cardio)) +  
 stat\_poly\_line(color="#0066CC") +#adding tendency line  
 stat\_poly\_eq(aes(label = paste(after\_stat(eq.label),  
 after\_stat(rr.label),   
 sep = "\*\", \"\*")))+ #adding equation and r2  
 geom\_point(color="#CC6666")+  
 geom\_line(color="#CC6666")+  
 ylim(0,1)+  
 theme\_half\_open(12)+  
 panel\_border()+  
 ggtitle("Age tendency")+ #Add title  
 theme(plot.title = element\_text(hjust = 0.5))+ #Center the title  
 labs(x = "Age", y = "Proportion" )  
  
  
#putting the two plots together  
  
ages.plot<-plot\_grid(age.plot+ theme(legend.position="none"),  
 age.tendency.plot+ theme(legend.position="none"),  
 ncol=2,  
 rel\_heights = c(1, .1))  
  
title\_ages <- ggdraw() + draw\_label("Age tendency in cardiovascular diseases",  
 fontface='bold')   
   
ages.plot<- plot\_grid(title\_ages,  
 ages.plot,  
 ncol = 1,  
 rel\_heights=c(0.1, 1))  
   
ages.plot<-plot\_grid(ages.plot,  
 legend,  
 ncol=1,  
 rel\_heights = c(1, .05))  
  
ages.plot

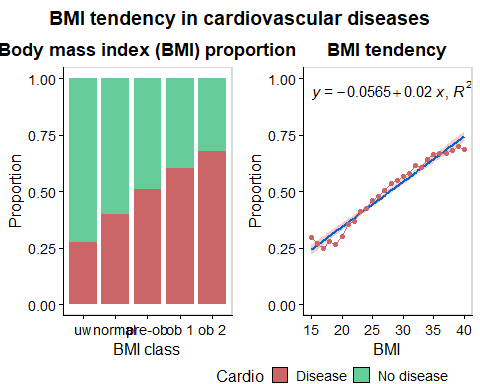


As can be seen in this graphs, there is a positive linear correlation between age and proportion of people with cardiovascular diseases. In older people, the proportion of people with cardiovascular disease is much higher than in the youngest: in the range from 60 to 65 years, the proportion of people with cardiovascular diseases is around 70%, meanwhile in the youngest group from 39 to 45 years old the proportion range of people with cardiovascular diseases is between 22 and 40%.

### Now working with the BMI variable

#Defining function for delete outliers with IQRmethod  
IQRmethod <- function(x,y){quartiles <- quantile(y, probs=c(.25, .75),  
 na.rm = FALSE)  
no\_outlier <- subset(x, y > quartiles[1] - 1.5\*IQR(y) & y < quartiles[2] + 1.5\*IQR(y))  
return(no\_outlier)}

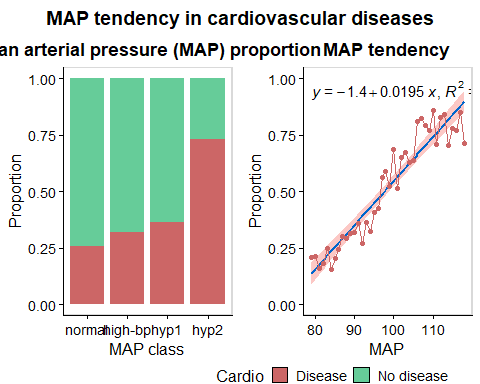
#Removing outliers  
bmi\_cutt <- IQRmethod(heart\_data, heart\_data$Bmi)  
  
#Building freq table without outliers  
bmi<-freq.table(bmi\_cutt, Bmi\_class, Cardio)  
  
  
bmi.plot<-ggplot(bmi, aes(x = Bmi\_class, y = freq, fill = Cardio))+  
 geom\_bar(stat = 'identity', width=0.85, position = position\_fill (reverse = TRUE))+  
 scale\_fill\_manual(values=c("#CC6666", "#66CC99"))+  
 theme\_half\_open(12)+  
 panel\_border()+  
 ggtitle("Body mass index (BMI) proportion")+ #Add title  
 theme(plot.title = element\_text(hjust = 0.5))+ #Center the title  
 labs(x = "BMI class", y = "Proportion" )  
  
#Now determining the correlation between BMI and the proportion of people with cardiovascular disease.  
  
#Obtaining round values  
bmi.round<-freq.table(bmi\_cutt, round(Bmi), Cardio)  
  
bmi.1<-subset(bmi.round, Cardio == "Disease") #Focusing only in the cases which have cardiovascular disease  
  
corbmi<-cor(bmi.1$`round(Bmi)`,bmi.1$freq) #exists a correlation of 0.98  
  
bmi.tendency.plot<-ggplot(data = bmi.1, aes(x = `round(Bmi)`, y = freq, fill = Cardio)) +  
 stat\_poly\_line(color="#0066CC") + #adding tendency line  
 stat\_poly\_eq(aes(label = paste(after\_stat(eq.label), #equation  
 after\_stat(rr.label), #r2  
 sep = "\*\", \"\*")))+  
 geom\_point(color="#CC6666")+  
 geom\_line(color="#CC6666")+  
 ylim(0,1)+  
 theme\_half\_open(12)+  
 panel\_border()+  
 ggtitle("BMI tendency")+ #Add title  
 theme(plot.title = element\_text(hjust = 0.5))+ #Center the title  
 labs(x = "BMI", y = "Proportion" )  
  
  
#Putting plots together  
  
bmi.plot1<-plot\_grid(bmi.plot+ theme(legend.position="none"),  
 bmi.tendency.plot+ theme(legend.position="none"),  
 ncol=2,  
 rel\_heights = c(1, .1))  
  
title\_bmi <- ggdraw() + draw\_label("BMI tendency in cardiovascular diseases", fontface='bold')   
   
bmi.plot1<- plot\_grid(title\_bmi,  
 bmi.plot1,  
 ncol = 1,  
 rel\_heights=c(0.1, 1))  
   
bmi.plot1<-plot\_grid(bmi.plot1,  
 legend,  
 ncol=1,  
 rel\_heights = c(1, .05))  
  
bmi.plot1



Here it can be seen a positive linear correlation. Higher BMI it’s correlated with higher proportion of people with cardiovascular diseases.

### Analyzing MAP

#Doing the same but with the MAP  
map\_cutt<-IQRmethod(heart\_data, heart\_data$Map)  
  
  
map<-freq.table(map\_cutt, Map\_class, Cardio)   
  
map.plot<-ggplot(map, aes(x = Map\_class, y = freq, fill = Cardio))+  
 geom\_bar(stat = 'identity', width=0.85, position = position\_fill (reverse = TRUE))+  
 scale\_fill\_manual(values=c("#CC6666", "#66CC99"))+  
 theme\_half\_open(12)+  
 panel\_border()+  
 ggtitle("Mean arterial pressure (MAP) proportion")+ #Add title  
 theme(plot.title = element\_text(hjust = 0.5))+ #Center the title  
 labs(x = "MAP class", y = "Proportion" )  
  
  
map.round<-freq.table(map\_cutt, round(Map), Cardio)  
  
map.1<-subset(map.round, Cardio == "Disease") #Focusing only in the cases which have cardiovascular disease  
  
cormap<-cor(map.1$`round(Map)`,map.1$freq) #exists a correlation of 0.95  
  
map.tendency.plot<-ggplot(data = map.1, aes(x = `round(Map)`, y = freq, fill = Cardio)) +  
 stat\_poly\_line(color="#0066CC") + #tendency line  
 stat\_poly\_eq(aes(label = paste(after\_stat(eq.label), #equation  
 after\_stat(rr.label), #r2  
 sep = "\*\", \"\*")))+  
 geom\_point(color="#CC6666")+  
 geom\_line(color="#CC6666")+  
 ylim(0,1)+  
 theme\_half\_open(12)+  
 panel\_border()+  
 ggtitle("MAP tendency")+   
 theme(plot.title = element\_text(hjust = 0.5))+   
 labs(x = "MAP", y = "Proportion" )  
  
  
  
map.plot1<-plot\_grid(map.plot+ theme(legend.position="none"),  
 map.tendency.plot+ theme(legend.position="none"),  
 ncol=2,  
 rel\_heights = c(1, .1))  
  
title\_map <- ggdraw() + draw\_label("MAP tendency in cardiovascular diseases",  
 fontface='bold')   
   
map.plot1<- plot\_grid(title\_map,  
 map.plot1,  
 ncol = 1,  
 rel\_heights=c(0.1, 1))  
   
map.plot1<-plot\_grid(map.plot1,  
 legend,  
 ncol=1,  
 rel\_heights = c(1, .05))  
  
map.plot1



Viewing the mean arterial pressure, it also can be seen a positive linear correlation. In the hypertension 2 group is where the proportion of people with cardiovascular diseases is really higher (73%).

So these three measurable variables (age, BMI and MAP) can help to determine the probability of getting a cardiovascular disease.

## Logistic regression model

Now I will see if a logistic regression model with that 3 measurable variables can help to predict the risk of getting cardiovascular disease.

#removing the outliers for both bmi, map  
no.outlier <- IQRmethod(heart\_data, heart\_data$Map)  
no.outlier <- IQRmethod(no.outlier, no.outlier$Bmi)  
  
#Creating boolean column  
no.outlier$Boolean <- no.outlier$Cardio=="Disease"  
  
  
# Split the data into training and test set  
set.seed(123)  
training.samples <- no.outlier$Boolean %>%   
 createDataPartition(p = 0.8, list = FALSE)  
train.data <- no.outlier[training.samples, ]  
test.data <- no.outlier[-training.samples, ]  
  
 #Analyzing the glm model  
glmMod <- glm(Boolean ~ Bmi + Map + Age\_years,  
 data = train.data,   
 family = "binomial")  
  
summary(glmMod) #Here it can be seen that all variables are significant when explaining the probability of getting a heart disease.

##   
## Call:  
## glm(formula = Boolean ~ Bmi + Map + Age\_years, family = "binomial",   
## data = train.data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -14.106442 0.158937 -88.75 <2e-16 \*\*\*  
## Bmi 0.039556 0.002386 16.58 <2e-16 \*\*\*  
## Map 0.102305 0.001389 73.67 <2e-16 \*\*\*  
## Age\_years 0.059249 0.001553 38.16 <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: 68522 on 49460 degrees of freedom  
## Residual deviance: 57938 on 49457 degrees of freedom  
## AIC: 57946  
##   
## Number of Fisher Scoring iterations: 3

The model adjusts really well. All variables are significant.

### Now testing the model accuracy

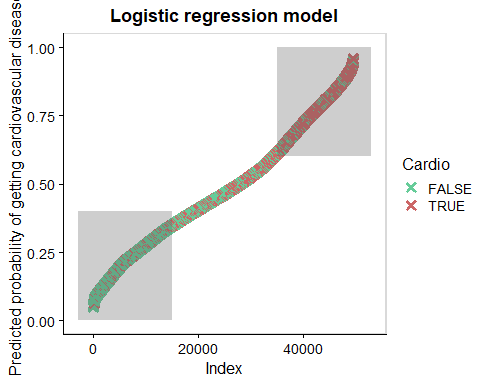
# Make predictions in the test data  
probabilities <- glmMod %>% predict(test.data, type = "response")  
predicted.classes <- ifelse(probabilities > 0.5, "TRUE", "FALSE")  
  
# Model accuracy  
mean(predicted.classes == test.data$Boolean)

## [1] 0.7041411

#Model accuracy ~70%

### Plotting the accuracy of the model

#Plot the accuracy of the model  
  
predicted.data <- data.frame(  
 probability.of.cd = glmMod$fitted.values,  
 Cardio=train.data$Boolean)  
  
predicted.data <- predicted.data[  
 order(predicted.data$probability.of.cd, decreasing=FALSE),]  
  
predicted.data$rank <- 1:nrow(predicted.data)  
  
  
# plot the predicted probabilities for each individual of having heart diseases and color by whether or not they actually had heart disease  
  
predicted.plot<-ggplot(data=predicted.data, aes(x=rank, y=probability.of.cd)) +  
 geom\_point(aes(color=Cardio), alpha=1, shape=4, stroke=2) +  
 scale\_color\_manual(values = c("FALSE" = "#66CC99", "TRUE" = "#CC6666"))+  
 theme\_half\_open(12)+  
 panel\_border()+  
 annotate("rect", xmin = 35000, xmax = 53000, ymin = 0.6, ymax = 1,  
 alpha = .3)+  
 annotate("rect", xmin = -3000, xmax = 15000, ymin = 0, ymax = 0.4,  
 alpha = .3)+  
 ggtitle("Logistic regression model")+ #Add title  
 theme(plot.title = element\_text(hjust = 0.5))+ #Center the title  
 xlab("Index") +  
 ylab("Predicted probability of getting cardiovascular disease")  
  
predicted.plot



In the lowest gray area, most of the people don’t have heart disease, and in the upper gray area most of the people have cardiovascular disease. So here it’s represented the 70% of accuracy of the model

### Final conclusions

From these dataset it can be concluded the next things for adults between 39 and 65 years old:

* Lifestyle: being an active person reduces the probabilities of having cardiovascular diseases.
* Gender: there isn’t difference between genders and risk of having cardiovascular diseases.
* Cholesterol and glucose levels: higher cholesterol and glucose levels are highly correlated with cardiovascular diseases.
* Age: Age and proportion of people with cardiovascular disease are highly correlated. In older people (from 60 to 65 years old) the risk is much higher than in the less-older ones (from 39 to 45 years old).
* BMI: positive correlation between BMI and proportion of people with cardiovascular disease. Obesity class 2 group is the most risky one.
* MAP: poositive correlation between MAP and proportion of people with cardiovascular disease. Hypertension stage 2 group is the most risky one.
* Predictive model: a logistic regression model with the age, BMI and MAP variables was made. This model has a 70% of accuracy when predicting the probability of having cardiovascular diseases.

### References

* 1: <https://online.stat.psu.edu/stat200/lesson/3/3.2>
* 2: <https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations>
* 3: <https://www.ahajournals.org/doi/10.1161/HYPERTENSIONAHA.120.14929>
* 4: <https://www.hopkinsmedicine.org/health/conditions-and-diseases/smoking-and-cardiovascular-disease>
* 5: <https://rpubs.com/GehadGad/854190>