Coronary Heart Disease Prediction Project

Alan Bauza

7/23/2021

1. CHD PROJECT INTRODUCTION

The data set that will be used is from the Framingham heart study, an ongoing study since 1948 in the American city of Framingham that has the objective of studying cardiovascular health.

In this case the data set will be used to predict if a person can have an event of coronary heart disease (CHD) in a 10-year period considering different factors that will be explained in this introduction section.

The data set has 4238 rows, and 16 columns.

Each column represents the following:

-"male", whether the person is a male or not, with value 1 if male and 0 if woman;

-"age", the age of the person;

-"education", the level of education attained by the individual: No high school (value = 1), high school degree (value = 2), some college (value = 3), college degree (value = 4);

-"currentSmoker" whether the person currently smokes, with value 1 if they smoke and 0 if they don't;

-"cigsPerDay", the amount of cigarettes the person smokes per day;

-"BPMeds", whether the person currently takes blood pressure medication, with value 1 if they do and 0 if they don't;

-"prevalentStroke", if the person has had a stroke, with value 1 if they had and 0 if they had not;

-"prevalentHyp", whether the person has hypertension, with value 1 if they have and 0 if they have not;

-"diabetes", if the person has diabetes or not, with value 1 if they have and 0 if they have not;

-"totChol", the amount of total cholesterol, which includes the "good" cholesterol (HDL) and the "bad" cholesterol (LDL), measured in milligrams per deciliter;

-"sysBP", the systolic blood pressure (when the heart contracts) at rest;

-"diaBP" the diastolic blood pressure (when the heart relaxes) at rest;

-"BMI", the body mass index, which indicates the weight compared to the height of the person and it is considered in the following way: underweight if the value is < 18.5, normal weight if the value is between 18.5 and 24.9, overweight if the value is between 25 and 29.9, obese if the value is between 30 and 34.9, and extremely obese if the value is $>= 35$;

-"heartRate", the heart rate of the person at rest, per minute;

-"glucose", the glucose level in the blood (also measured in milligrams per deciliter);

-"TenYearCHD", if the person had coronary heart disease (CHD) in a 10-year lapse. This will be the target variable.

The key steps performed were installing and/or loading the necessary libraries, wrangling the data set, analyzing different variables, balancing the data set, preparing the proper data sets for training and testing,

predicting if a person will have a coronary heart disease event in a 10-year period using a logistic regression model first and then a knn model, and analyzing and comparing the model performances considering their accuracy.

The dataset has been taken from the following kaggle link, from the user Naveen: https://www.kaggle.com/naveengowda16/logistic-regression-heart-disease-prediction?select=framingham_heart_disease.csv

2. CHD PROJECT ANALYSIS

The process and techniques used are, in the following order: installing and/or loading the necessary libraries; downloading the data set; cleaning the data by dropping NAs, eliminating correlated independent variables and balancing the data(the balancing was done at the end, after analyzing the data set), data exploration and visualization of different variables like male, age, level of study attained, current smokers, people taking blood pressure medicine, prevalent strokes, prevalent hypertension, diabetes, cholesterol, BMI, people who had any disease or issue depending on their gender, and coronary heart disease (CHD).

The insights gained will be presented immediately after each exploration or visualization activity is performed.

The modeling approach will be using both logistic regression and knn to predict whether a person will have a CHD. These models fit the needs of this project, since they both can calculate whether a result is 1 (the person had a CHD) or 0 (the person hadn't any CHD) just as the variable TenYearCHD indicates, by using a probabilistic approach (more than 0.5 is 1 and less is 0).

Let's start with the analysis section activities previously indicated.

Installing and/or loading the necessary libraries

```
if(!require(tidyverse)) install.packages("tidyverse",
                                         repos = "http://cran.us.r-project.org")
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
if(!require(caret)) install.packages("caret",
                                     repos = "http://cran.us.r-project.org")
```

```
## Loading required package: caret
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
if(!require(corrplot)) install.packages("corrplot",
                                        repos = "http://cran.us.r-project.org")
```

```
## Loading required package: corrplot
```

```
## corrplot 0.89 loaded
```

```
library(tidyverse)

library(caret)

library(corrplot)
```

Downloading the necessary data to a database called chd_db and having a glimpse of it

```
chd_db <- read.csv("https://raw.githubusercontent.com/Alwabau/Coronary-Heart-Disease-Project-with-R/mai

glimpse(chd_db)
```

```
## Rows: 4,238
## Columns: 16
## $ male            <int> 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, ~
## $ age             <int> 39, 46, 48, 61, 46, 43, 63, 45, 52, 43, 50, 43, 46, 41~
## $ education       <int> 4, 2, 1, 3, 3, 2, 1, 2, 1, 1, 1, 2, 1, 3, 2, 2, 3, 2, ~
## $ currentSmoker   <int> 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, ~
## $ cigsPerDay      <int> 0, 0, 20, 30, 23, 0, 0, 20, 0, 30, 0, 0, 15, 0, 9, 20,~
## $ BPMeds          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, ~
## $ prevalentStroke <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ prevalentHyp    <int> 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, ~
## $ diabetes        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ totChol         <int> 195, 250, 245, 225, 285, 228, 205, 313, 260, 225, 254,~
## $ sysBP           <dbl> 106.0, 121.0, 127.5, 150.0, 130.0, 180.0, 138.0, 100.0~
## $ diaBP           <dbl> 70.0, 81.0, 80.0, 95.0, 84.0, 110.0, 71.0, 71.0, 89.0,~
## $ BMI             <dbl> 26.97, 28.73, 25.34, 28.58, 23.10, 30.30, 33.11, 21.68~
## $ heartRate       <int> 80, 95, 75, 65, 85, 77, 60, 79, 76, 93, 75, 72, 98, 65~
## $ glucose         <int> 77, 76, 70, 103, 85, 99, 85, 78, 79, 88, 76, 61, 64, 8~
## $ TenYearCHD      <int> 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, ~
```

Checking if there are any NA values

```
sum(is.na(chd_db))
```

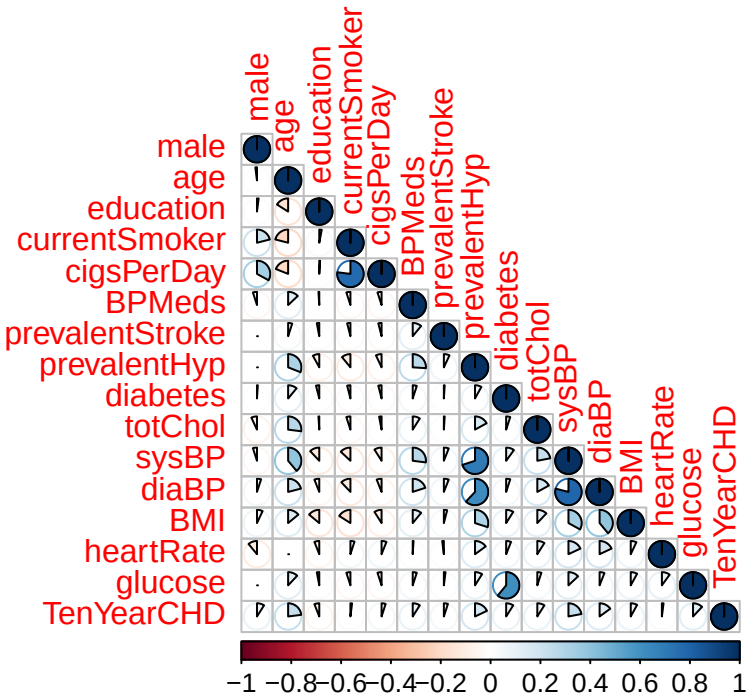
```
## [1] 645
```

Dropping the 645 NA values

```
chd_db <- drop_na(chd_db)
```

Checking the correlation between the variables, since the ones highly correlated might lead to multicollinearity and they need to be dropped

```
corrplot(round(cor(chd_db), digits = 2), method = "pie", type = "lower",
         title = "Correlation among all the variables", mar = c(0,0,1,0))
```



We can see there are a few strong correlations (more than 0.5):

a) cigsPerDay and currentSmoker

b) sysBP and diaBP between themselves and both of them with prevalentHyp

c) glucose and diabetes

I will drop the following columns to simplify the future calculations and avoid multicollinearity: cigsPerDay, sysBP, diaBP and glucose, since I would like to see if being a smoker, having diabetes or hypertension affect the results

```
chd_db <- chd_db %>% select(-cigsPerDay, -sysBP, -diaBP, -glucose)
```

Checking how the data set looks now

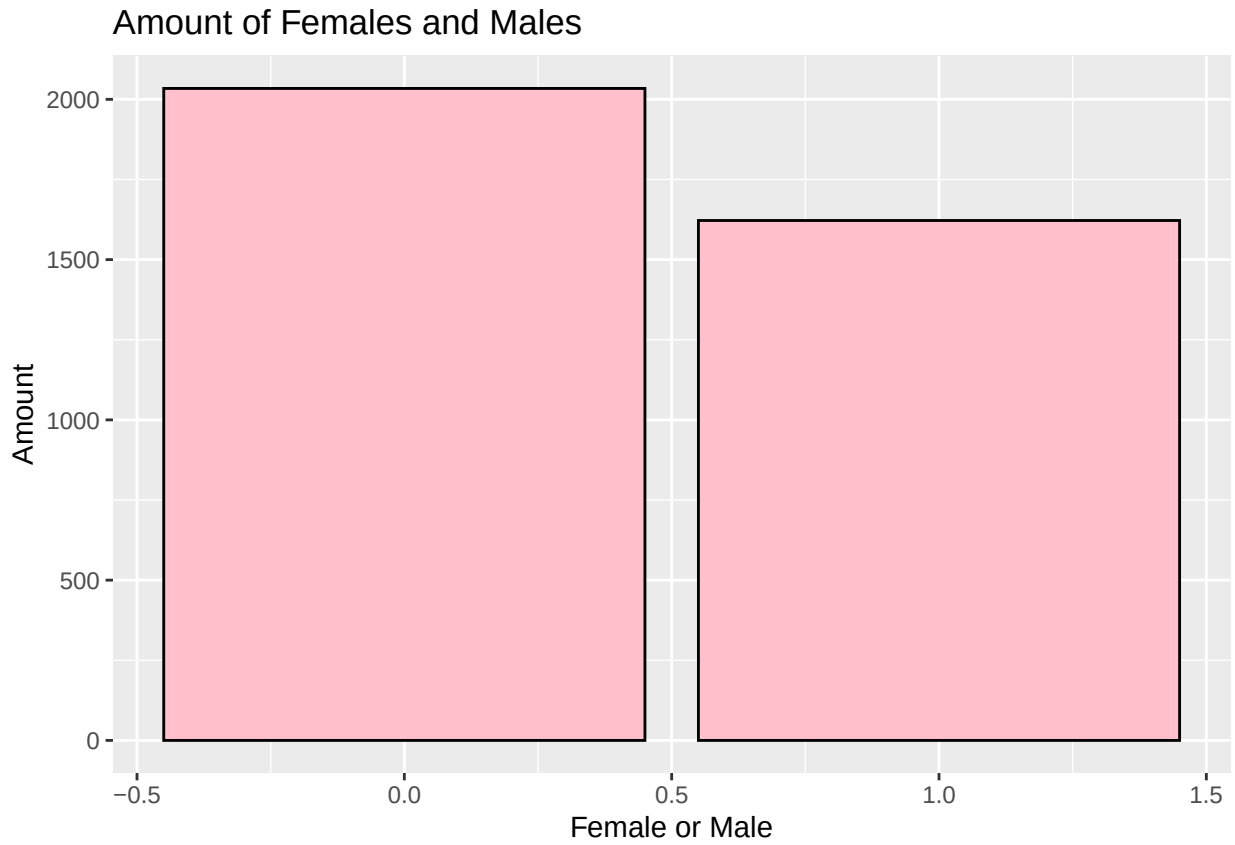
```
glimpse(chd_db)
```

```
## Rows: 3,656
## Columns: 12
## $ male            <int> 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, ~
## $ age             <int> 39, 46, 48, 61, 46, 43, 63, 45, 52, 43, 50, 43, 46, 41~
## $ education       <int> 4, 2, 1, 3, 3, 2, 1, 2, 1, 1, 1, 2, 1, 3, 2, 3, 2, 2, ~
## $ currentSmoker   <int> 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, ~
## $ BPMeds          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ~
## $ prevalentStroke <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ prevalentHyp    <int> 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, ~
## $ diabetes        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ totChol         <int> 195, 250, 245, 225, 285, 228, 205, 313, 260, 225, 254,~
## $ BMI             <dbl> 26.97, 28.73, 25.34, 28.58, 23.10, 30.30, 33.11, 21.68~
## $ heartRate       <int> 80, 95, 75, 65, 85, 77, 60, 79, 76, 93, 75, 72, 98, 65~
## $ TenYearCHD      <int> 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, ~
```

The proper changes have been made, there are four less columns

Checking how many males (value = 1) and how many females (value = 0) there are

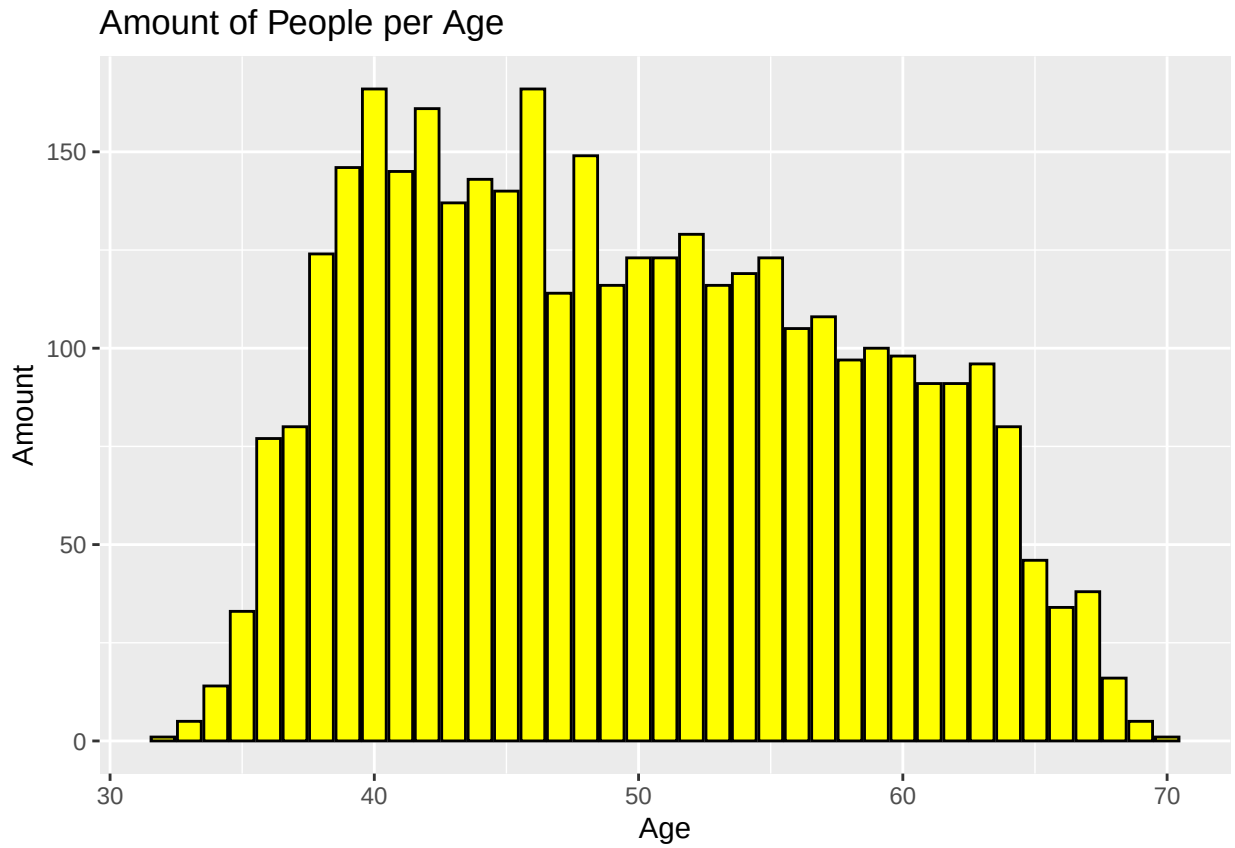
```
chd_db %>% group_by(male) %>% ggplot(aes(male)) +
  geom_bar(fill= "pink", colour = "black") +
  labs(x = "Female or Male", y = "Amount", title = "Amount of Females and Males")
```



There are fewer males, but the difference is not very large

Investigating the amount of people per age

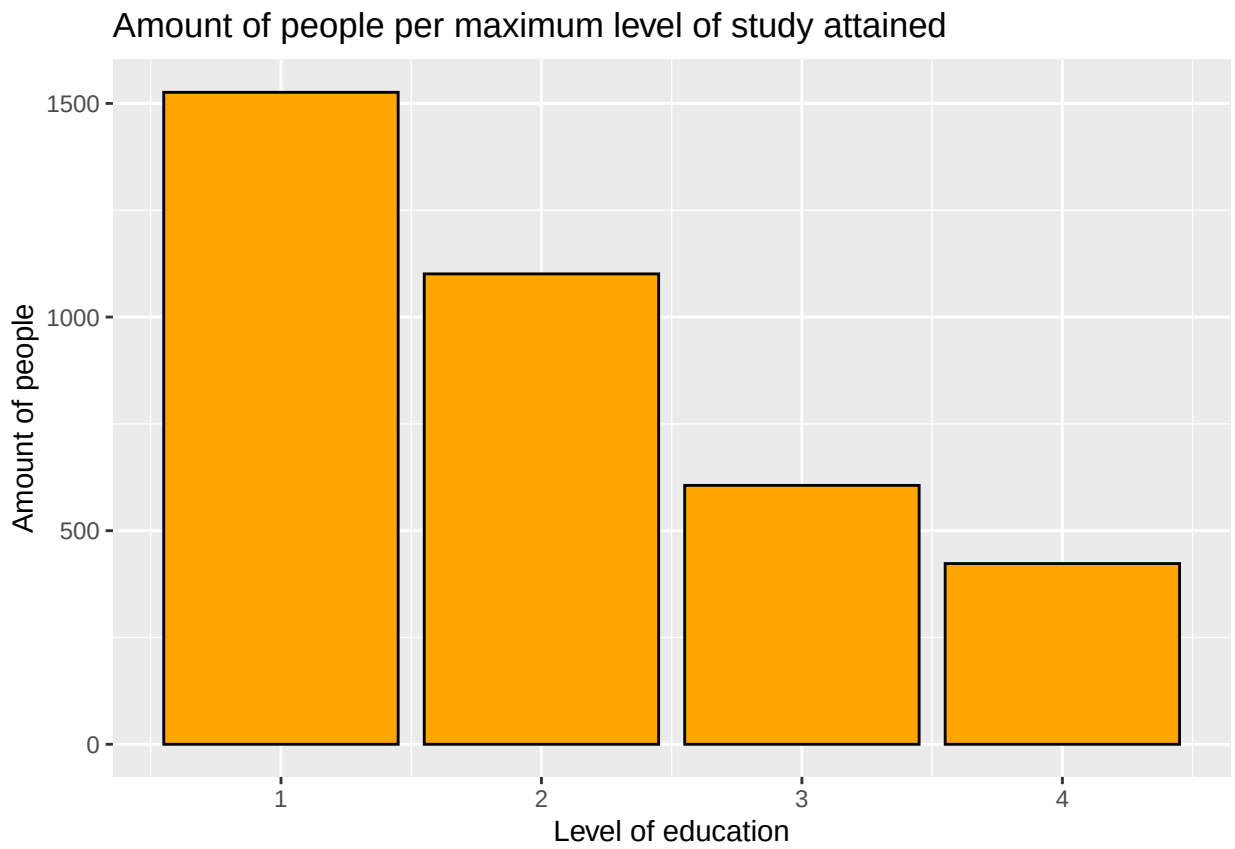
```
chd_db %>% group_by(age) %>% ggplot(aes(age)) +
  geom_bar(fill= "yellow", colour = "black") +
  labs(x = "Age", y = "Amount", title = "Amount of People per Age")
```



The distribution is relatively normal, with a slight right skew

Checking the amount of people according to their maximum level of education: No high school (value = 1), high school degree (value = 2), some college (value = 3), college degree (value = 4)

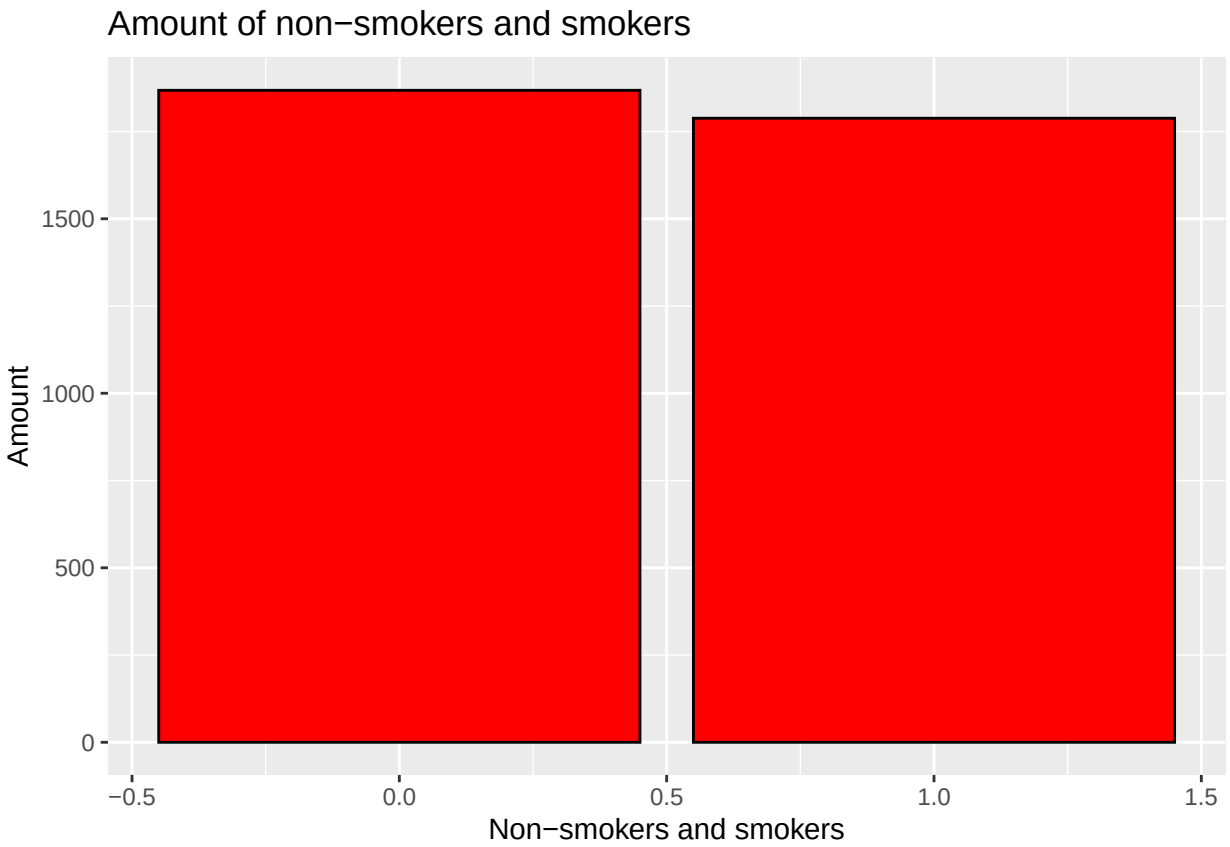
```
chd_db %>% group_by(education) %>% ggplot(aes(education)) +
  geom_bar(fill= "orange", colour = "black") +
  labs(x = "Level of education", y = "Amount of people",
       title = "Amount of people per maximum level of study attained")
```



The graph shows that the higher maximum level of education attained, the fewer the people

Examining the amount of people that smoke (value = 1) or do not (value = 0)

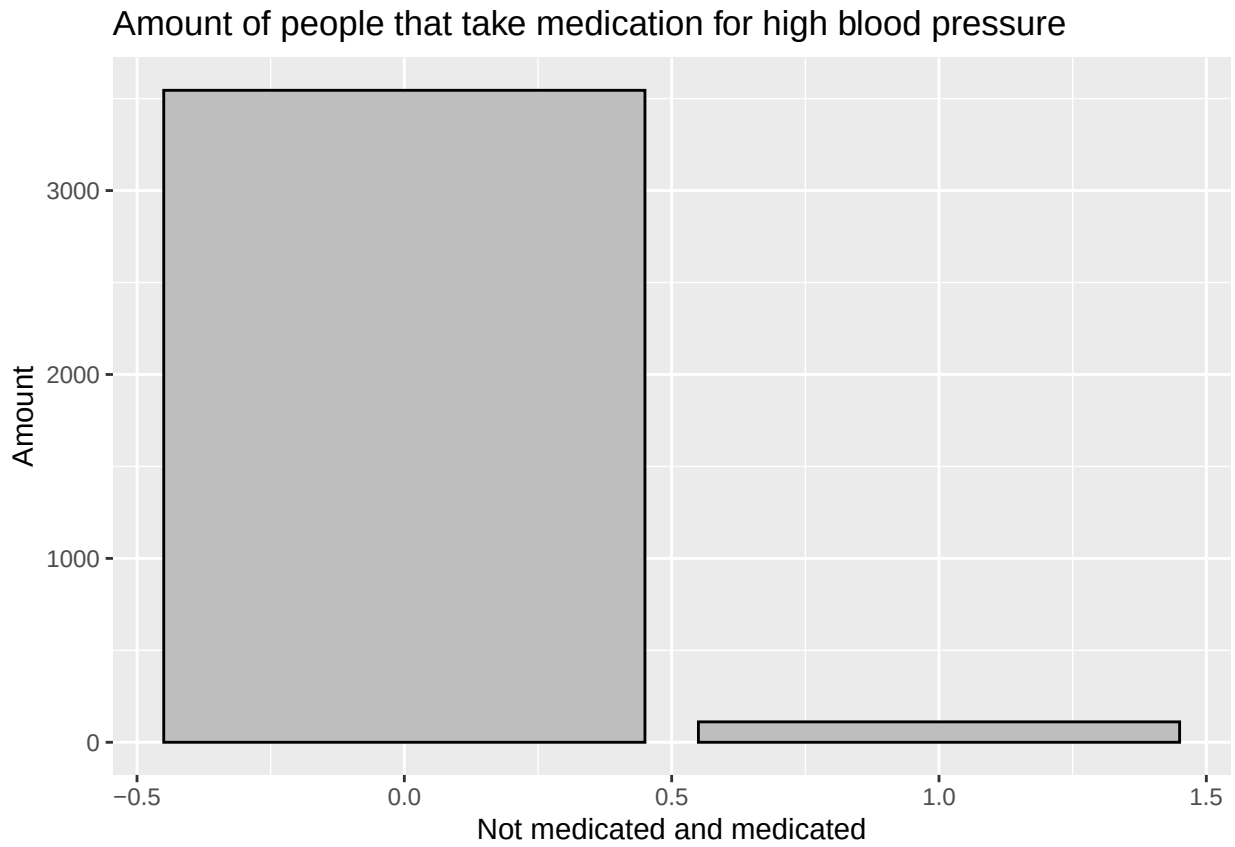
```
chd_db %>% group_by(currentSmoker) %>% ggplot(aes(currentSmoker)) +
  geom_bar(fill= "red", colour = "black") +
  labs(x = "Non-smokers and smokers", y = "Amount",
       title = "Amount of non-smokers and smokers")
```



The amount of people that smoke and don't smoke are relatively even

Inspecting the amount of people that take (value = 1) or don't take medication for high blood pressure (value = 0)

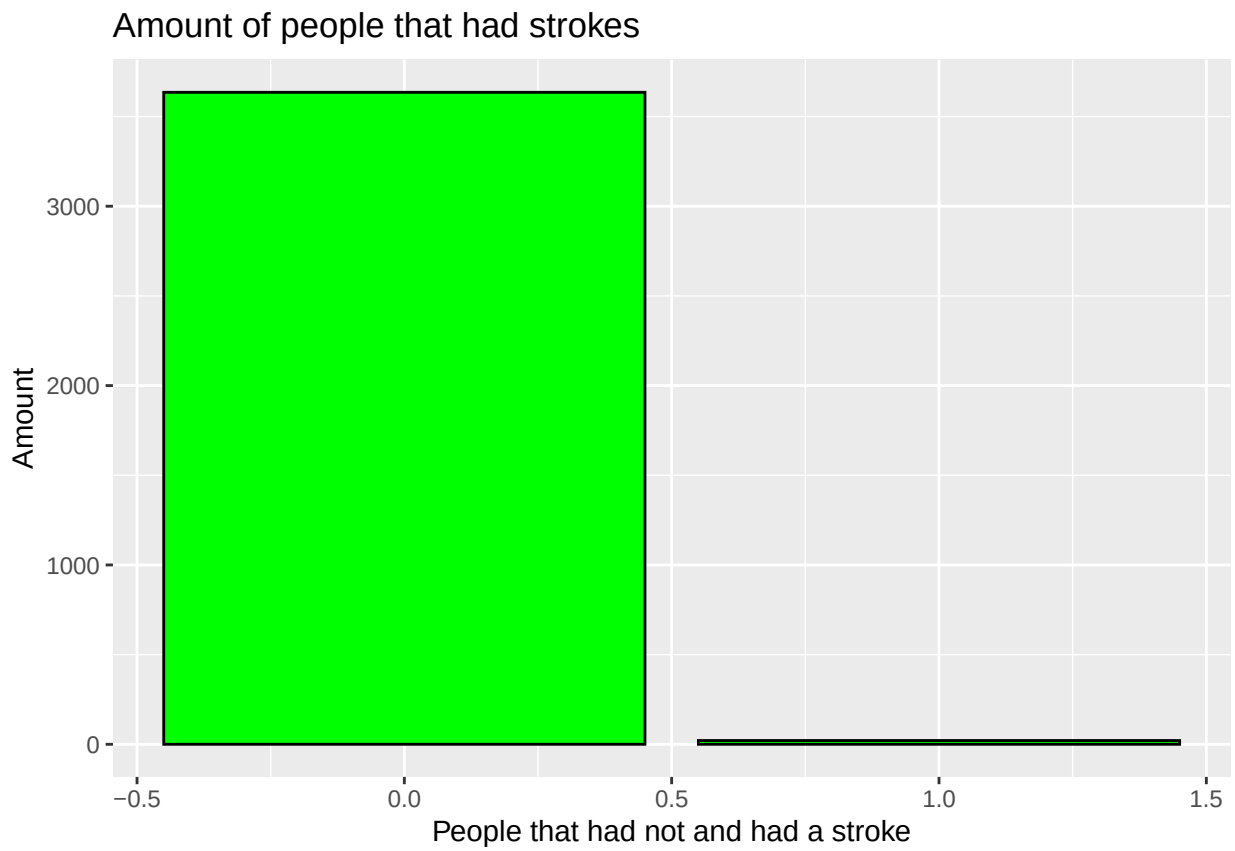
```
chd_db %>% group_by(BPMeds) %>% ggplot(aes(BPMeds)) +
  geom_bar(fill= "gray", colour = "black") +
  labs(x = "Not medicated and medicated", y = "Amount",
       title = "Amount of people that take medication for high blood pressure")
```



The proportion of people who are taking blood pressure medication is very small compared to those who are not

Checking the amount of people that had (value = 1) or had not any strokes (value = 0)

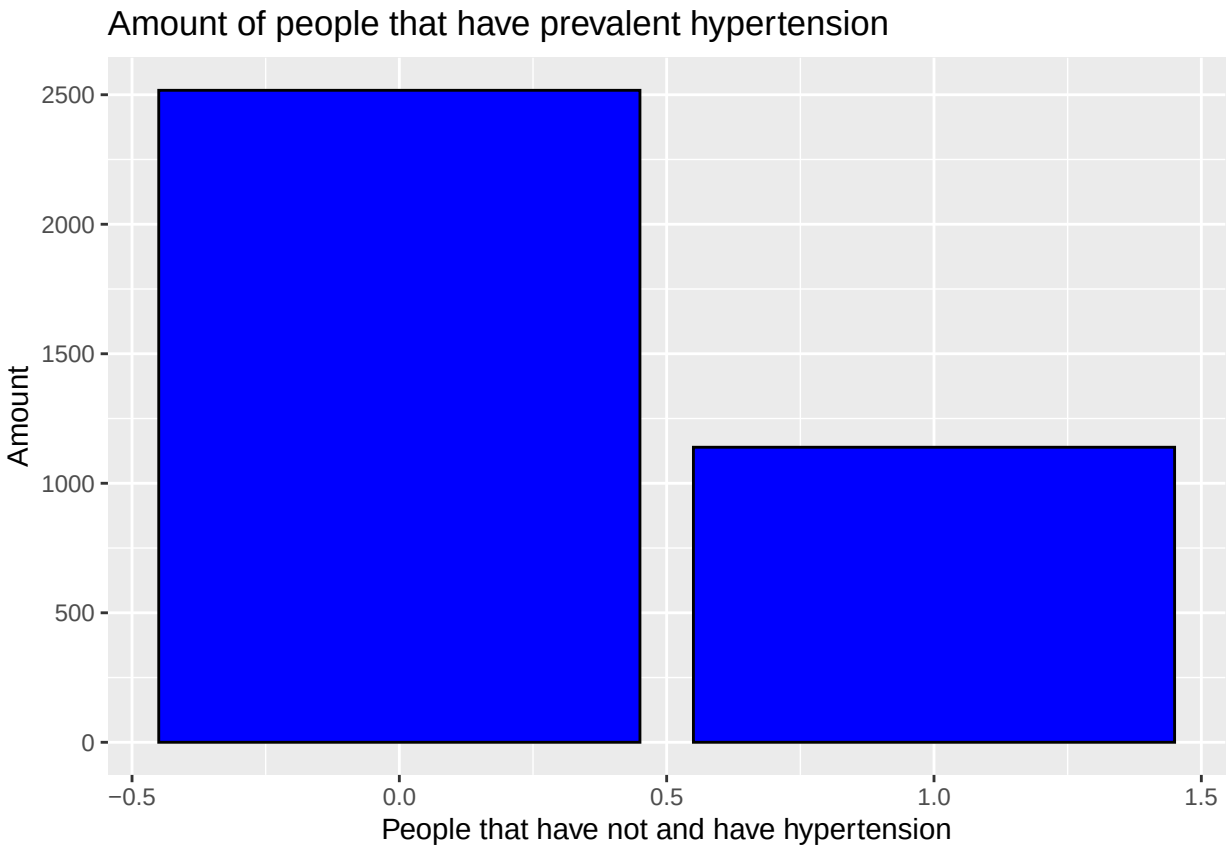
```
chd_db %>% group_by(prevalentStroke) %>% ggplot(aes(prevalentStroke)) +
  geom_bar(fill= "green", colour = "black") +
  labs(x = "People that had not and had a stroke", y = "Amount",
       title = "Amount of people that had strokes")
```



The proportion of people who had a stroke is minuscule

Analyzing the amount of people who have (value = 1) or don't have prevalent hypertension (value = 0)

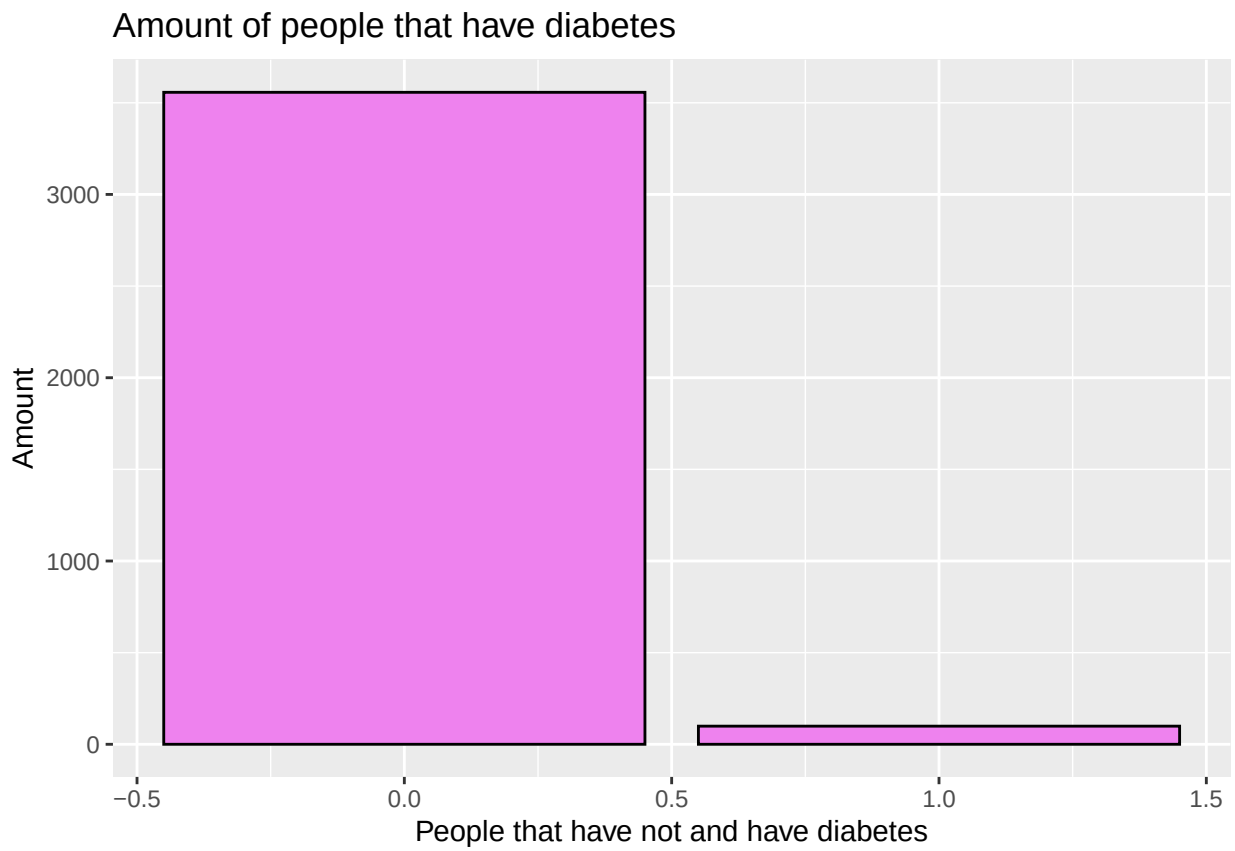
```
chd_db %>% group_by(prevalentHyp) %>% ggplot(aes(prevalentHyp)) +
  geom_bar(fill= "blue", colour = "black") +
  labs(x = "People that have not and have hypertension", y = "Amount",
       title = "Amount of people that have prevalent hypertension")
```



The proportion of people with hypertension is about 1/3 (one third) of the total quantity

Examining the amount of people that have (value = 1) or don't have diabetes (value = 0)

```
chd_db %>% group_by(diabetes) %>% ggplot(aes(diabetes)) +
  geom_bar(fill= "violet", colour = "black") +
  labs(x = "People that have not and have diabetes", y = "Amount",
       title = "Amount of people that have diabetes")
```



The amount of people with diabetes is quite small compared to those who don't have it

Checking the range of the total cholesterol variable

```
range(chd_db$totChol)
```

```
## [1] 113 600
```

The range goes from 113 to 600. It would be interesting to see how many people are under the limit of cholesterol, who are on it and who exceed it

Calculating the percentage of people that are under the total cholesterol limit (cholesterol equal to or less than 200, under the name of under_chol), on the limit (cholesterol more than 200 and under 240, under the name of limit_chol) and above it(cholesterol equal to or over 240, under the name of exceed_chol) to later graph their distribution

```
under_chol <- round(sum(chd_db$totChol <= 200) * 100 / NROW(chd_db$totChol),
                    digits = 2)

limit_chol <- round(sum(chd_db$totChol < 240 & chd_db$totChol > 200) * 100 /
                      NROW(chd_db$totChol), digits = 2)

exceed_chol <- round(sum(chd_db$totChol >= 240) * 100 / NROW(chd_db$totChol),
                     digits = 2)
```

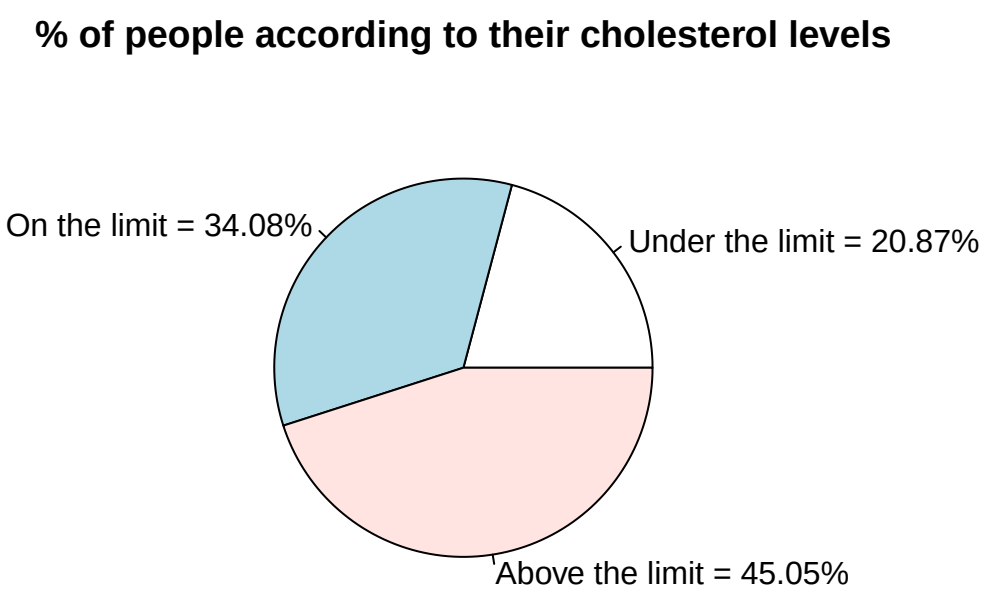
Creating a pie chart with the distribution of the people according to their cholesterol levels, that were just calculated in under_chol, limit_chol and exceed_chol

```r
col_data <- c(under_chol, limit_chol, exceed_chol)

col_label <- c("Under the limit", "On the limit", "Above the limit")

col_labels <- paste0(col_label, " = ", col_data, "%")

pie(col_data, col_labels,
    main = "% of people according to their cholesterol levels")
```



We can see that around 79% of the people are on the cholesterol limit or above it, and only around 21% are under it

Checking the range of the body mass index (BMI) variable

```r
range(chd_db$BMI)
```

```
## [1] 15.54 56.80
```

Calculating the percentage of people that are underweight (BMI less than 18.5, under the name of underweight), those who have a normal weight(BMI equal to or more than 18.5 and under 25, under the name of normal_weight) and those who are overweight or obese (BMI equal to or over 25, under the name of overweight) to later graph their distribution taking into consideration their BMI

```r
underweight <- round(sum(chd_db$BMI < 18.5) * 100 / NROW(chd_db$BMI), digits = 2)
```

```r
normal_weight <- round(sum(chd_db$BMI >= 18.5 & chd_db$BMI < 25) * 100 /
                         NROW(chd_db$BMI), digits = 2)

overweight <- round(sum(chd_db$BMI >= 25) * 100 / NROW(chd_db$BMI), digits = 2)
```

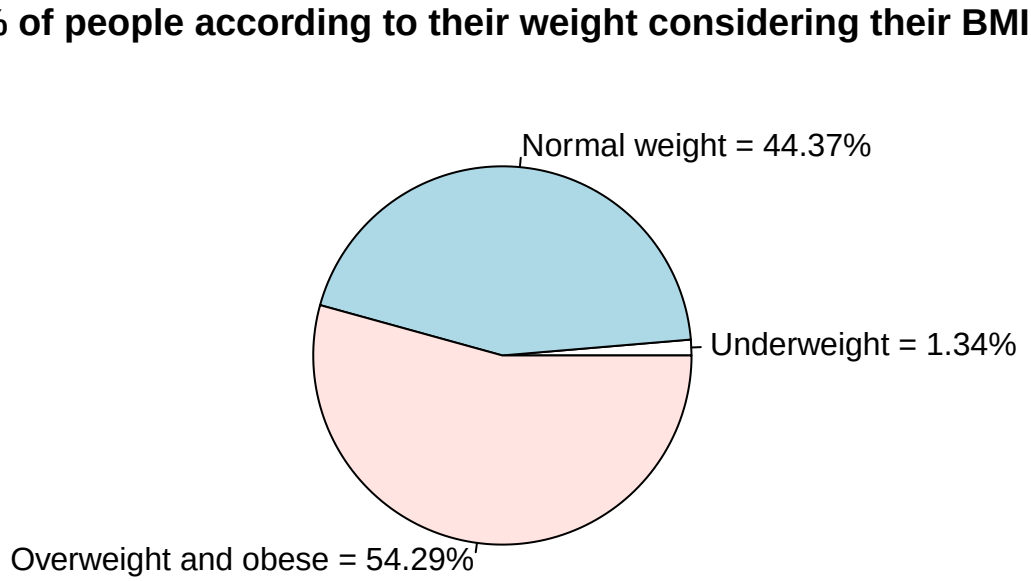
Creating a pie chart with the distribution of the people according to their cholesterol levels, that were just calculated in underweight, normal_weight and overweight

```r
BMI_data <- c(underweight, normal_weight, overweight)

BMI_label <- c("Underweight", "Normal weight", "Overweight and obese")

BMI_labels <- paste0(BMI_label, " = ", BMI_data, "%")

pie(BMI_data, BMI_labels,
    main = "% of people according to their weight considering their BMI")
```



We can see that only around 44.4% of the people have a normal weight. More than half of them (only a very small percentage are underweight) have a weight that could potentially lead to health issues

Calculating the amount of males that had any type of disease or issue considered in the data set (a stroke, hypertension, diabetes or coronary heart disease)

```r
male_disease <- NROW(chd_db[(chd_db$male == 1) & (chd_db$prevalentStroke == 1 |
                                                    chd_db$prevalentHyp == 1 |
                                                    chd_db$diabetes == 1 |
```

```
                                                    chd_db$TenYearCHD == 1),])

male_disease
```

```
## [1] 692
```

Calculating the percentage of males who had any of the previously mentioned problems

```
male_disease * 100 / NROW(chd_db[chd_db$male == 1,])
```

```
## [1] 42.66338
```

Around 42.7% of the males had or has one of the mentioned problems

Calculating the amount of females that had any type of disease or issue considered in the data set (a stroke, hypertension, diabetes or coronary heart disease)

```
female_disease <- NROW(chd_db[(chd_db$male == 0) & (chd_db$prevalentStroke == 1 |
                                                      chd_db$prevalentHyp == 1 |
                                                      chd_db$diabetes == 1 |
                                                      chd_db$TenYearCHD == 1),])

female_disease
```

```
## [1] 759
```

Calculating the percentage of females who had any of the previously mentioned problems

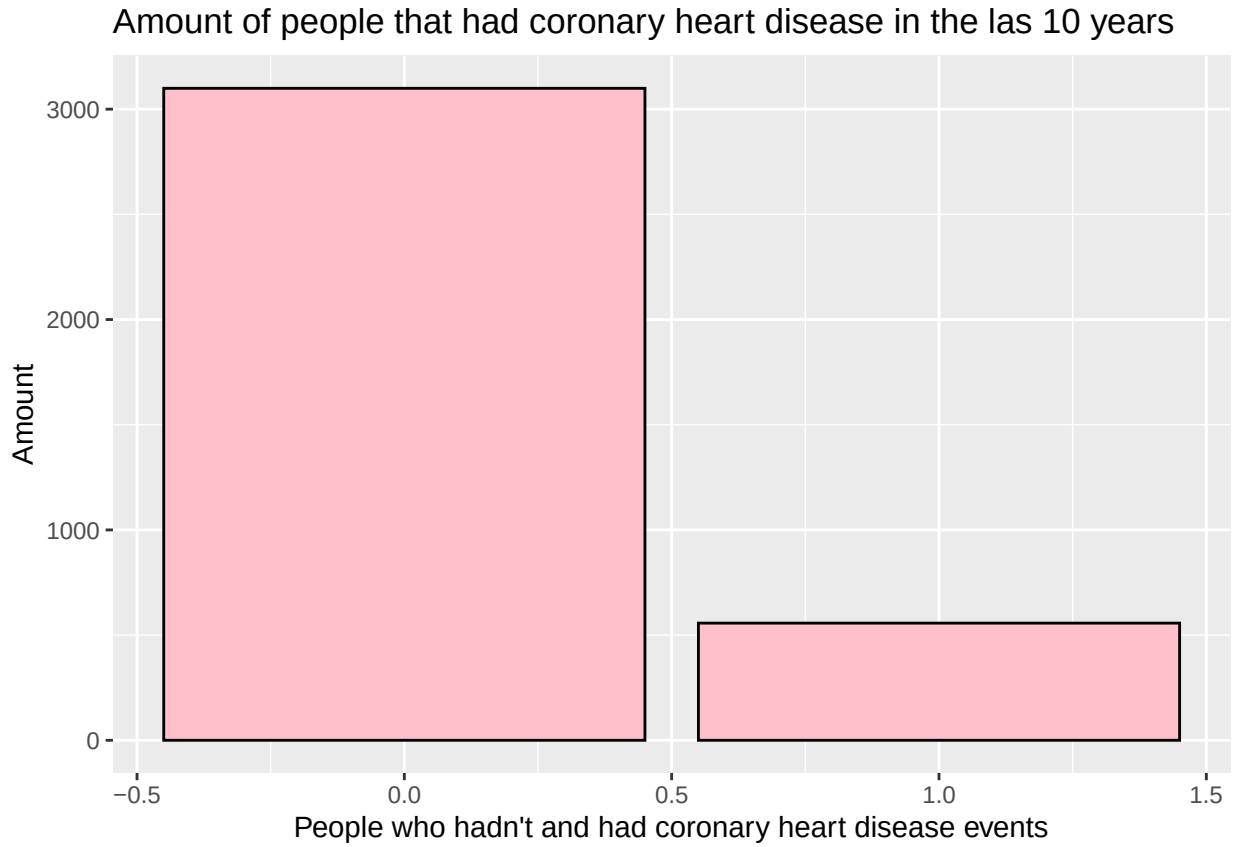
```
female_disease * 100 / NROW(chd_db[chd_db$male == 0,])
```

```
## [1] 37.31563
```

Around 37.3% of the males had or has one of the mentioned problems, which is around 5% less than males

Inspecting the amount of people that had coronary heart disease (CHD) in the last 10 years (value = 1) and those who had not (value = 0)

```
chd_db %>% group_by(TenYearCHD) %>% ggplot(aes(TenYearCHD)) +
  geom_bar(fill= "pink", colour = "black") +
  labs(title = "Amount of people that had coronary heart disease in the las 10 years",
       x = "People who hadn't and had coronary heart disease events", y = "Amount")
```



There are many more cases that didn't have a CHD event that the ones that did. Let's see what percentage of the cases did have

Calculating what percentage of cases with CHD there are in the data set

```
round(sum(chd_db$TenYearCHD)/nrow(chd_db)*100, digits = 2)
```

```
## [1] 15.24
```

Only 15.24% of people had CHD. Since this is the target variable we should have around the same amount of cases that had and that had not any CHD in a 10-year period to train and test the models properly. The data set is unbalanced and this issue needs to be addressed. I will double the amount of cases for the target value 1 and then reduce the target value 0 cases to the same amount of the value 1

Doubling the amount of cases of the target variable of value 1 and checking the amount

```
chd_db_y_1 <- chd_db[(chd_db$TenYearCHD == 1),]

chd_db_y_1 <- chd_db_y_1 %>% slice(rep(1:n(), each = 2))

NROW(chd_db_y_1)
```

```
## [1] 1114
```

Reducing the amount of cases of the target variable of value 0 to 1114 and checking the amount

```r
chd_db_y_0 <- chd_db[(chd_db$TenYearCHD == 0),]

chd_db_y_0 <- chd_db_y_0[(1:1114),]

NROW(chd_db_y_0)
```

```
## [1] 1114
```

Choosing around 70% of cases for the train set from each target variable value set (chd_db_y_0 and chd_db_y_1) and creating a training set with this information under the name of train_set

```r
train_set_0 <- chd_db_y_0[1:780,]

train_set_1 <- chd_db_y_1[1:780,]

train_set <- rbind(train_set_0,train_set_1)
```

Choosing around 30% of cases for the test set from each target variable value set (chd_db_y_0 and chd_db_y_1) and creating a test set with this information under the name of test_set

```r
test_set_0 <- chd_db_y_1[781:1114,]

test_set_1 <- chd_db_y_1[781:1114,]

test_set <- rbind(test_set_0,test_set_1)
```

Since we have to choose between 2 categorical values (0 or 1, or not having CHD or having it) a logistic regression model will be used to fit and predict the probabilities of having CHD in a 10-year period first, and then a knn model. Their accuracy will be compared later.

Fitting the logistic regression model with the train set under the name of log_reg_fit

```r
log_reg_fit <- glm(TenYearCHD ~ ., data = train_set, family=binomial)
```

Predicting the probabilities of the test set values to be either 0 or 1. For that I will also be transforming the probabilities into 1 if they are more than 0.5 or to 0 if they are less than that threshold

```r
log_prediction <- predict(log_reg_fit, test_set, type = "response")

log_prediction <- ifelse(log_prediction > 0.5,1,0)
```

Calculating the accuracy of the model

```r
round(mean(log_prediction == test_set$TenYearCHD), digits = 2)
```

```
## [1] 0.72
```

The logistic regression model has been able to accurately predict 72% of the time if a person would have coronary heart disease or not

Now let's move to the knn model

Fitting the knn model with the train set under the name of knn_fit

```
knn_fit <- knn3(TenYearCHD ~ ., data = train_set, k = 1)
```

Predicting the probabilities of the test set values to be either 0 or 1. This will create a data frame with 2 columns filled with ones (this indicates positive) or zeros (this indicates negative): X0 that shows the prediction of the model being 0, and X1 that shows the prediction of the model being 1. Both columns are of course mutually exclusive, as shown below the code

```
knn_prediction <- data.frame(predict(knn_fit, test_set, type = "prob"))

head(knn_prediction)
```

```
##   X0 X1
## 1  1  0
## 2  1  0
## 3  0  1
## 4  0  1
## 5  0  1
## 6  0  1
```

We can see for example that in the first two cases the prediction is 0 and the other next four cases it is 1. Then, column X1 is the one we need to compare with the real results from the test set (since we want to know if the prediction is 1, just like the TenYearCHD target variable)

Calculating the accuracy of the model

```
round(mean(knn_prediction$X1 == test_set$TenYearCHD), digits = 2)
```

```
## [1] 0.42
```

The knn model has been able to accurately predict 42% of the time if a person would have coronary heart disease or not

3. CHD PROJECT RESULTS

The logistic regression model has performed considerably better (72% of accuracy) than the knn one (42% of accuracy) at predicting if a person had a coronary heart disease event in a 10-year period, correctly predicting 30% more of cases.

The knn has actually performed poorly, below the 50% accuracy mark.

If any of these two models should be considered again to predict CHD events in the future the logistic regression one should be used.

4. CHD PROJECT CONCLUSION

This report has shown that a logistic regression model can predict having a CHD event in a 10-year period with an accuracy of 72% using the following variables: gender, age, level of education, being a current smoker, taking blood pressure medication, having a stroke, having hypertension, having diabetes, the total cholesterol levels, the body mass index and the heart rate.

The potential impact of this could be knowing beforehand if a person has chances of having a coronary heart disease and thus correcting or controlling if possible those variables that affect CHD.

The limitations of this report are, on the one hand, that it still fails to accurately predict 28% of the cases, and on the other hand, that our health is affected by a wide range of factors that are not included here, and that could also potentially affect the result of having a CHD or not.

The Framingham study should start considering more variables in the near future to see if there are any other that might be important to determine the chances of having CHD.