Data Exploration - Cardiovascular Study Dataset

Autor: “Ali Abdorrahimi”
Studiengang: “Digital Engineering”
Fach: “Data Mining”
Semester : “Wintersemester 2022”

Projektbericht

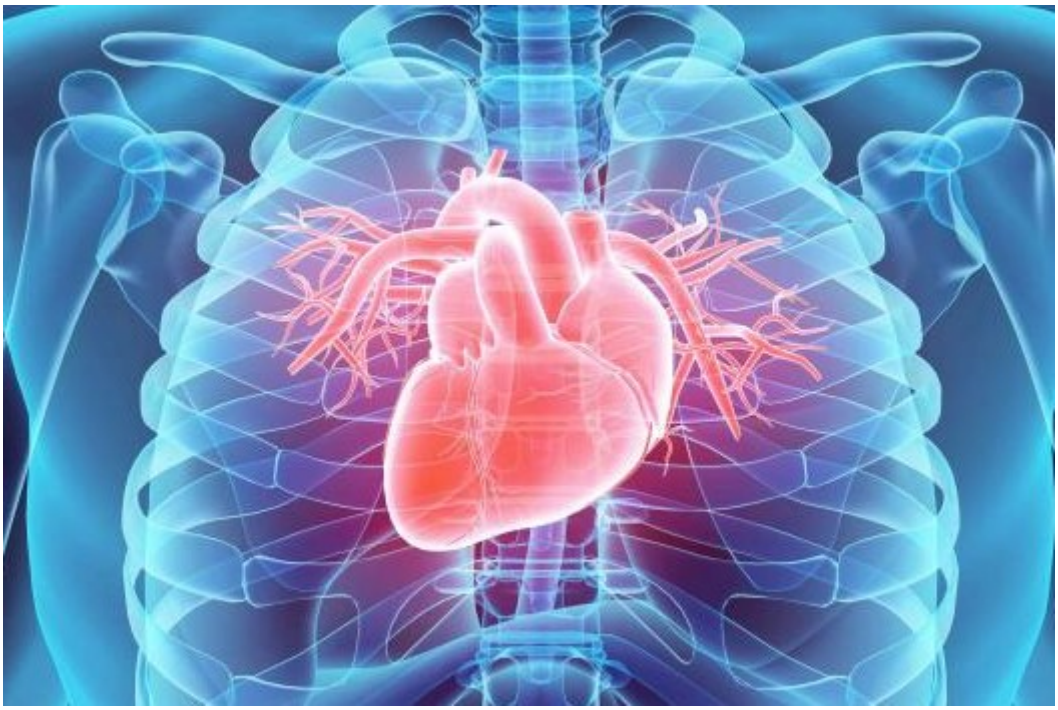


Figure 1: Cardio

Die WHO schätzt, dass jedes Jahr weltweit 12 Millionen Todesfälle auf Herzkrankheiten zurückzuführen sind. Die Hälfte der Todesfälle in den USA und anderen entwickelten Ländern ist auf cardio vascular Krankheiten zurückzuführen. Die frühzeitige Prognose von Cardio-Erkrankungen kann dazu beitragen, dass bei Risikopatienten Entscheidungen zur Änderung der Lebensweise getroffen werden, was wiederum zu einer Verringerung der Komplikationen führt. Ziel dieser Untersuchung ist es, die wichtigsten Risikofaktoren für Herzkrankheiten zu ermitteln und das Gesamtrisiko mithilfe Prediction model vorherzusagen.

Load Libraries

```
library(magrittr)
library(plyr)
library(dplyr)
library(ggplot2)
library(grid)
library(gridExtra)
library(stringr)
library(here)
library(VIM)
```

Load Data

```
cardio_raw= read.csv(here::here ("D:\\DataMining\\train.csv"))
```

Datensatzbeschreibung

Der Datensatz ist auf der Kaggle-Website öffentlich verfügbar und stammt aus einer laufenden kardiovaskulären Studie über Einwohner der Stadt Framingham, Massachusetts. Ziel der Klassifizierung ist die Vorhersage, ob ein Patient ein 10-Jahres-Risiko für eine künftige koronare Herzkrankheit (CHD) hat. Der Datensatz enthält Informationen über die Patienten. Er umfasst über 4.000 Datensätze und 15 Attribute.

Variablen

Jedes Attribut ist ein potenzieller Risikofaktor. Es gibt sowohl demografische, verhaltensbezogene als auch medizinische Risikofaktoren.

Demographic:

- Sex: male or female(“M” or “F”)
- Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous) Behavioral
- is_smoking: whether or not the patient is a current smoker (“YES” or “NO”)
- Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

Medical(history)

- BP Meds: whether or not the patient was on blood pressure medication (Nominal)
- Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
- Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
- Diabetes: whether or not the patient had diabetes (Nominal)

Medical(current) • Tot Chol: total cholesterol level (Continuous)

- Sys BP: systolic blood pressure (Continuous)
- Dia BP: diastolic blood pressure (Continuous)
- BMI: Body Mass Index (Continuous)

• Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)

• Glucose: glucose level (Continuous) Predict variable (desired target) • 10 year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”)

Descriptive Statistics

R’s str function gibt uns einen Blick auf die Datentypen im Datensatz , die head function druckt die ersten 5 Zeilen. Mit der summary-function können wir grundlegende Zusammenfassungsstatistiken für jede Spalte anzeigen.

Die ersten 5 Zeilen anzeigen.

```
head(cardio_raw)
```

```
##   id age education sex is_smoking cigsPerDay BPMeds prevalentStroke
## 1  0  64         2   F        YES          3      0               0
## 2  1  36         4   M         NO          0      0               0
## 3  2  46         1   F        YES         10      0               0
## 4  3  50         1   M        YES         20      0               0
## 5  4  64         1   F        YES         30      0               0
## 6  5  61         3   F         NO          0      0               0
##   prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose TenYearCHD
## 1            0        0     221 148.0    85    NA        90      80          1
## 2            1        0     212 168.0    98 29.77        72      75          0
## 3            0        0     250 116.0    71 20.35        88      94          0
## 4            1        0     233 158.0    88 28.26        68      94          1
## 5            0        0     241 136.5    85 26.42        70      77          0
## 6            1        0     272 182.0   121 32.80        85      65          1
```

Zeigt Strukturinformationen über den Datenrahmen an.

```
str(cardio_raw)
```

```
## 'data.frame':    3390 obs. of  17 variables:
##  $ id             : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ age            : int  64 36 46 50 64 61 61 36 41 55 ...
##  $ education      : num  2 4 1 1 1 3 1 4 2 2 ...
##  $ sex            : chr  "F" "M" "F" "M" ...
##  $ is_smoking     : chr  "YES" "NO" "YES" "YES" ...
##  $ cigsPerDay     : num  3 0 10 20 30 0 0 35 20 0 ...
##  $ BPMeds         : num  0 0 0 0 0 0 0 0 NA 0 ...
##  $ prevalentStroke: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ prevalentHyp   : int  0 1 0 1 0 1 1 0 0 1 ...
##  $ diabetes       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ totChol        : num  221 212 250 233 241 272 238 295 220 326 ...
##  $ sysBP          : num  148 168 116 158 136 ...
##  $ diaBP          : num  85 98 71 88 85 121 136 68 78 81 ...
##  $ BMI            : num  NA 29.8 20.4 28.3 26.4 ...
##  $ heartRate      : num  90 72 88 68 70 85 75 60 86 85 ...
##  $ glucose        : num  80 75 94 94 77 65 79 63 79 NA ...
##  $ TenYearCHD     : int  1 0 0 1 0 1 0 0 0 0 ...
```

Zusammenfassende Statistiken pro Spalte anzeigen.

```
summary(cardio_raw)
```

```
##        id              age          education         sex
##  Min.   :   0.0   Min.   :32.00   Min.   :1.000   Length:3390
##  1st Qu.: 847.2   1st Qu.:42.00   1st Qu.:1.000   Class :character
##  Median :1694.5   Median :49.00   Median :2.000   Mode  :character
##  Mean   :1694.5   Mean   :49.54   Mean   :1.971
##  3rd Qu.:2541.8   3rd Qu.:56.00   3rd Qu.:3.000
##  Max.   :3389.0   Max.   :70.00   Max.   :4.000
##                                   NA's   :87
##   is_smoking          cigsPerDay         BPMeds        prevalentStroke
##  Length:3390        Min.   : 0.000   Min.   :0.00000   Min.   :0.00000
##  Class :character   1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.00000
##  Mode  :character   Median : 0.000   Median :0.00000   Median :0.00000
##                     Mean   : 9.069   Mean   :0.02989   Mean   :0.00649
##                     3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.00000
##                     Max.   :70.000   Max.   :1.00000   Max.   :1.00000
##                     NA's   :22       NA's   :44
##   prevalentHyp       diabetes          totChol          sysBP
##  Min.   :0.0000   Min.   :0.00000   Min.   :107.0   Min.   : 83.5
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:206.0   1st Qu.:117.0
##  Median :0.0000   Median :0.00000   Median :234.0   Median :128.5
##  Mean   :0.3153   Mean   :0.02566   Mean   :237.1   Mean   :132.6
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:264.0   3rd Qu.:144.0
##  Max.   :1.0000   Max.   :1.00000   Max.   :696.0   Max.   :295.0
##                                     NA's   :38
##      diaBP             BMI          heartRate         glucose
##  Min.   : 48.00   Min.   :15.96   Min.   : 45.00   Min.   : 40.00
##  1st Qu.: 74.50   1st Qu.:23.02   1st Qu.: 68.00   1st Qu.: 71.00
##  Median : 82.00   Median :25.38   Median : 75.00   Median : 78.00
##  Mean   : 82.88   Mean   :25.79   Mean   : 75.98   Mean   : 82.09
##  3rd Qu.: 90.00   3rd Qu.:28.04   3rd Qu.: 83.00   3rd Qu.: 87.00
##  Max.   :142.50   Max.   :56.80   Max.   :143.00   Max.   :394.00
##                   NA's   :14      NA's   :1        NA's   :304
##    TenYearCHD
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.1507
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
```

Cleaning and Preparing the Data

Aufgrund der Ergebnisse der obigen Funktion wurden mehrere Probleme mit dem Import der Daten durch die Funktion read.csv festgestellt, die vor einer tiefer gehenden Analyse behoben werden müssen:

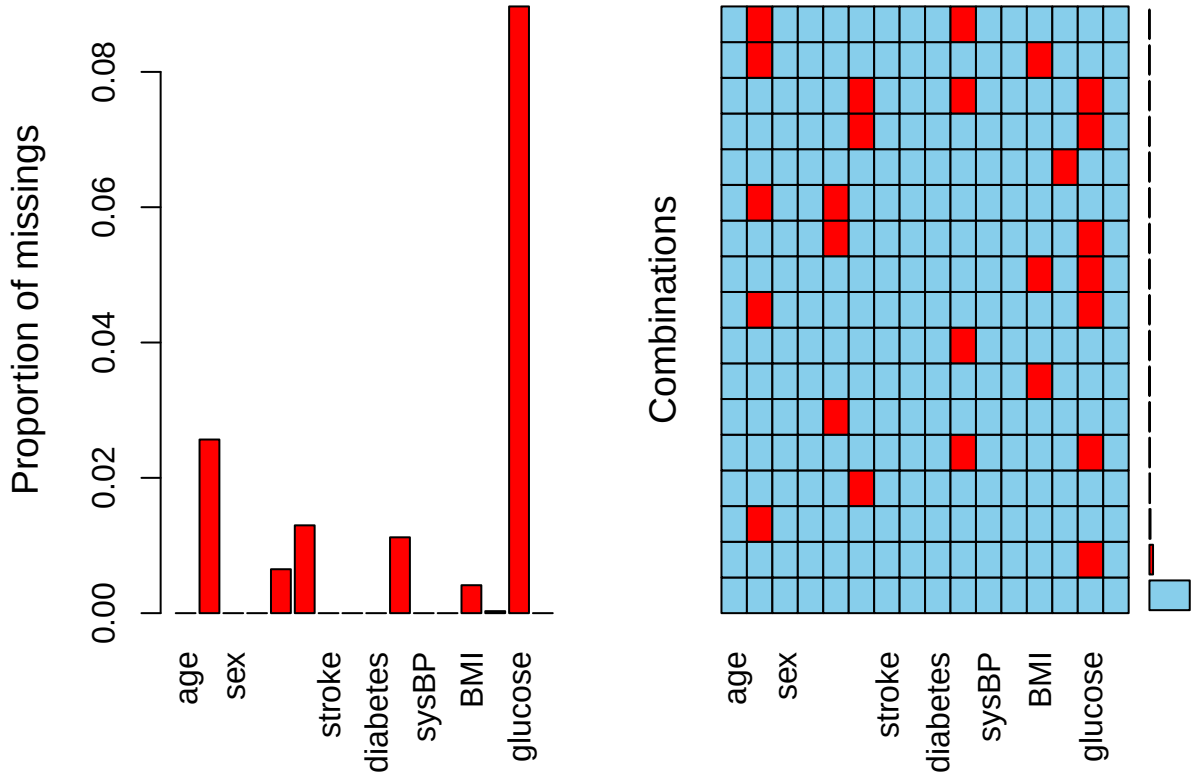
```
workPath ="D:\\DataMining\\"

cardio <- cardio_raw[ , -1 ]

cardio$education <- factor(cardio$education)

cardio$sex <- ifelse (cardio$sex == "F", "female", "male")

cardio$is_smoking <- ifelse (cardio$is_smoking == "YES", "smoking", "not smoking")

colnames(cardio)[4] <- "smoking"

colnames(cardio)[6] <- "BloodPresMed"

cardio$BloodPresMed <- ifelse (cardio$BloodPresMed == 0, "no", "yes")

cardio$prevalentStroke <- ifelse (cardio$prevalentStroke == 0, "no stroke", "stroke")
colnames(cardio)[7] <- "stroke"

cardio$TenYearCHD <- ifelse (cardio$TenYearCHD == 0, "healthy", "CHD")

colnames(cardio)[8] <- "hypertensive"
cardio$hypertensive <- ifelse (cardio$hypertensive == 0, "no hypertensive", "hypertensive")

cardio$diabetes <- ifelse (cardio$diabetes == 0, "no diabetes", "diabetes")

colnames(cardio)[ncol(cardio)] <- "target"


## hilfsdatensätze
cardio_chd = subset(cardio,target == "CHD" )
cardio_healthy = subset(cardio,target == "healthy" )

saveRDS(cardio, file  =paste (workPath , "cardio.rds", sep =""))
```

Missing Data

Glukose hat die meisten fehlenden Werte, etwas mehr als 8 % des gesamten Datensatzes. Es gibt auch einige Beobachtungen, bei denen wichtige Variablen wie BloodPresMed und Education fehlen.

Die gute Nachricht ist, dass die Variablen age, sex , stroke ,smoking ,hypertensive, diabetes, sysBP und diaBP sind und keine fehlenden Werte aufweisen.

```
aggr(cardio)
```



Univariate Plots

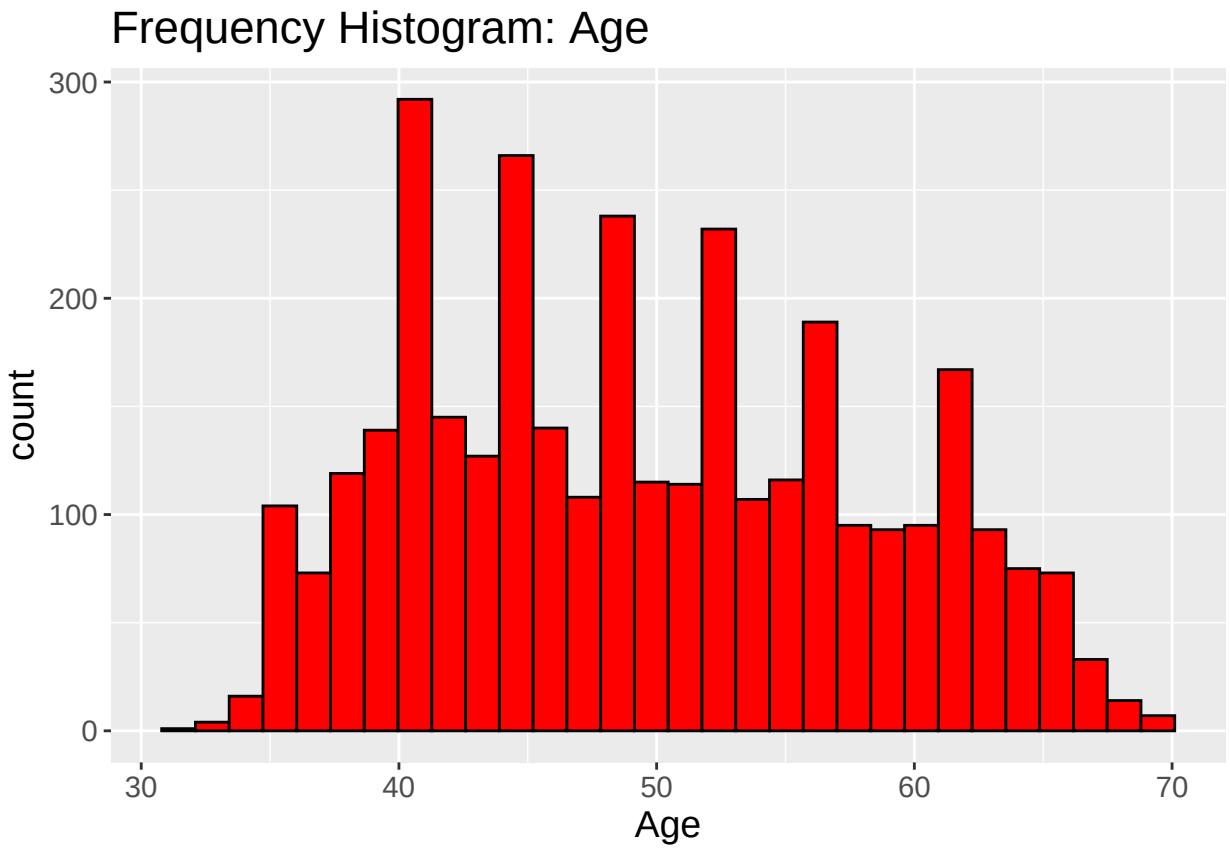
In diesem Abschnitt werde ich einen Blick auf die Verteilung der Werte für jede Variable im Datensatz werfen, indem ich Histogramme mit der ggplot-Funktion von ggplot2 erstelle. Ich versuche herauszufinden, ob es mehr Daten gibt, die bereinigt werden müssen, einschließlich Ausreißer oder fremde Werte. Dies könnte mir auch helfen, Beziehungen zwischen Variablen zu erkennen, die es wert sind, weiter untersucht zu werden.

Alter

Frequency Histogram

```
# By Age
cardio %>%
  ggplot( aes(x = age) ) +
    geom_histogram(color="black", fill="red")+
  theme(text = element_text(size=14)) +
    labs ( title = "Frequency Histogram: Age" )  +
          xlab ("Age") +
          ylab  ("count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

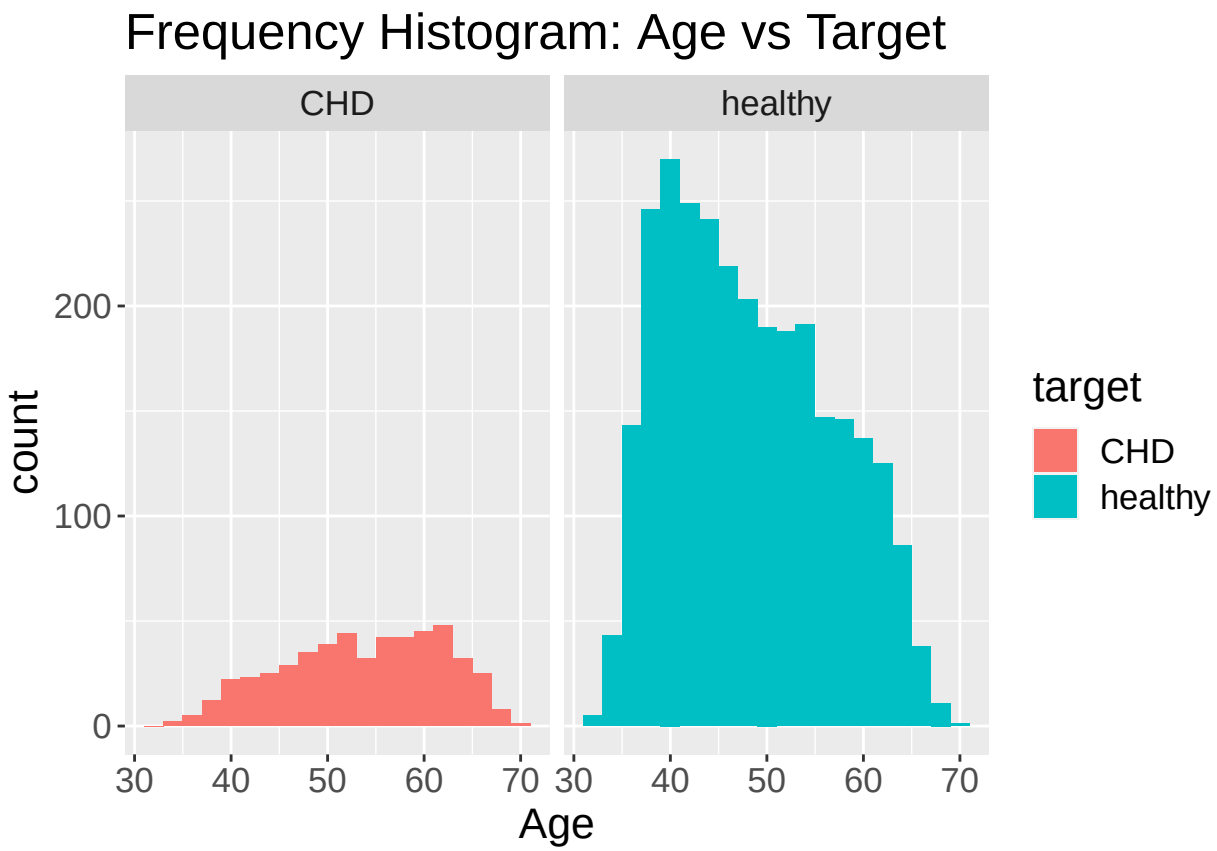


Interpretation:

Der Großteil der Patienten ist zw. 40 und 60 Jahren. Nur sehr wenige sind unter 35 bzw. über 65.

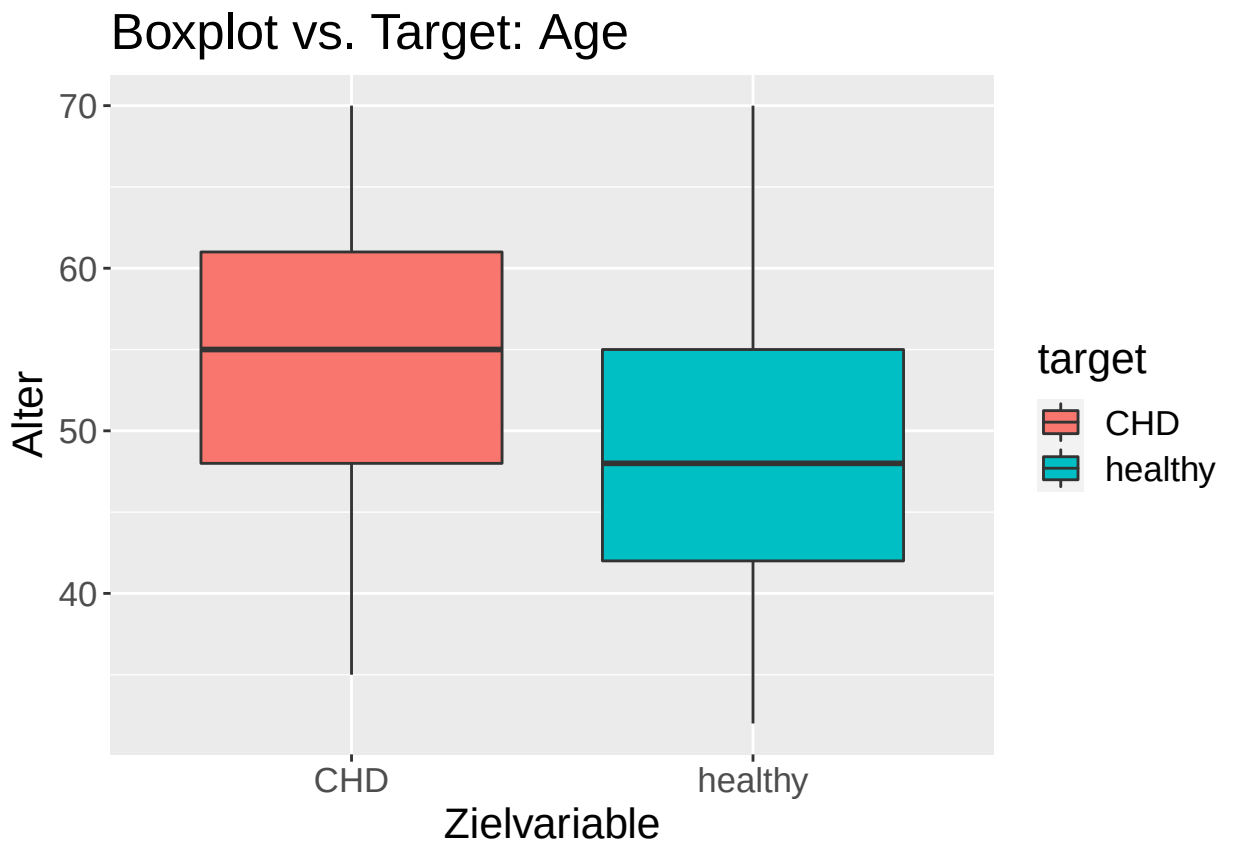
Histogram

```
# By Age and Target
cardio %>%
  ggplot( aes(x = age, fill = target)) +
    geom_histogram(binwidth =2)+
    facet_wrap(~ target) +
    theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: Age vs Target" ) +
          xlab ("Age") +
          ylab ("count")
```



Boxplot

```
cardio %>%
  ggplot( aes(x = target, y= age, fill = target)) +
    geom_boxplot()+
    theme(text = element_text(size=16)) +
    labs ( title = "Boxplot vs. Target: Age") +
    xlab ("Zielvariable") +
    ylab ("Alter")
```




```
age_pvalue = t.test(cardio_chd$age, cardio_healthy$age)$p.value
```

Interpretation:

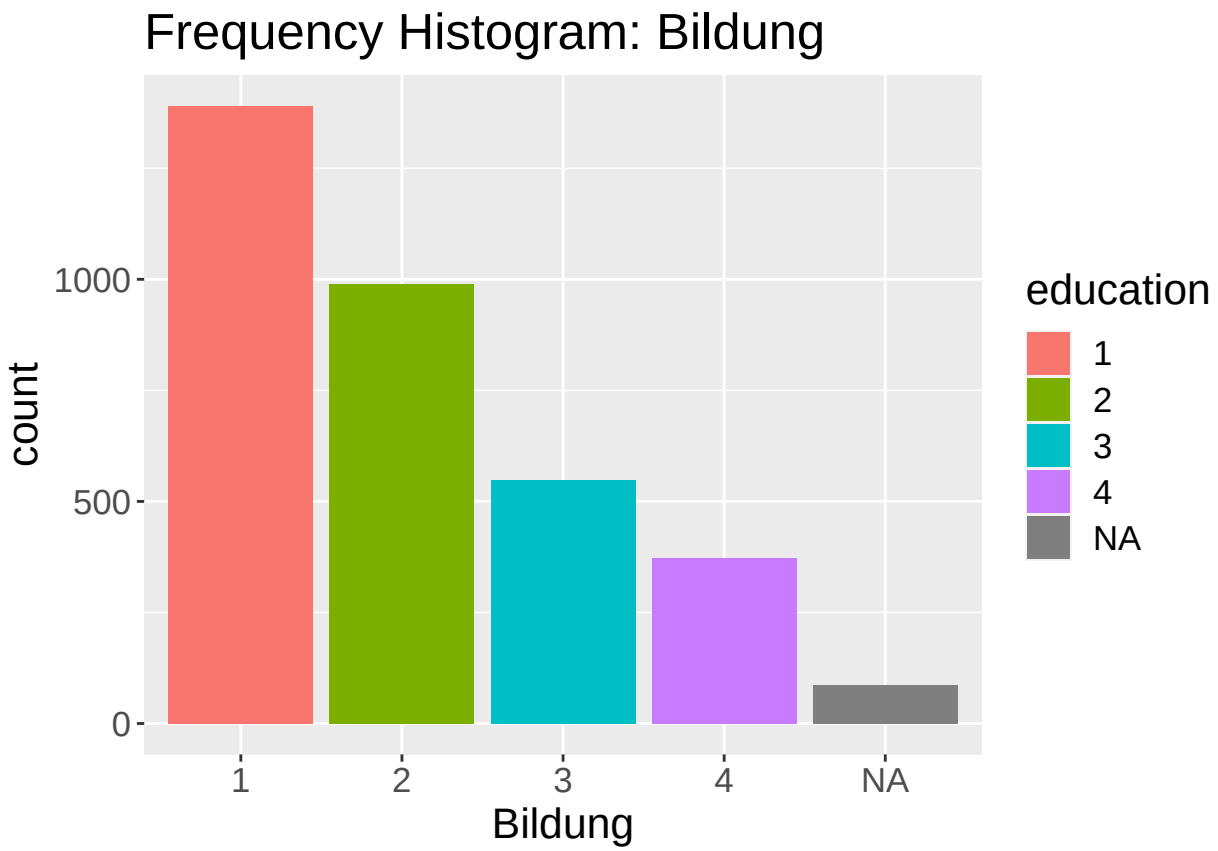
Das Alter scheint einen Einfluss auf den Gesundheitszustand zu haben. Ältere Patienten (ca. 50 bis 60 Jahre) sind im Vergleich zu jüngeren Patienten (< 50 Jahre) häufiger betroffen.

Das Alter hat jedoch nur eine begrenzte Aussagekraft, da sich die beiden Verteilungen im Bereich zwischen 45 und 55 Jahren für Gesunde und Kranke eindeutig überlappen.

Das Alter ist statistisch stark signifikant. Der p-value liegt bei 1.84555411177536e-38.

Bildungsgrad

```
# By education
cardio %>%
  ggplot( aes(x = education ,fill=education)) +
    geom_bar()+
  theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: Bildung " )  +
          xlab ("Bildung") +
          ylab  ("count")
```




```
education_level1 =round(table(cardio$education)["1"]/nrow(cardio) *100,1)
education_level2 =round(table(cardio$education)["2"]/nrow(cardio) *100,1)
education_level4 =round(table(cardio$education)["4"]/nrow(cardio) *100,1)
```

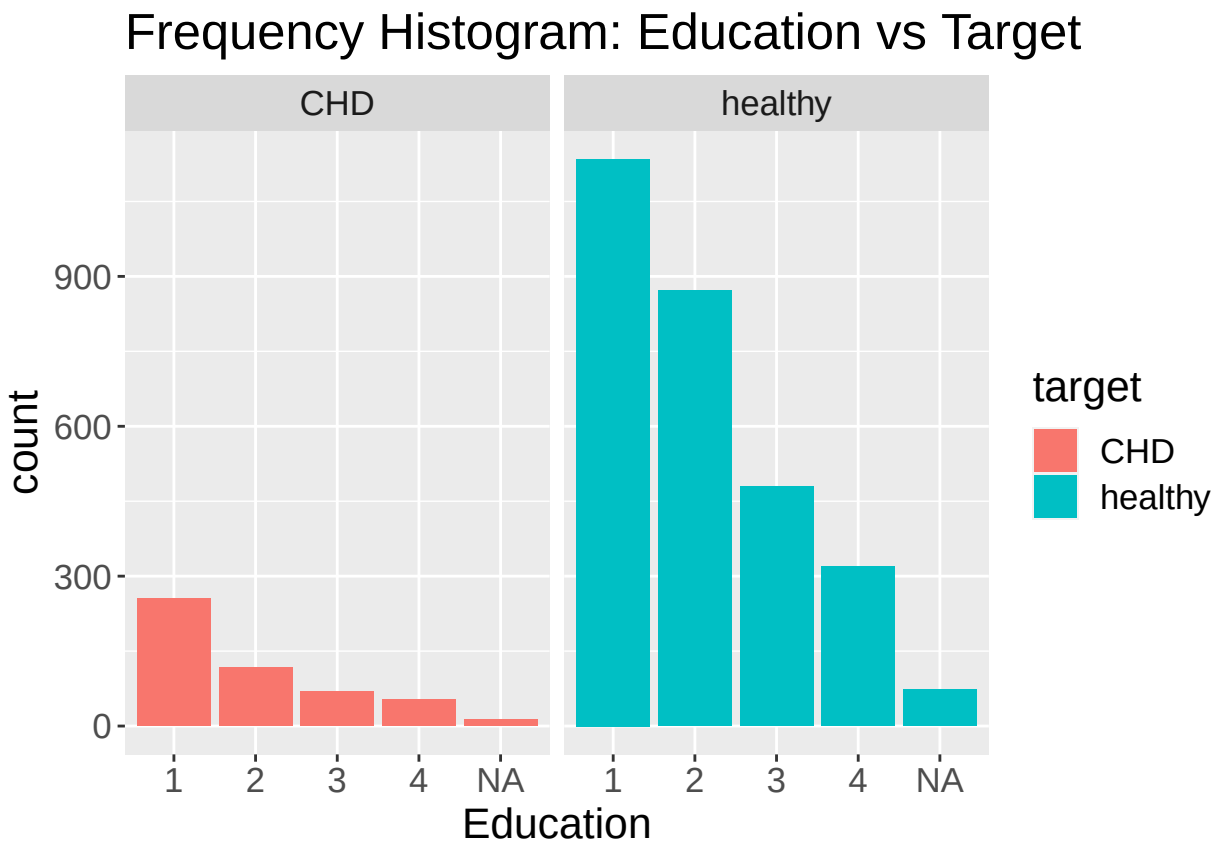
Interpretation:

Wie die Tabelle zeigt, 41% der Teilnehmer haben Level 1 und 29.2% Level 2.

Level 4 zeigt die Personen mit höherer Bildung und der Anteil beträgt 11%.

Frequency Histogram

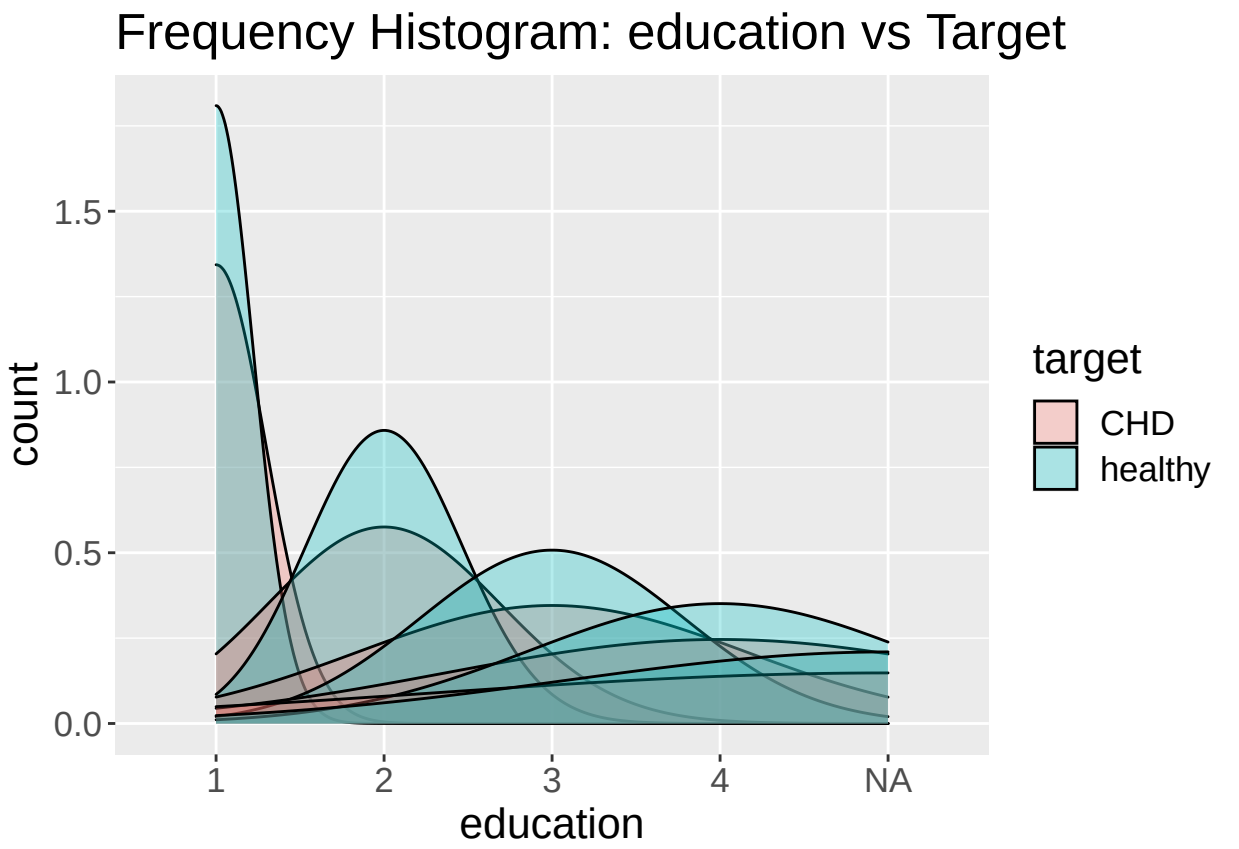
```
# By Education  and Target
cardio %>%
  ggplot( aes(x = education, fill = target)) +
    geom_bar()+
    facet_wrap(~ target) +
    theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: Education vs Target" ) +
          xlab ("Education") +
          ylab ("count")
```




```
education_p_value <- fisher.test(table (cardio$target, cardio$education))$p.value
```

Density plot

```
cardio %>%
  ggplot( aes(x = education, fill = target)) +
    geom_density(alpha = 0.3)+
    theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: education vs Target" ) +
          xlab ("education") +
          ylab ("count")
```



Interpretation:

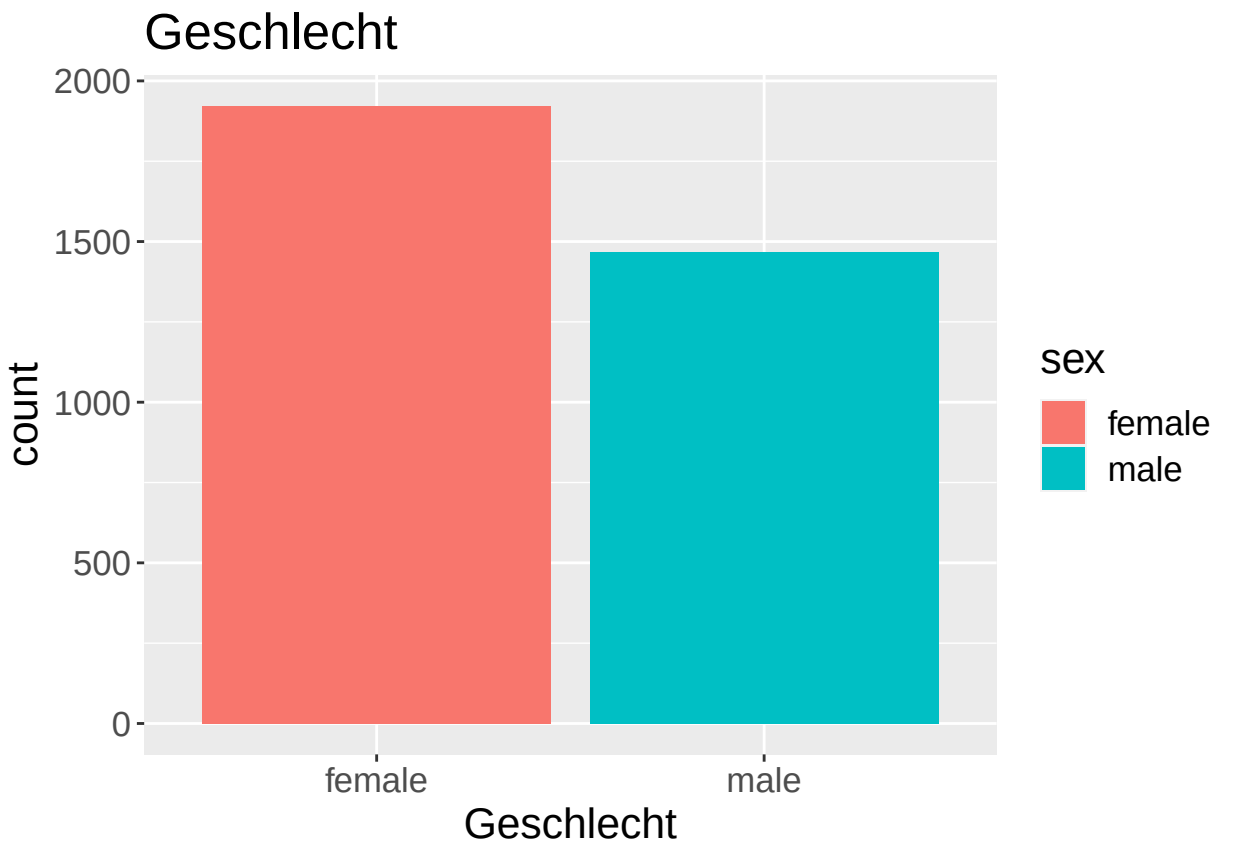
```
table(cardio$target, cardio$education)
```

```
##
##              1    2    3    4
##   CHD      256  118   70   54
##   healthy 1135  872  479  319
```

18,40 % der Telnehmer in Level 1 , 11,91 % der Telnehmer in Level 2, 12,75 % der Telnehmer in Level 3 und 14,47 % der Telnehmer in Level 4 sind an Chd erkrankt . Das heißt die Telnehmer in Level 4 , die Personen mit höherer Bildung ,haben mehr Chance um diese Krankheit zu bekommen.

Geschlecht

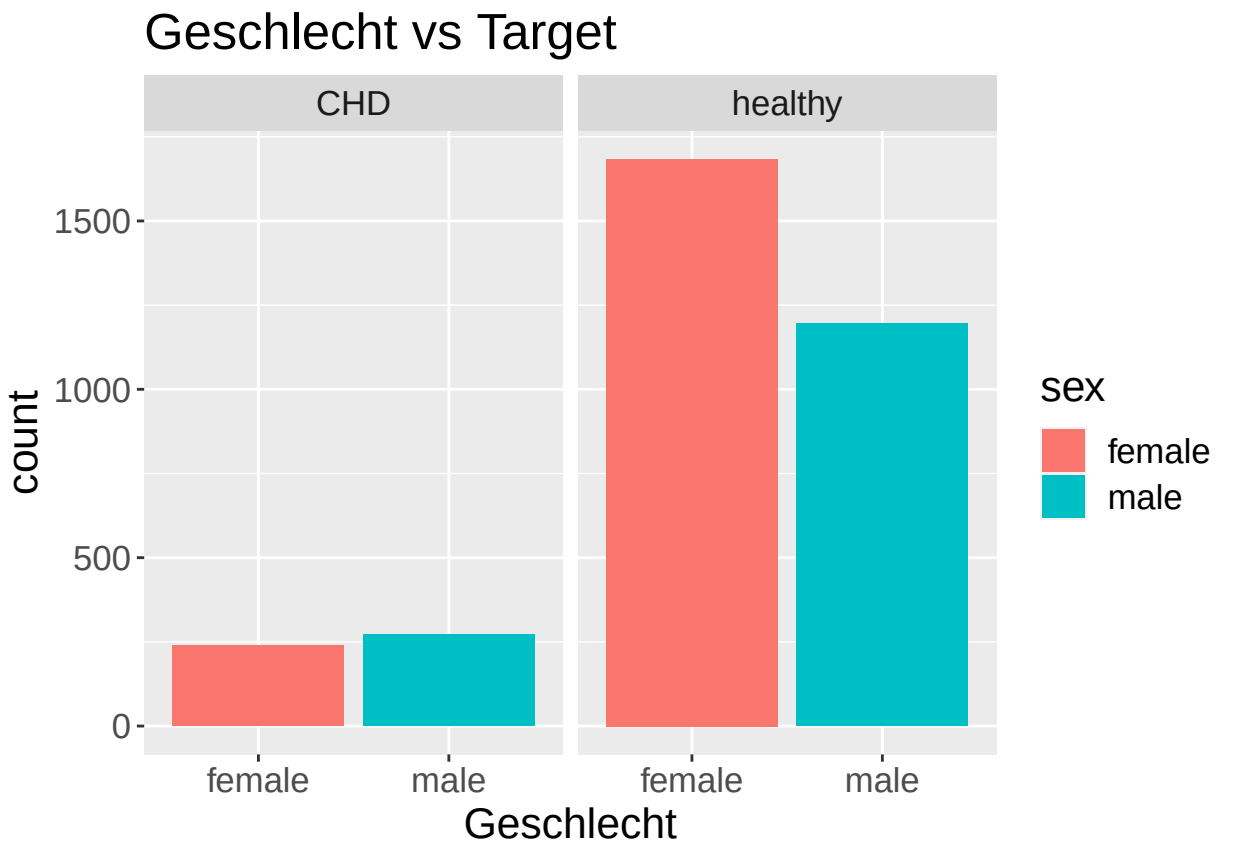
```
# By sex
cardio %>%
  ggplot( aes(x = sex, fill = sex)) +
    geom_bar()+
  theme(text = element_text(size=16)) +
    labs ( title = "Geschlecht" )  +
          xlab ("Geschlecht") +
          ylab  ("count")
```




```
t_femail_count =round(table(cardio$sex)["female"]/nrow(cardio) *100,1)
```

Interpretation: Es sind 56.7 % der Patienten weiblich.

```
# By sex and target
cardio %>%
  ggplot( aes(x = sex, fill = sex)) +
    geom_bar()+
    facet_wrap(~target) +
    theme(text = element_text(size=16)) +
    labs ( title = "Geschlecht vs Target" )  +
          xlab ("Geschlecht") +
          ylab  ("count")
```




```
t_femail_chd_count =round(table(cardio_chd$sex )["female"]/nrow(cardio_chd) *100,1)


t_femail_healthy_count =round(table(cardio_healthy$sex )["female"]/nrow(cardio_healthy) *100,1)


sex_pvalue = fisher.test(table(cardio$target, cardio$sex))$p.value
```

Interpretation:

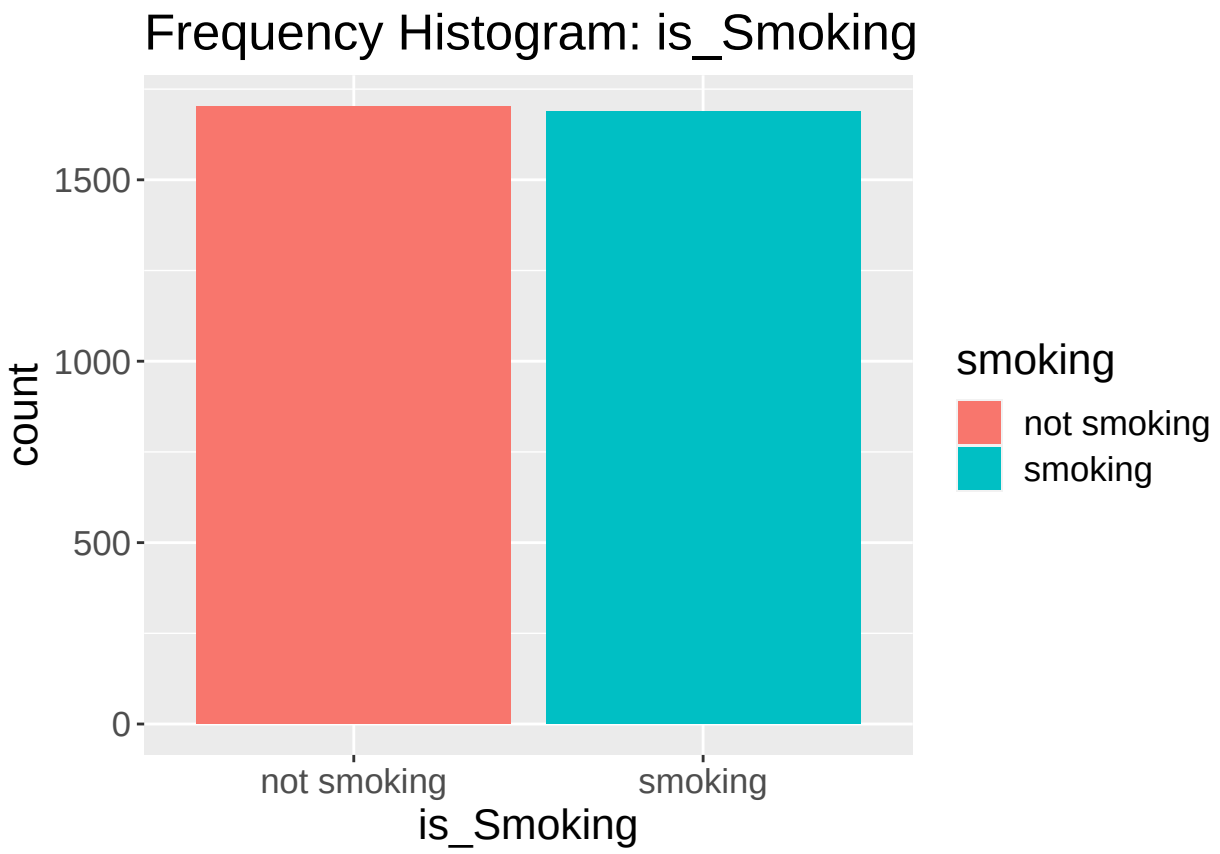
Männer haben ein erhöhtes Risiko die Erkrankung zu entwickeln. Obwohl Frauen in dem Datensatz mit 56.7 % vorkommen liegt der Anteil der Frauen bei den erkrankten bei nur 46.8 % und bei den gesunden bei 58.5 %.

Der Effekt des Geschlechtes ist statistisch signifikant. Der P-Value liegt bei 9.50443889414983e-07.

Smoking

```
# By smoking
cardio %>%
  ggplot( aes(x = smoking,fill=smoking)) +
    geom_bar()+

  theme(text = element_text(size=16)) +
```

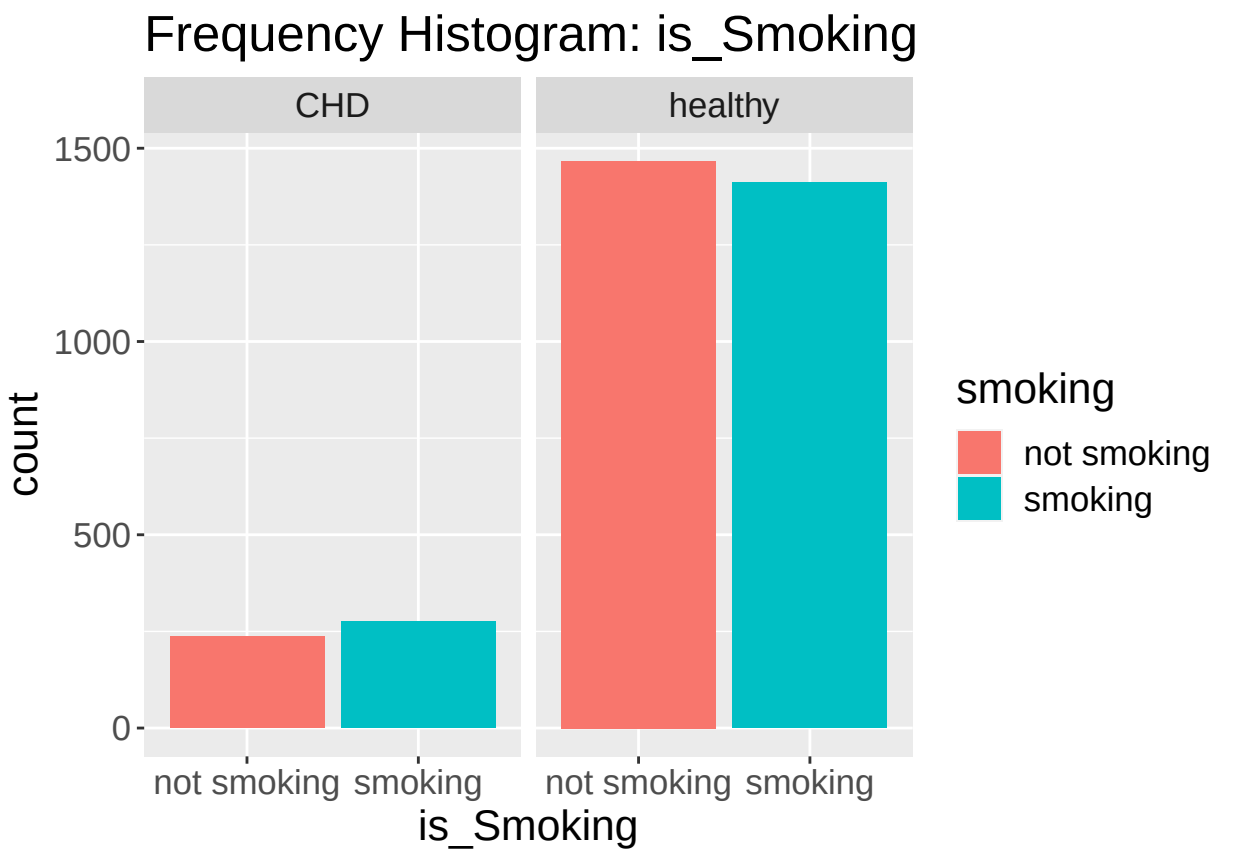
```
    labs ( title = "Frequency Histogram: is_Smoking" )  +
          xlab ("is_Smoking") +
          ylab  ("count")
```



Interpretation:

Der Anteil von Rauchern und nicht Rauchern ist fast gleich, 49% der Teilnehmer rauchen und 51% der Teilnehmer rauchen nicht.

```
cardio %>%
  ggplot( aes(x = smoking,fill=smoking)) +
    geom_bar()+
  facet_wrap(~target) +
  theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: is_Smoking" )  +
          xlab ("is_Smoking") +
          ylab  ("count")
```




```
t_smoking_chd_count =round(table(cardio_chd$smoking )["smoking"]/nrow(cardio_chd) *100,1)
```

```
t_smoking_healthy_count =round(table(cardio_healthy$smoking )["smoking"]/nrow(cardio_healthy) *100,1)
```

Statistischer Test:

```
table(cardio$target, cardio$smoking)
```

```
##
##           not smoking smoking
##   CHD              236     275
##   healthy         1467    1412
```

```
smoking_pvalue=fisher.test(table(cardio$target, cardio$smoking))$p.value
```

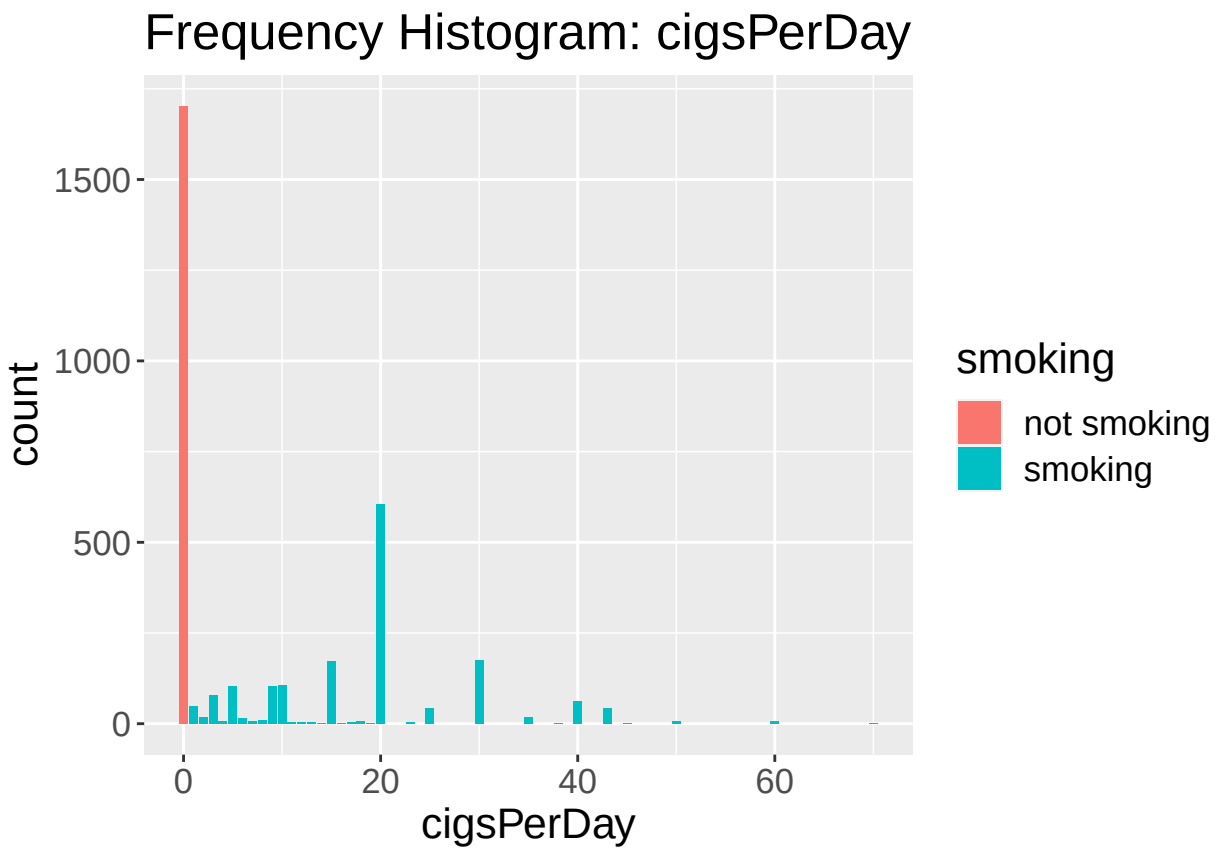
Interpretation:

Der Anteil von Rauchern ist bei den Gesunden 49% und bei den Kranken 53.8% Die Raucher haben ein leichtes Risiko die Erkrankung zu entwickeln.

Cigs Per Day

```
# By Cigs Per Day
cardio %>%
  ggplot( aes(x =cigsPerDay ,fill=smoking))  +
    geom_bar()+

  theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: cigsPerDay" )  +
          xlab ("cigsPerDay") +
          ylab  ("count")
```

```
## Warning: Removed 22 rows containing non-finite values (stat_count).
```



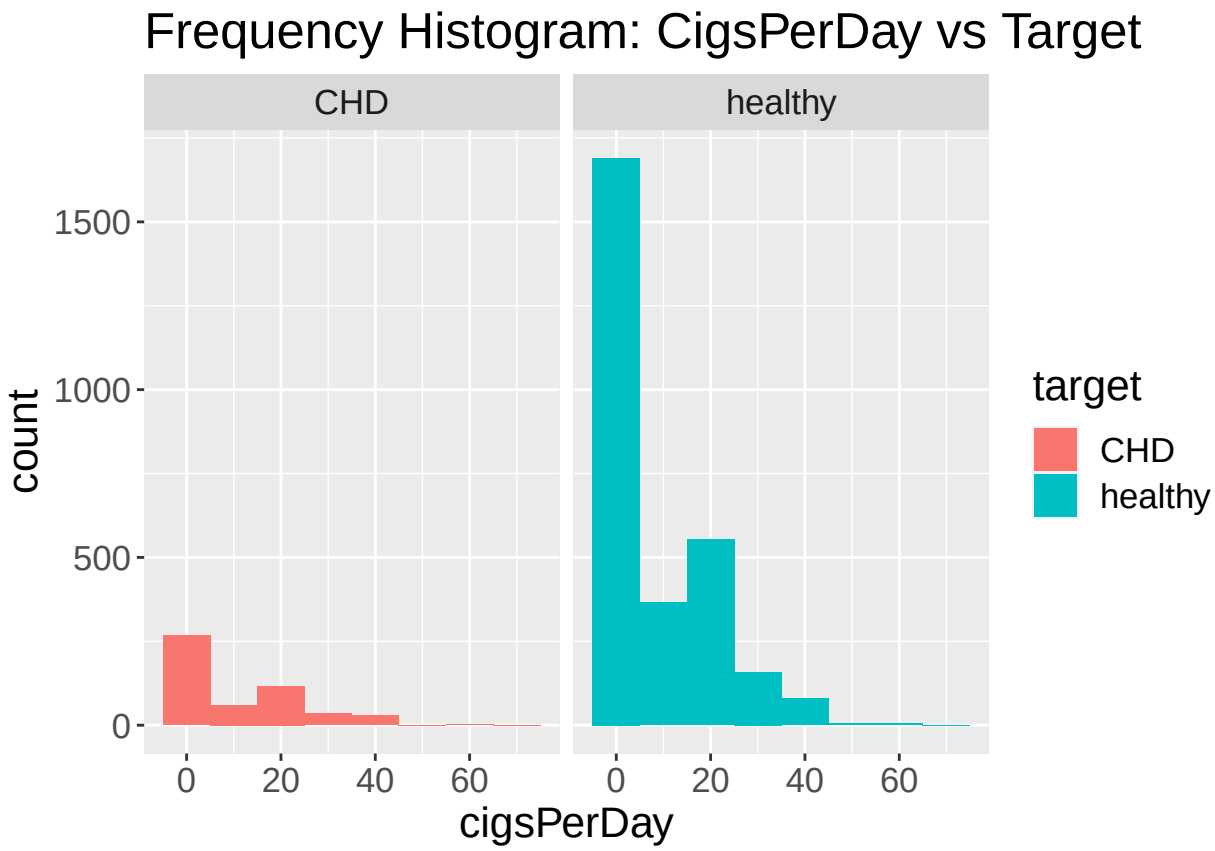
Interpretation:

" Cigs Per Day" zeigt : Anzahl der Zigaretten, die die Person im Durchschnitt an einem Tag geraucht hat. Die Mehrheit der Raucher raucht 20 Zigaretten pro Tag.

```
cardio %>%
  ggplot( aes(x = cigsPerDay, fill = target)) +
    geom_histogram(binwidth =   10)+
    facet_wrap(~ target) +
    theme(text = element_text(size=16)) +
```

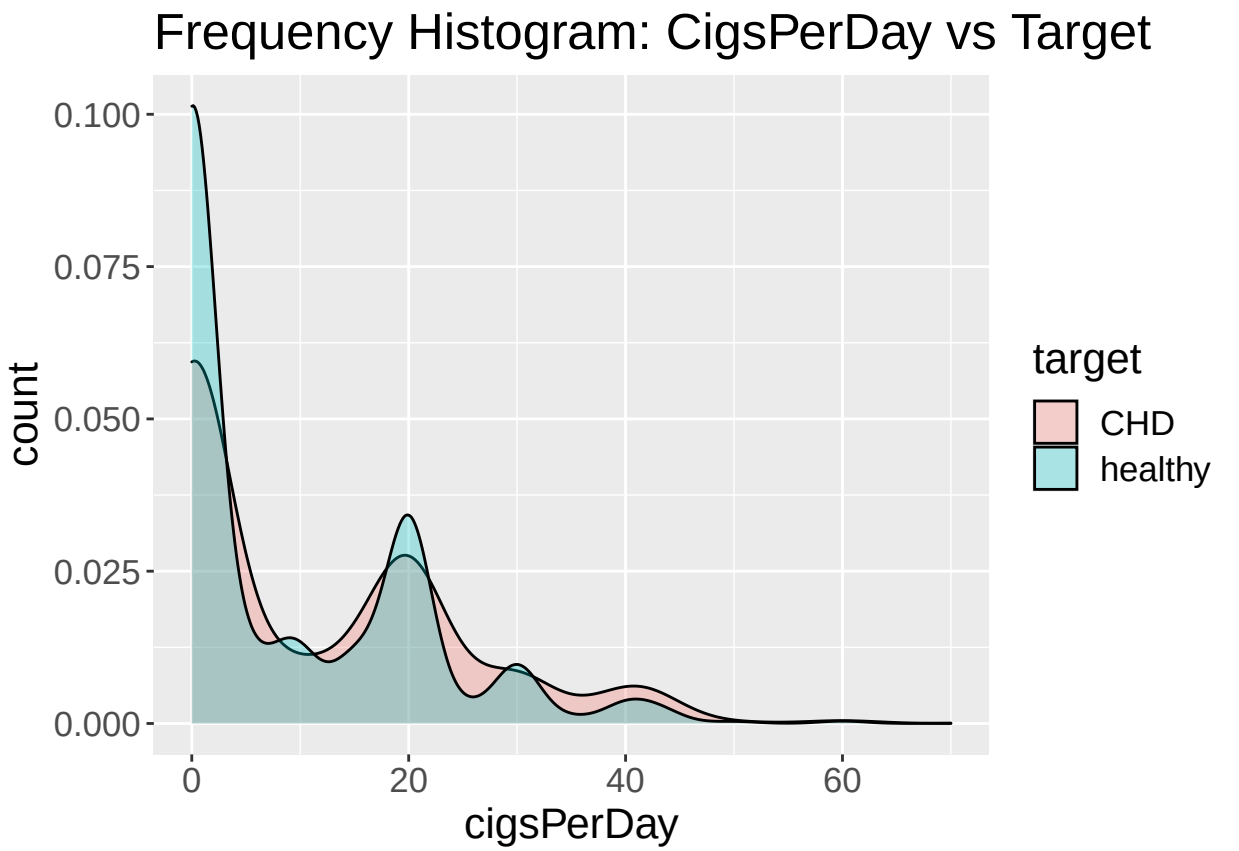
```
    labs ( title = "Frequency Histogram: CigsPerDay vs Target" ) +
          xlab ("cigsPerDay") +
          ylab ("count")
```

```
## Warning: Removed 22 rows containing non-finite values (stat_bin).
```




```
cardio %>%
  ggplot( aes(x = cigsPerDay, fill = target)) +
    geom_density(alpha = 0.3)+
    #facet_wrap(~ target) +
    theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: CigsPerDay vs Target" ) +
          xlab ("cigsPerDay") +
          ylab ("count")
```

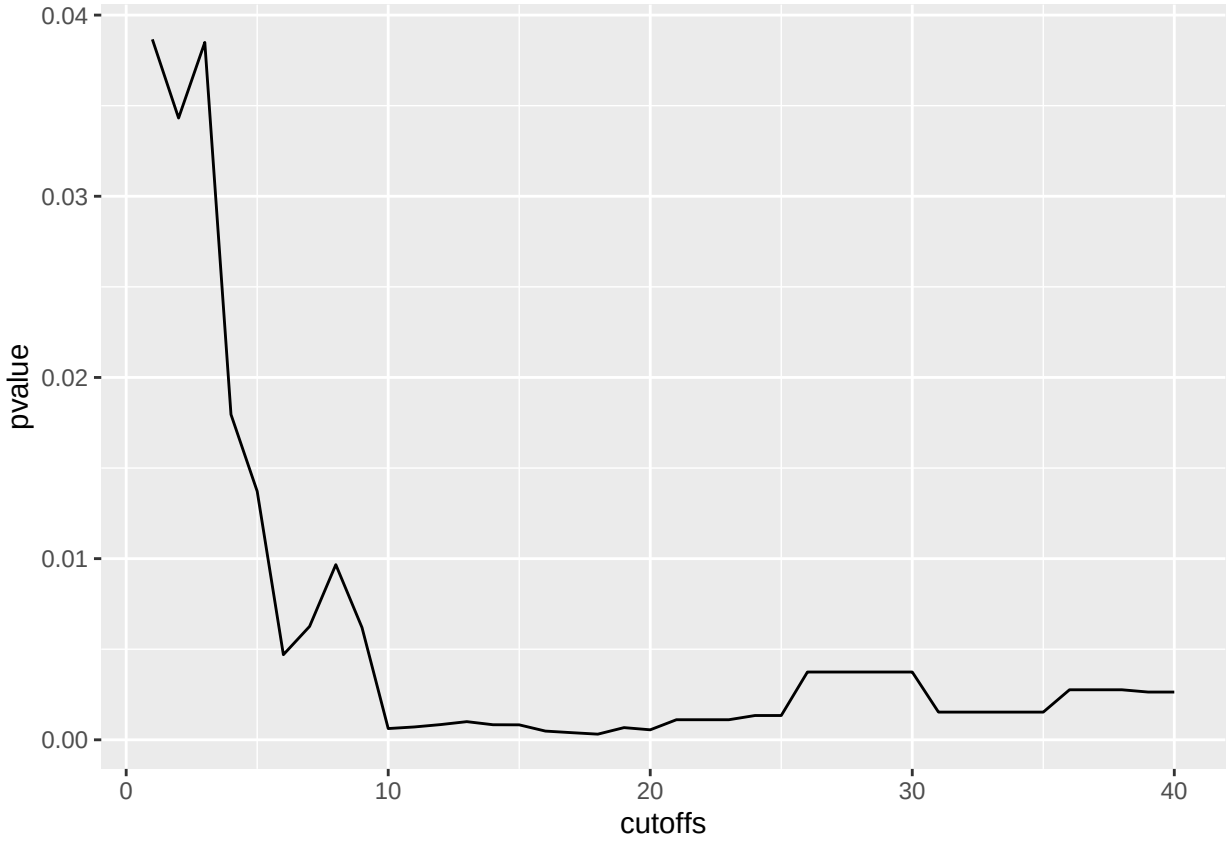
```
## Warning: Removed 22 rows containing non-finite values (stat_density).
```




```
cigsPerDay_pvalue = t.test(cardio_chd$cigsPerDay, cardio_healthy$cigsPerDay)$p.value
```

Die Anzahl an Zigaretten pro Tag hat einen Effekt auf das Risiko zu erkranken. Es wird die Anzahl an Zigaretten gesucht, welche den Effekt auf das Krankheitsrisiko maximiert.

```
#fisher.test(table(cardio$target, cardio$cigsPerDay >= 10 ))$p.value

cutoffs <- seq(1,40)
cigsPerDay_pvalues = rep (NA, length(cutoffs))
for (i in 1:length(cutoffs)) {
   cigsPerDay_pvalues[i]= fisher.test(table(cardio$target, cardio$cigsPerDay >= cutoffs[i] ))$p.value
}

data.frame(cutoffs = cutoffs, pvalue = cigsPerDay_pvalues ) %>%
  ggplot(aes (cutoffs,pvalue )) +
  geom_line()
```




```
min_idx <- which.min (cigsPerDay_pvalues)

print (paste("Bei einem Cutoff von <= " ,cutoffs[min_idx] , " ergibt sich der särkste statistische Effe
```

```
## [1] "Bei einem Cutoff von <=  18  ergibt sich der särkste statistische Effekt. Hier ergibt sich ein
```

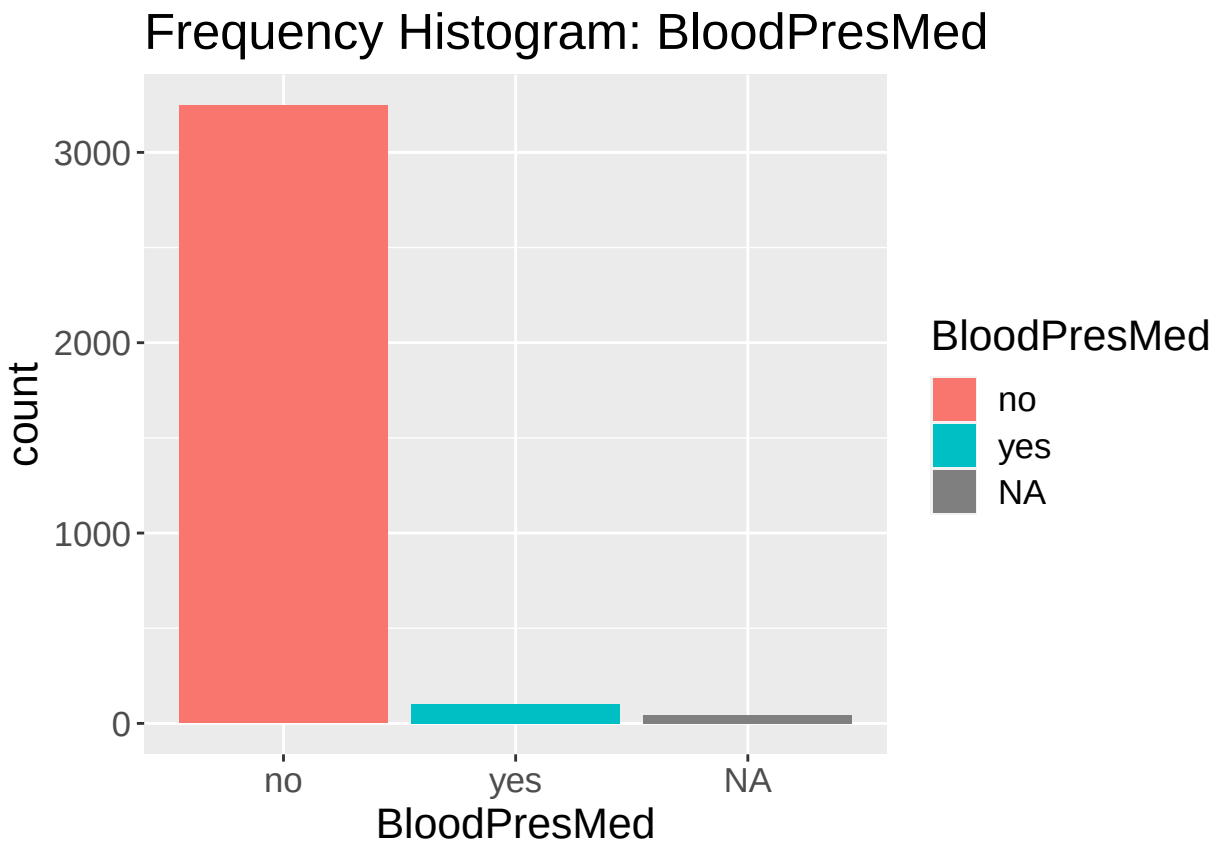
```
cardio[1:10, c("target","smoking","cigsPerDay")]
```

```
##     target     smoking cigsPerDay
## 1      CHD     smoking          3
## 2  healthy not smoking          0
## 3  healthy     smoking         10
## 4      CHD     smoking         20
## 5  healthy     smoking         30
## 6      CHD not smoking          0
## 7  healthy not smoking          0
## 8  healthy     smoking         35
## 9  healthy     smoking         20
## 10 healthy not smoking          0
```

Interpretation: Die Personen, die weniger als 18 Zigaretten pro Tag rauchen, haben ein geringes Risiko, an CHD zu erkranken.

BloodPresMed

```
# BloodPresMed
cardio %>%
  ggplot( aes(x = BloodPresMed,fill=BloodPresMed)) +
    geom_bar()+

  theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: BloodPresMed" )  +
          xlab ("BloodPresMed") +
          ylab  ("count")
```



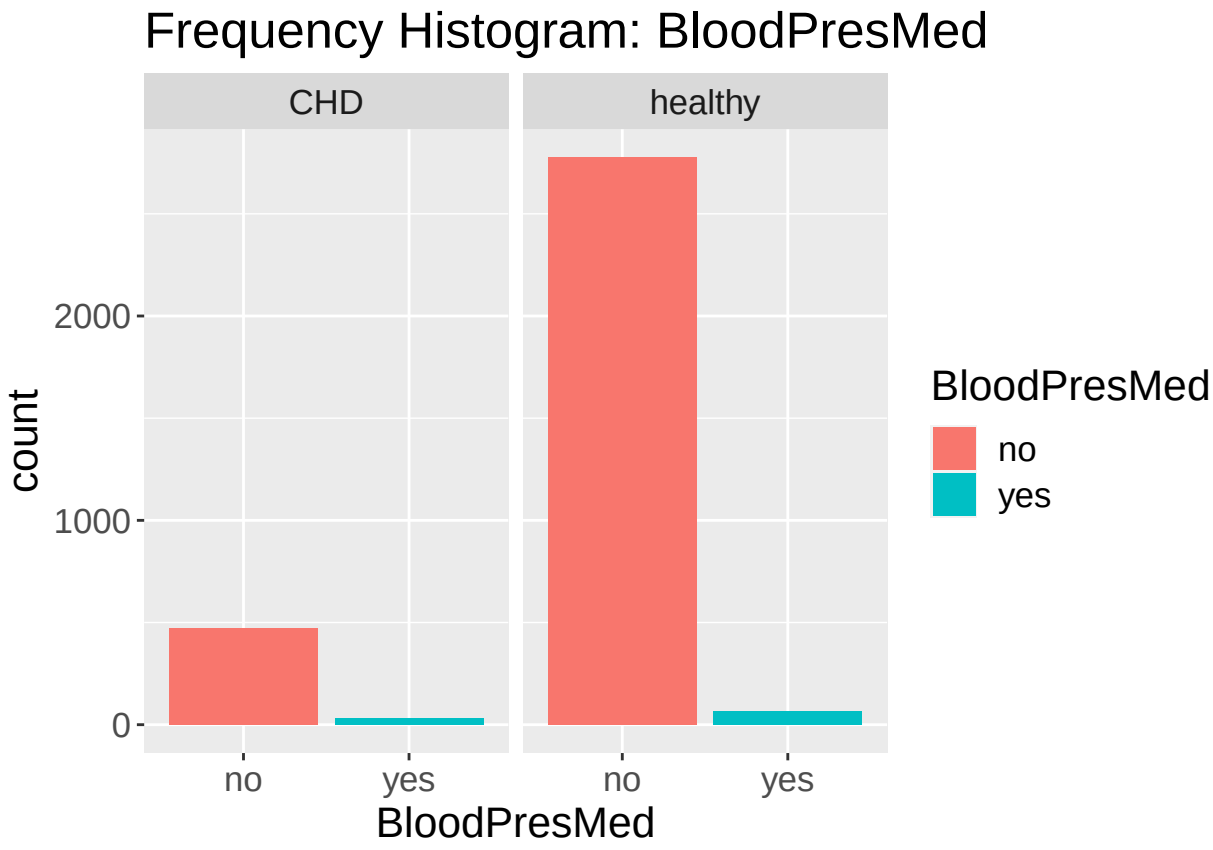

```
t_keinBPmed_count =round(table(cardio$BloodPresMed)["no"]/nrow(cardio) *100,1)
```

Interpretation:

95.8der Teilnehmer nehmen keine Blutdruckmedikamente ein.

```
subset(cardio, !is.na(BloodPresMed)) %>%
  ggplot( aes(x = BloodPresMed,fill=BloodPresMed)) +
    geom_bar()+
  facet_wrap(~target) +
  theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: BloodPresMed" )  +
```

```
          xlab ("BloodPresMed") +
          ylab  ("count")
```




```
t_keinBPmed_chd_count =round(table(cardio_chd$BloodPresMed )["no"]/nrow(cardio_chd) *100,1)


t_keinBPmed_healthy_count =round(table(cardio_healthy$BloodPresMed )["no"]/nrow(cardio_healthy) *100,1)


BloodPresMed_pvalue=fisher.test(table(cardio$target, cardio$BloodPresMed))$p.value
```

Interpretation:

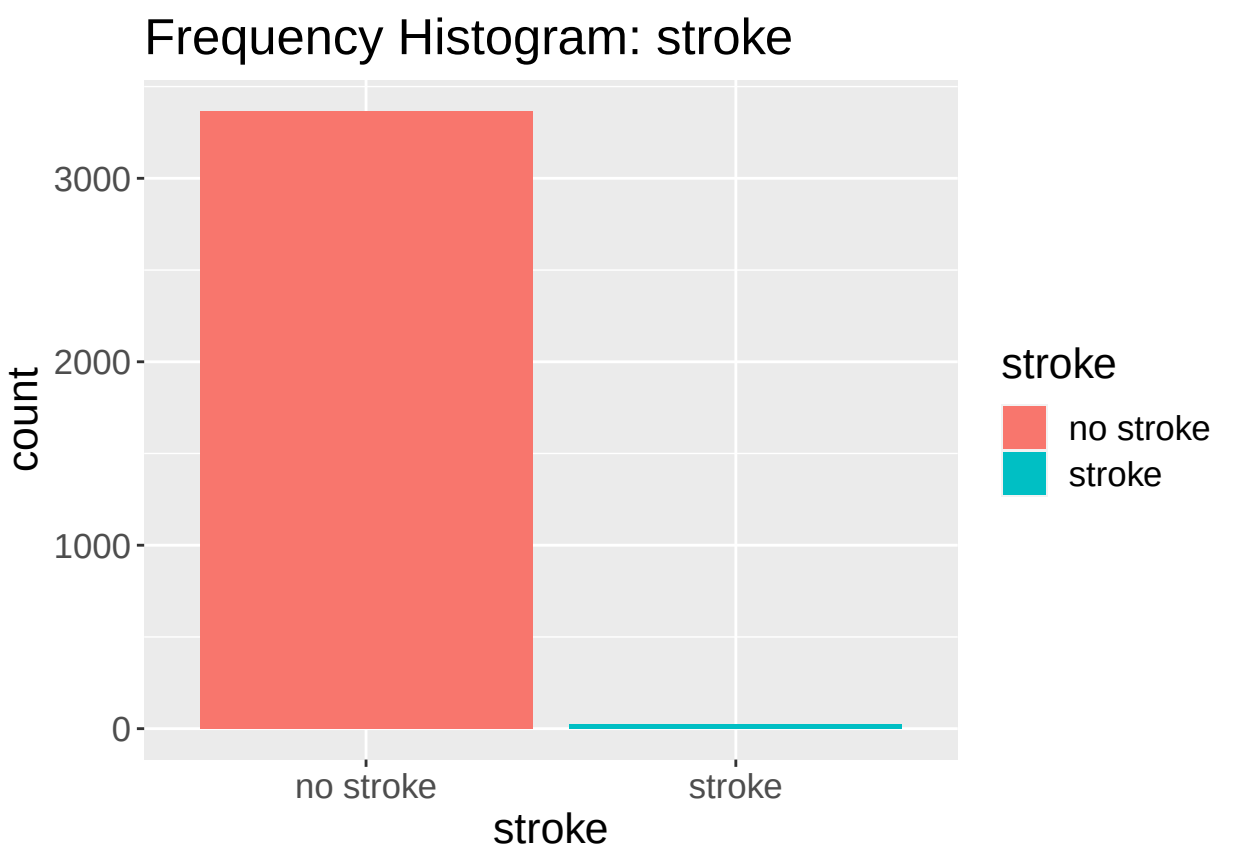
Der Anteil von der Teilnehmer, die keine Blutdruckmedikamente einnehemn, ist bei den Gesunden 92.2% und bei den Kranken 96.4%

Der Bluthochdruck hat einen statistisch signifikaten Effekt. Der P-value liegt bei 5.234109654326645e-06.

stroke

```
# stroke
cardio %>%
  ggplot( aes(x = stroke,fill=stroke)) +
```

```
    geom_bar()+

  theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: stroke" )  +
          xlab ("stroke") +
          ylab  ("count")
```

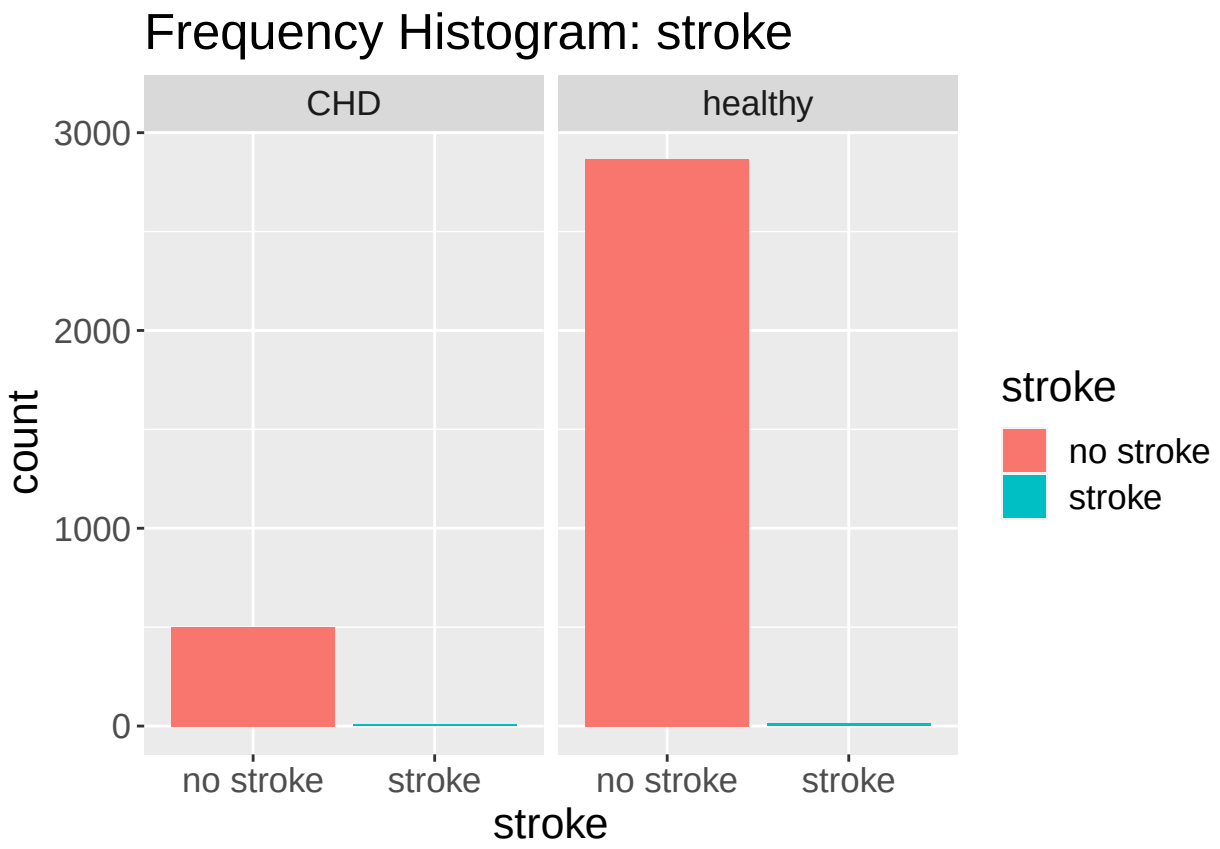



```
t_stroke_count =round(table(cardio$stroke)["stroke"]/nrow(cardio) *100,1)
```

Interpretation:

" stroke" zeigt : ob der Teilnehmer zuvor einen Schlaganfall hatte oder nicht . 0.6 % der Teilnehmer hatten zuvor einen Schlaganfall gehabt.

```
cardio %>%
  ggplot( aes(x = stroke,fill=stroke)) +
    geom_bar()+
  facet_wrap(~target) +
  theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: stroke" )  +
          xlab ("stroke") +
          ylab  ("count")
```




```
t_stroke_chd_count =round(table(cardio_chd$stroke )["stroke"]/nrow(cardio_chd) *100,1)


t_stroke_healthy_count =round(table(cardio_healthy$stroke )["stroke"]/nrow(cardio_healthy) *100,1)

stroke_pvalue =fisher.test(table(cardio$target, cardio$stroke))$p.value
```

Interpretation:

Der Anteil von der Teilnehmer, die zuvor einen Schlaganfall hatten, ist bei den Gesunden 0.4% und bei den Kranken 0.4%

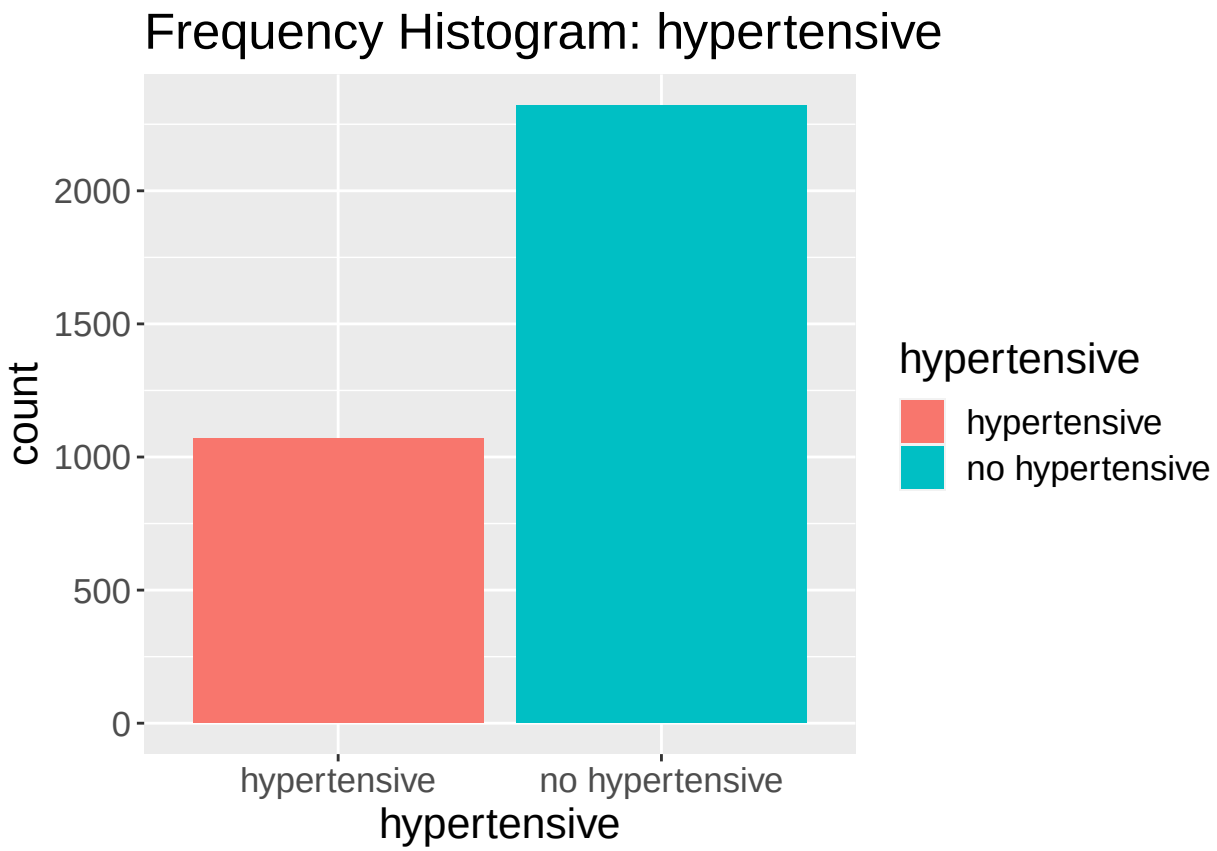
Der Anteil von der Teilnehmer, die zuvor einen Schlaganfall hatten, ist bei den Gesunden und bei den Kranken ist fast gleich.

Ein vorheriger Schlaganfall hat einen statistisch signifikaten Effekt. Der P-value liegt bei 0.000647033881655412.

hypertensive

```
# hypertensive
cardio %>%
  ggplot( aes(x = hypertensive,fill=hypertensive)) +
    geom_bar()+
```

```
  theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: hypertensive" )  +
          xlab ("hypertensive") +
          ylab  ("count")
```




```
t_hypertensive_count =round(table(cardio$hypertensive)["prevalent hypertensive"]/nrow(cardio) *100,1)
```

Interpretation:

" hypertensive" zeigt : ob der Teilnehmer einen Bluthochdruck hatten oder nicht . NA % der Teilnehmer leiden an Bluthochdruck

```
t_hypertensive_chd_count =round(table(cardio_chd$hypertensive )["hypertensive"]/nrow(cardio_chd) *100,1

t_hypertensive_healthy_count =round(table(cardio_healthy$hypertensive )["hypertensive"]/nrow(cardio_heal

hypertensive_pvalue=fisher.test(table(cardio$target, cardio$hypertensive))$p.value
```

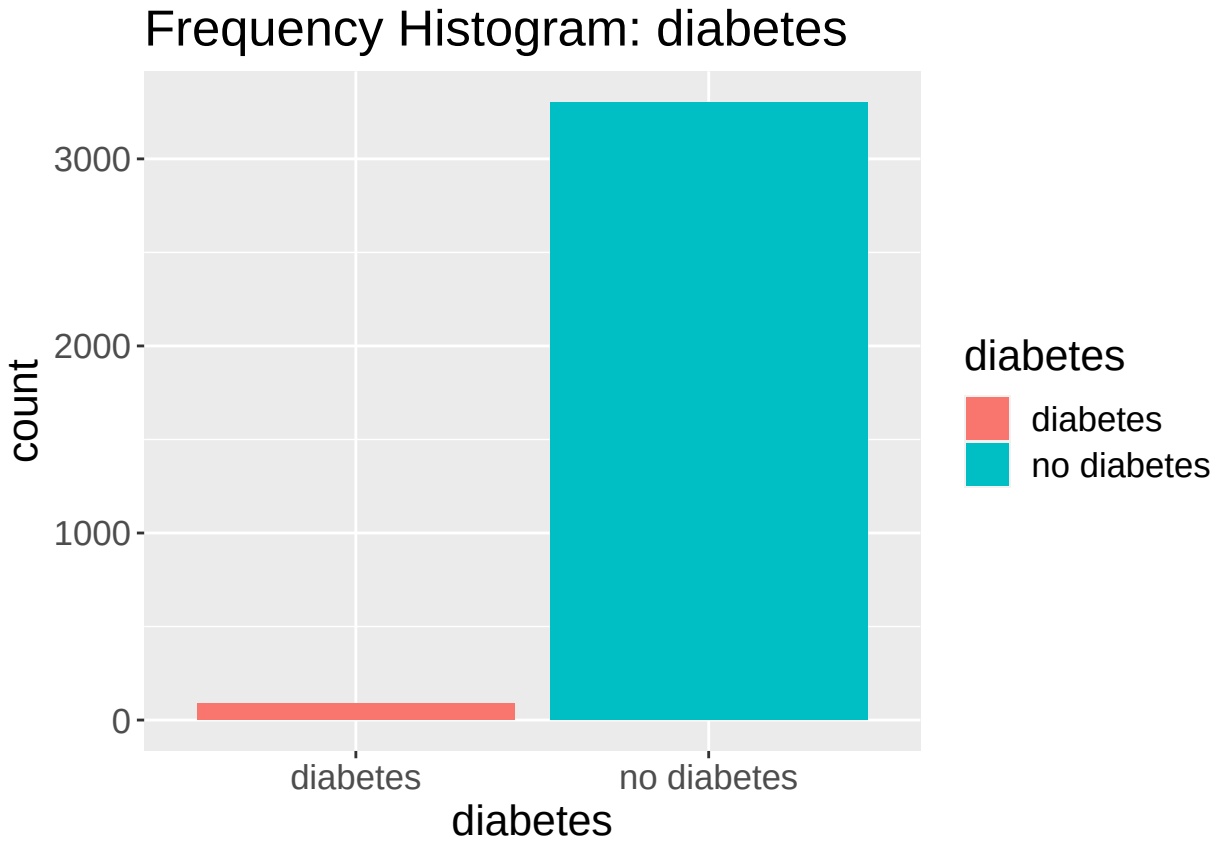
Interpretation:

Der Anteil von der Teilnehmer, die an Bluthochdruck leiden, ist bei den Gesunden 28.3% und bei den Kranken 49.9%

Ein vorhandener Bluthochdruck ist statistisch stark signifikant. Der P-Value liegt bei 4.91775745429724e-21.

diabetes

```
# diabetes
cardio %>%
  ggplot( aes(x = diabetes,fill=diabetes)) +
    geom_bar()+

  theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: diabetes" )  +
          xlab ("diabetes") +
          ylab  ("count")
```




```
t_diabetes_count =round(table(cardio$diabetes)["diabetes"]/nrow(cardio) *100,1)
```

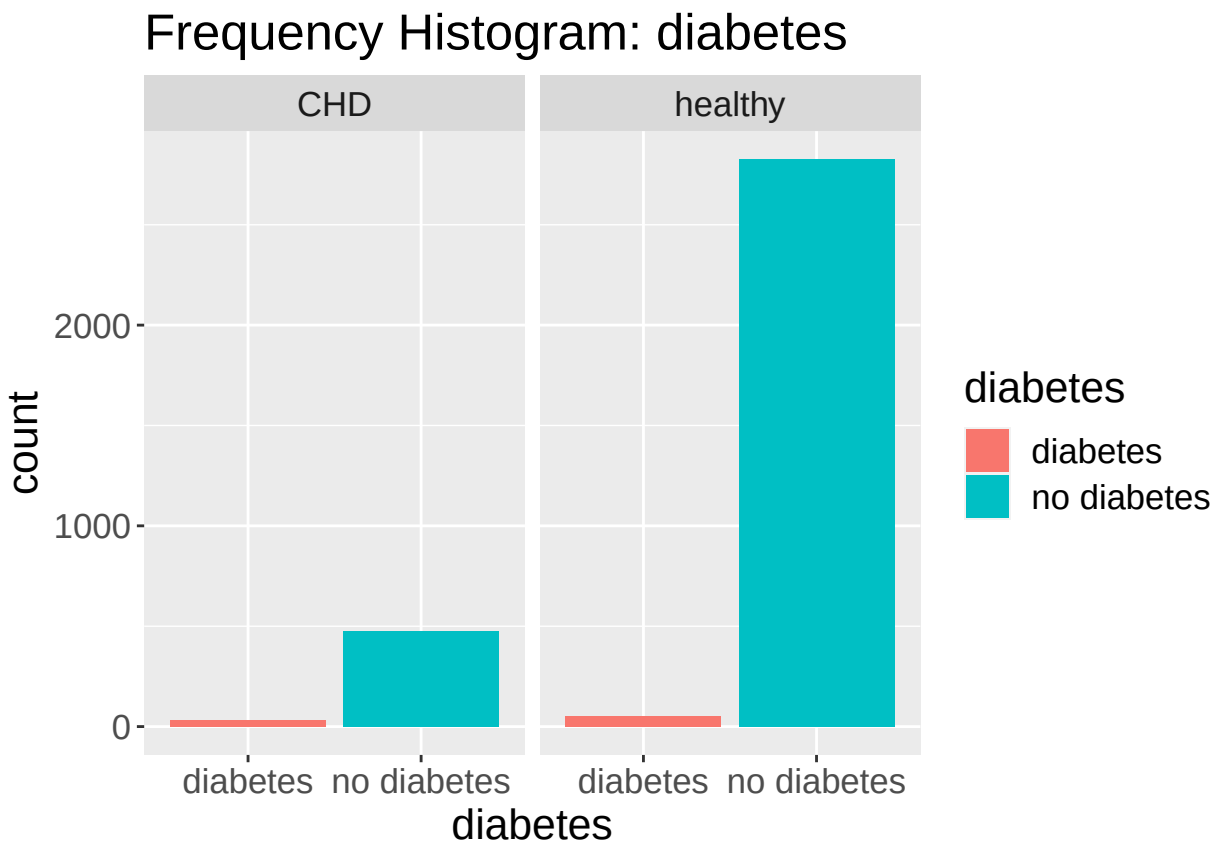
Interpretation:

" diabetes" zeigt : ob der Teilnehmer Diabetes hatten oder nicht .

Nur2.6 % der Teilnehmer leiden an Diabetes.

```
cardio %>%
  ggplot( aes(x = diabetes,fill=diabetes)) +
    geom_bar()+
  facet_wrap(~target) +
  theme(text = element_text(size=16)) +
```

```
    labs ( title = "Frequency Histogram: diabetes" )  +
          xlab ("diabetes") +
          ylab  ("count")
```




```
t_diabetes_healthy_count =round(table(cardio_healthy$diabetes )["diabetes"]/nrow(cardio_healthy) *100,1

t_diabetes_chd_count =round(table(cardio_chd$diabetes )["diabetes"]/nrow(cardio_chd) *100,1)

diabetes_pvalue = fisher.test(table(cardio$target, cardio$diabetes))$p.value
```

Interpretation:

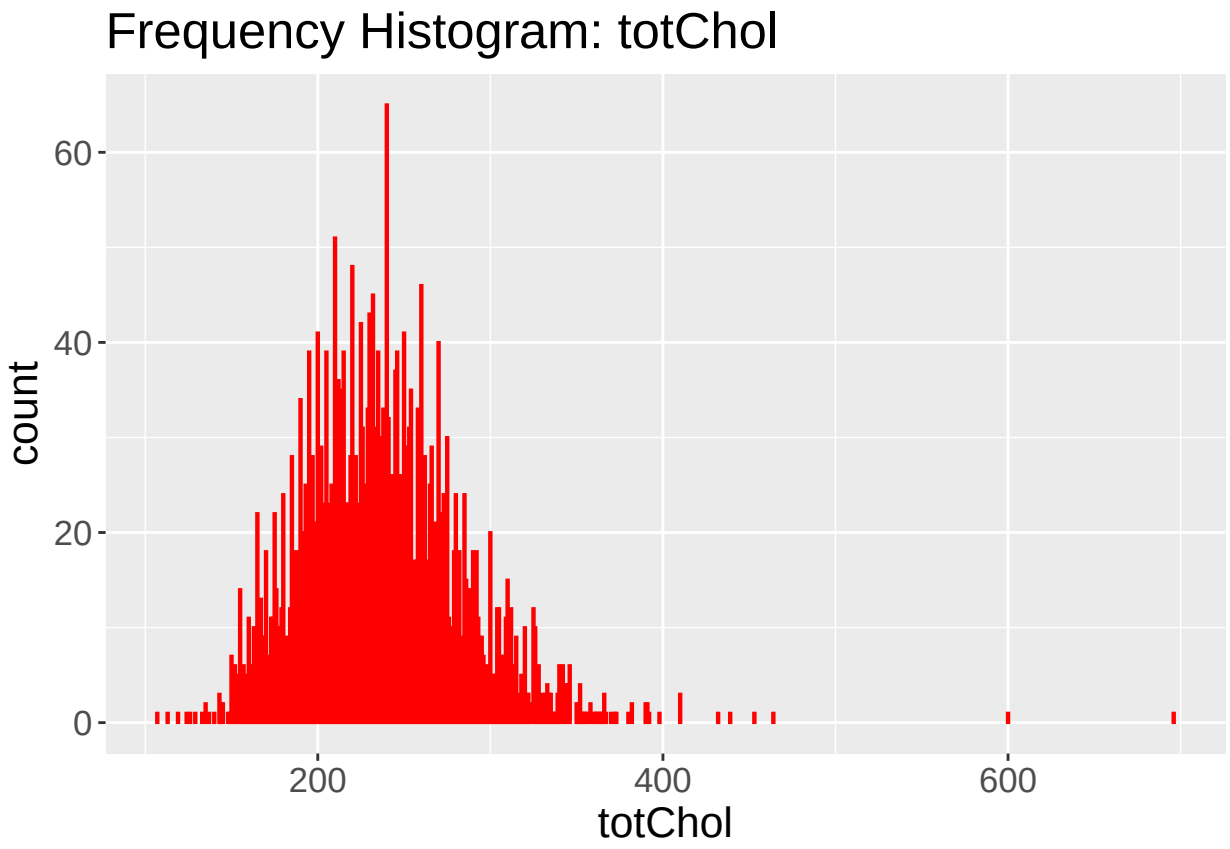
Der Anteil von der Teilnehmer, die an Diabetes leiden, ist bei den Gesunden 1.9% und bei den Kranken 6.5%

totChol

```
cardio %>%
  ggplot( aes(x =totChol ,fill=totChol))  +
    geom_bar(color="red")+

  theme(text = element_text(size=16)) +
```

```
    labs ( title = "Frequency Histogram: totChol" )  +
          xlab ("totChol") +
          ylab  ("count")
```

```
## Warning: Removed 38 rows containing non-finite values (stat_count).
```

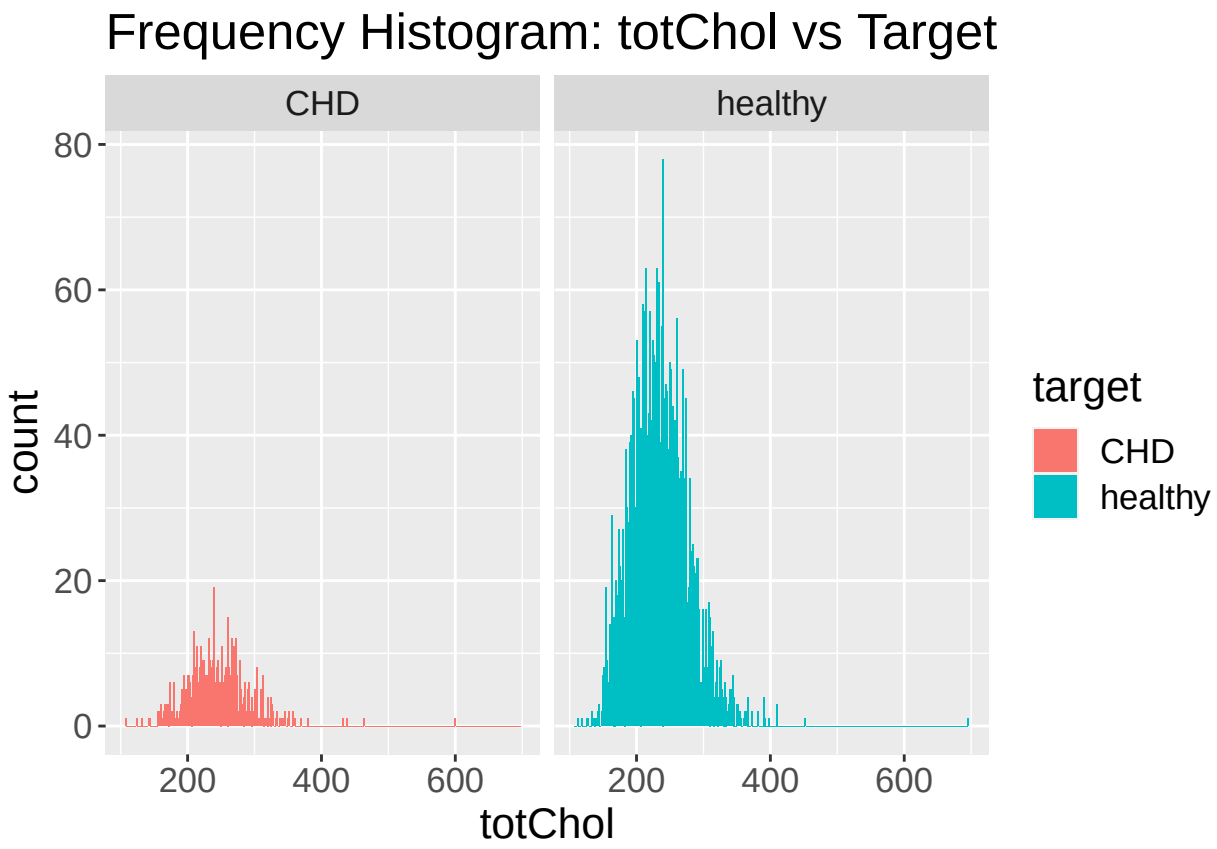


Interpretation:

Der durchschnittliche Gesamtcholesterinspiegel liegt bei etwa 236 .

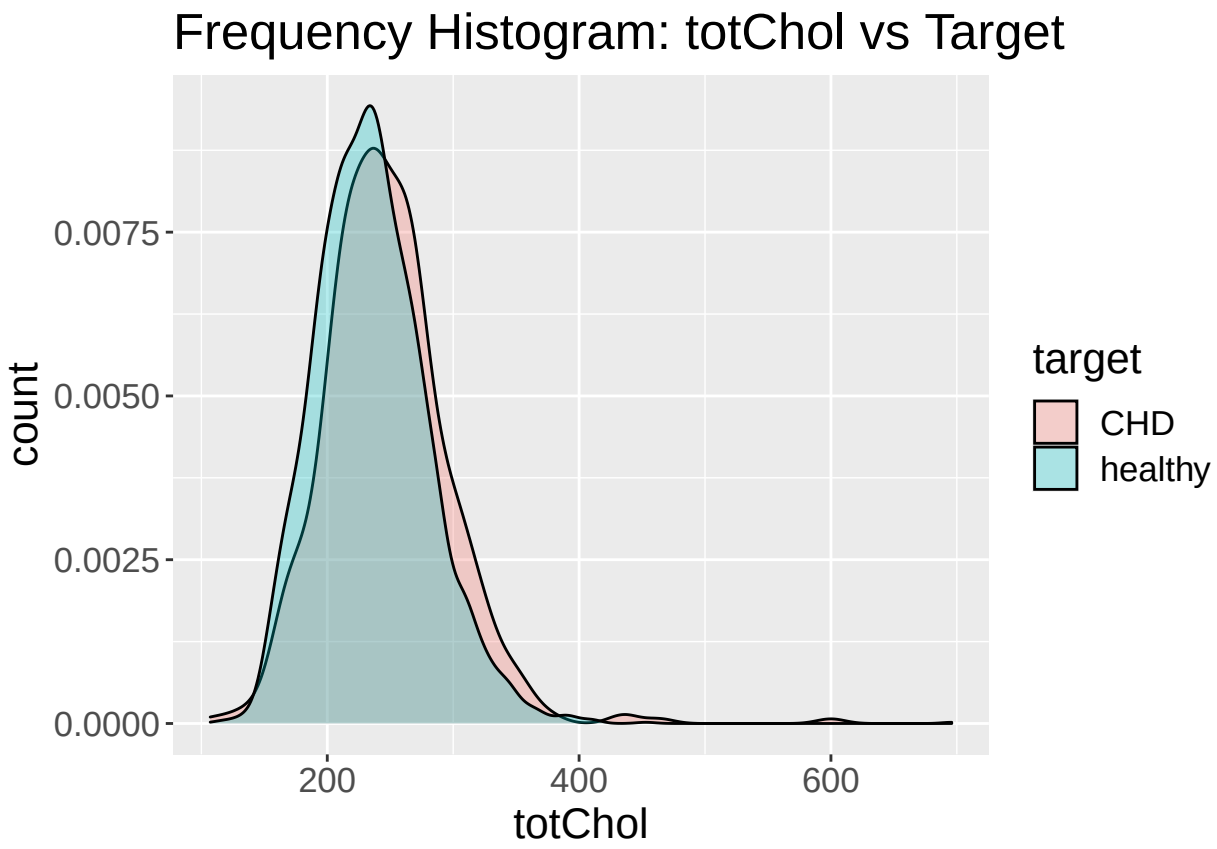
```
cardio %>%
  ggplot( aes(x = totChol, fill = target)) +
    geom_histogram(binwidth =2)+
    facet_wrap(~ target) +
    theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: totChol vs Target" ) +
          xlab ("totChol") +
          ylab ("count")
```

```
## Warning: Removed 38 rows containing non-finite values (stat_bin).
```



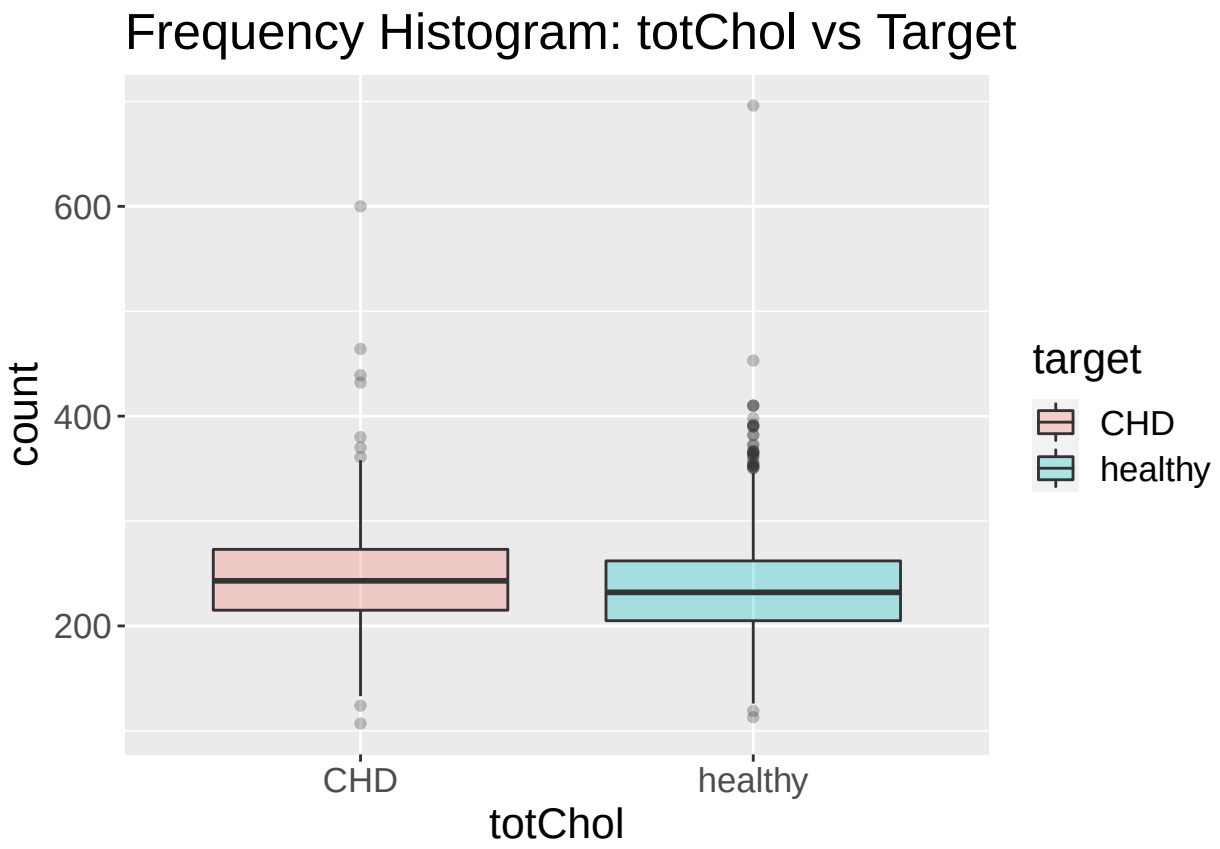

```
cardio %>%
  ggplot( aes(x = totChol, fill = target)) +
    geom_density(alpha = 0.3)+
    theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: totChol vs Target" ) +
          xlab ("totChol") +
          ylab ("count")
```

```
## Warning: Removed 38 rows containing non-finite values (stat_density).
```




```
cardio %>%
  ggplot( aes(x = target ,y= totChol, fill = target)) +
    geom_boxplot(alpha = 0.3)+
    theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: totChol vs Target" ) +
          xlab ("totChol") +
          ylab ("count")
```

```
## Warning: Removed 38 rows containing non-finite values (stat_boxplot).
```



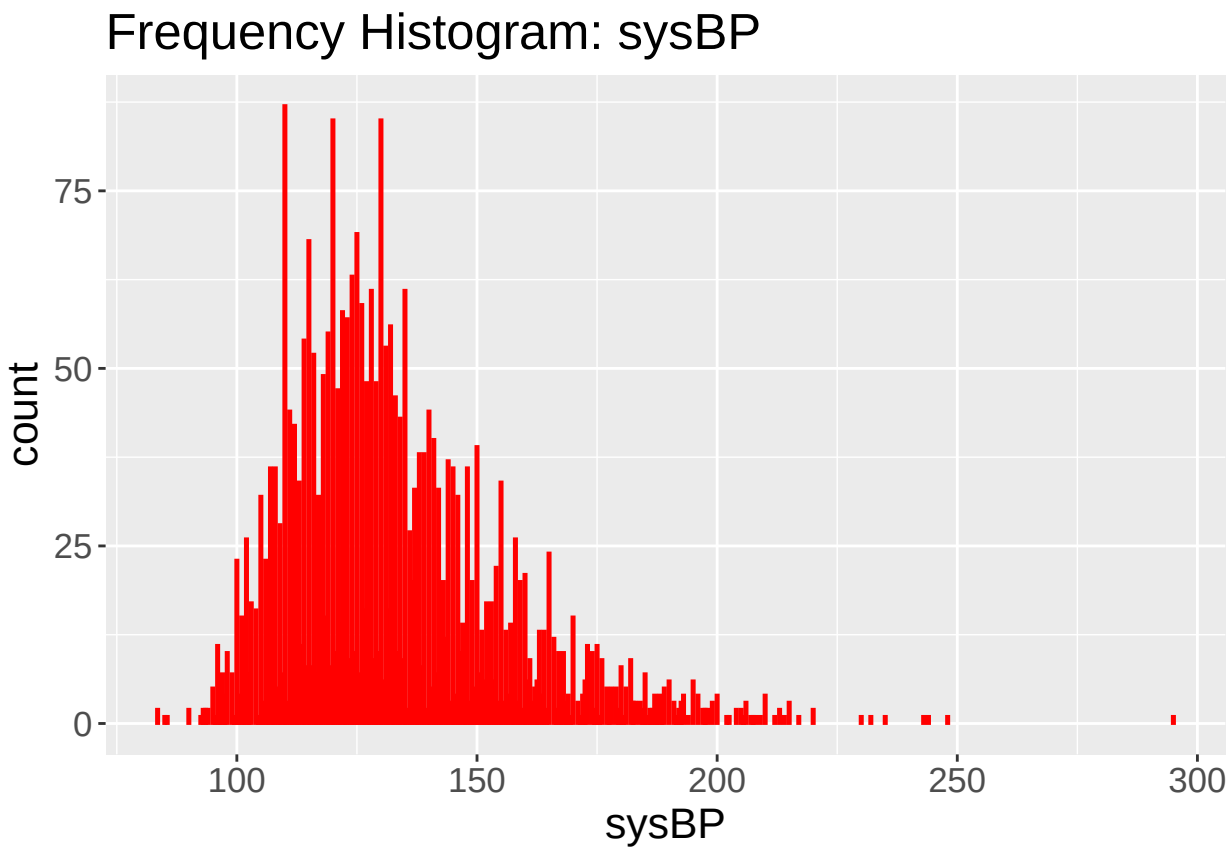

```
totChol_pvalue = t.test (cardio_healthy$totChol, cardio_chd$totChol)$p.value
```

Interpretation: Die Mehrheit der Teilnehmer hat einen Gesamtcholesterinspiegel zwischen 200 und 250.

Der Gesamtcholesterinspiegel hat einen statistisch signifikanten Effekt. Der P-value liegt bei 5.182689460121 33e-07.

sysBP

```
cardio %>%
  ggplot( aes(x =sysBP ,fill=sysBP))  +
    geom_bar(color="red")+

  theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: sysBP" )  +
          xlab ("sysBP") +
          ylab  ("count")
```



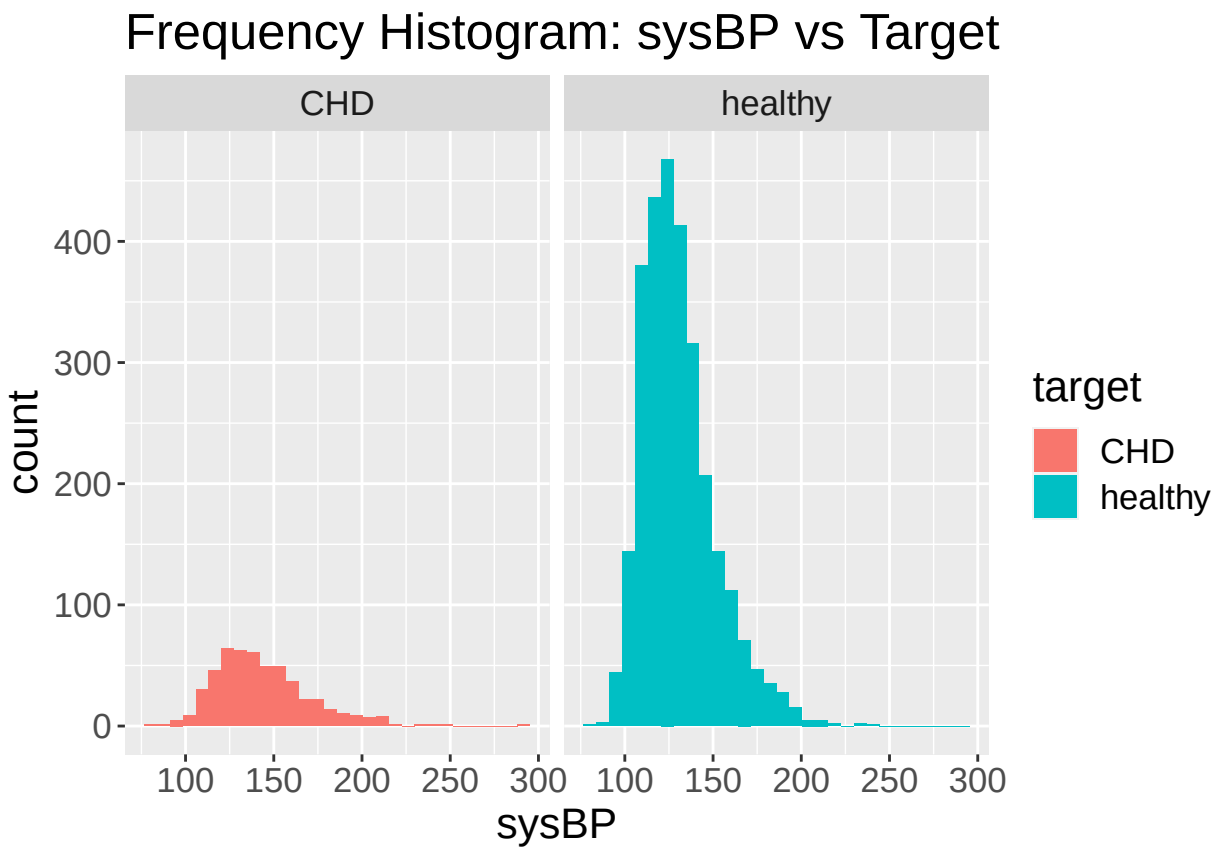

```
t_sysBP_count =round(table(cardio$sysBP)["sysBP"]/nrow(cardio) *100,1)
```

Interpretation: Der systolische Blutdruck, die obere Zahl, misst die Kraft, die Ihr Herz bei jedem Schlag auf die Wände Ihrer Arterien ausübt.

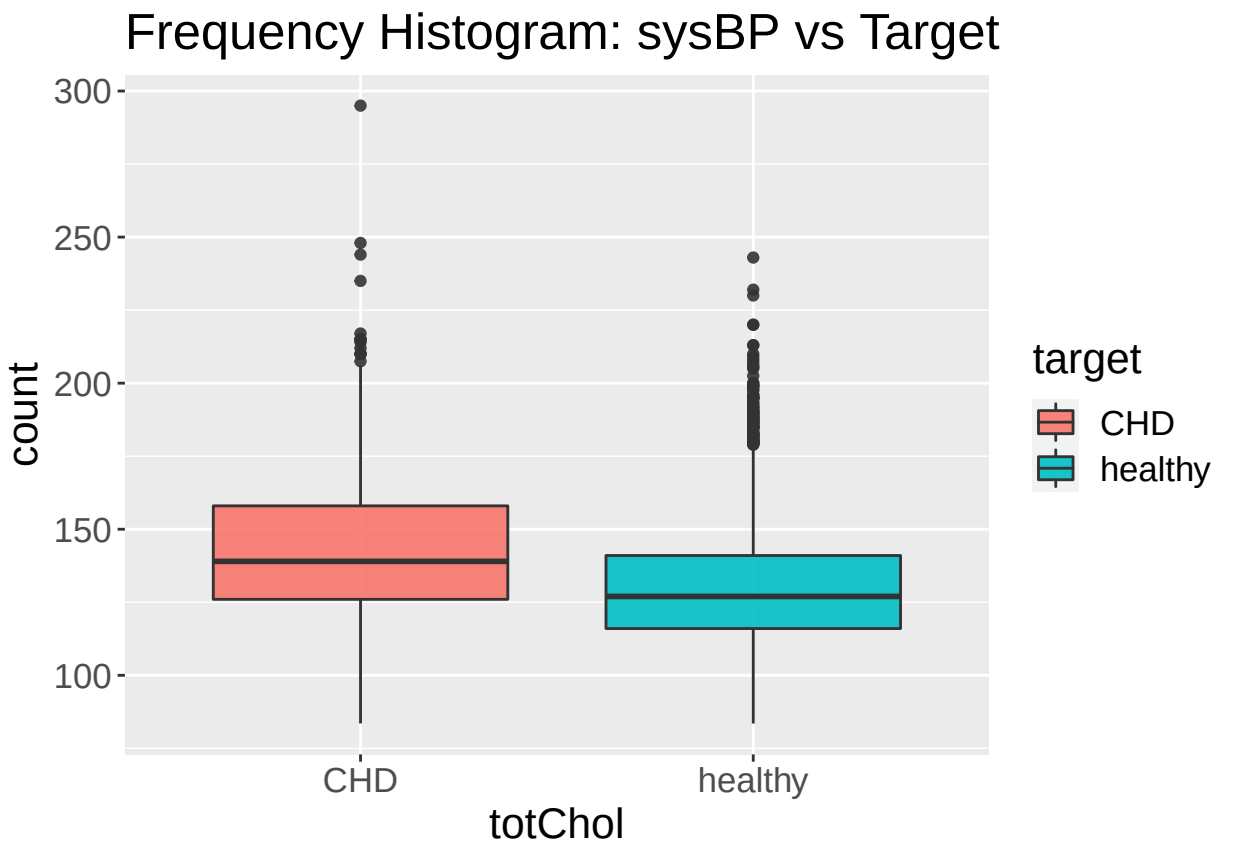
Der systolische Blutdruck liegt bei den meisten Teilnehmern im Bereich von 110-130 mmHg.

```
cardio %>%
  ggplot( aes(x = sysBP, fill = target)) +
    geom_histogram()+
    facet_wrap(~ target) +
    theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: sysBP vs Target" ) +
          xlab ("sysBP") +
          ylab ("count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
cardio %>%
  ggplot( aes(x = target ,y= sysBP, fill = target)) +
    geom_boxplot(alpha = 0.9)+
    theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: sysBP vs Target" ) +
          xlab ("totChol") +
          ylab ("count")
```




```
sysBP_pvalue = t.test (cardio_healthy$sysBP, cardio_chd$sysBP)$p.value
```

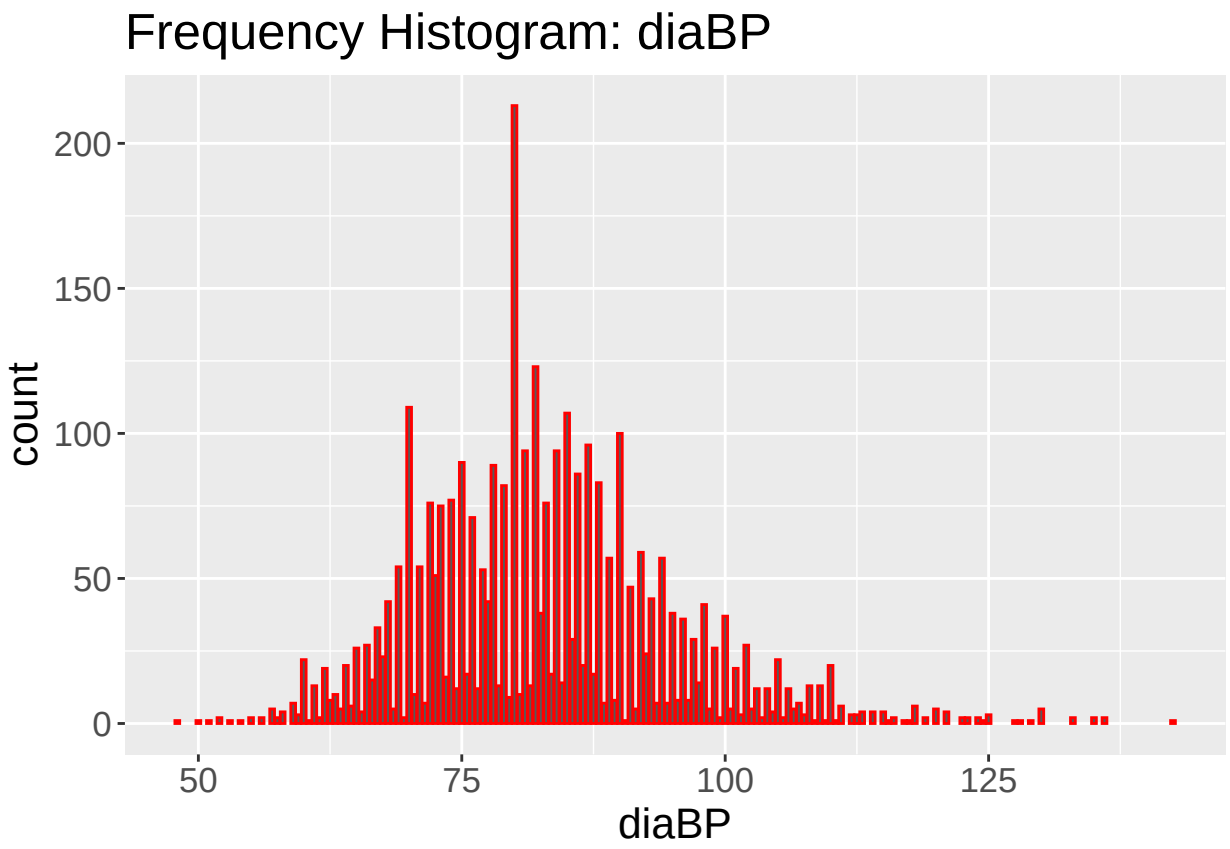
Interpretation:

Der systolische Blutdruck ist bei Kranken und Gesunden fast gleich hoch.

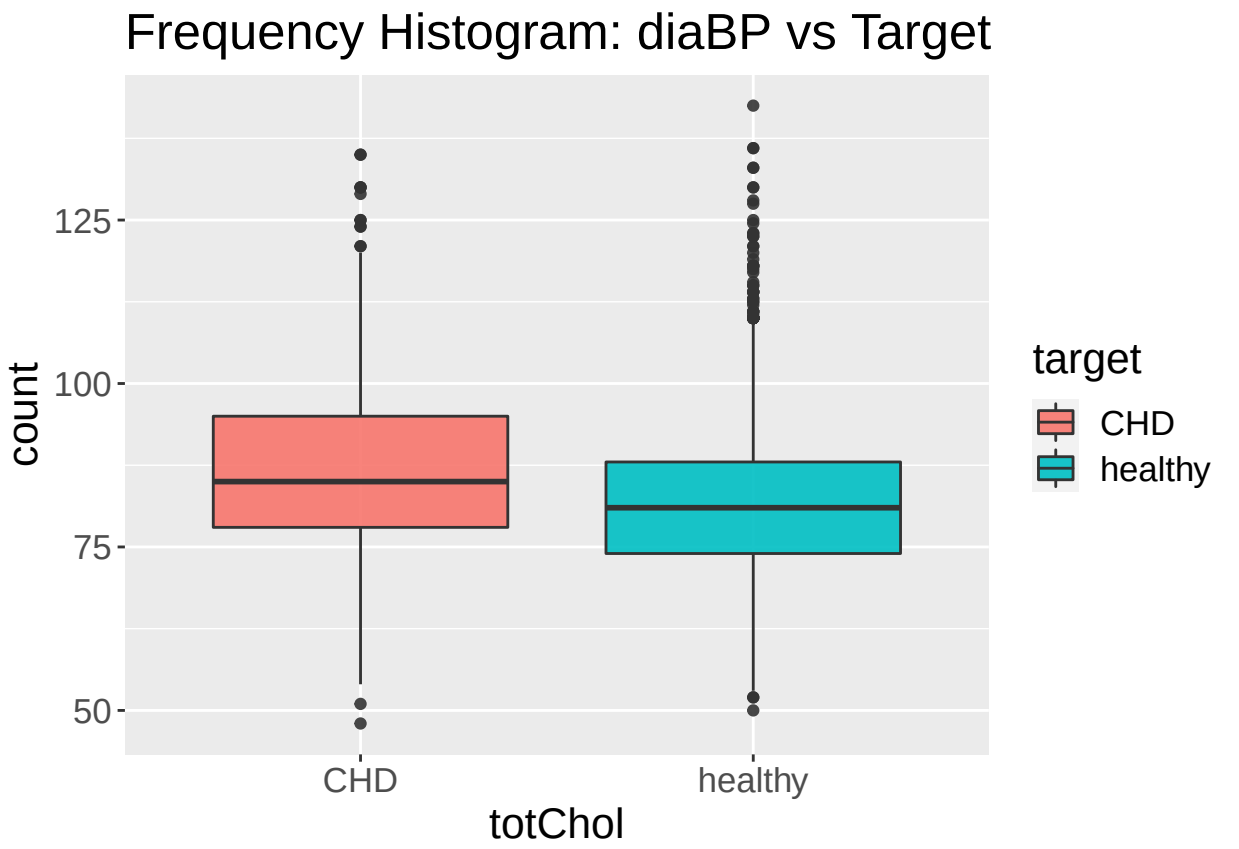
Der systolische Blutdruck hat einen statistisch signifikaten Einfluss. Der P-Value liegt bei 5.51533444666561e-24.

diaBP

```
cardio %>%
  ggplot( aes(x =diaBP ,fill=diaBP))  +
    geom_bar(color="red")+

  theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: diaBP" )  +
          xlab ("diaBP") +
          ylab  ("count")
```




```
cardio %>%
  ggplot( aes(x = target ,y= diaBP, fill = target)) +
    geom_boxplot(alpha = 0.9)+
    theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: diaBP vs Target" ) +
          xlab ("totChol") +
          ylab ("count")
```



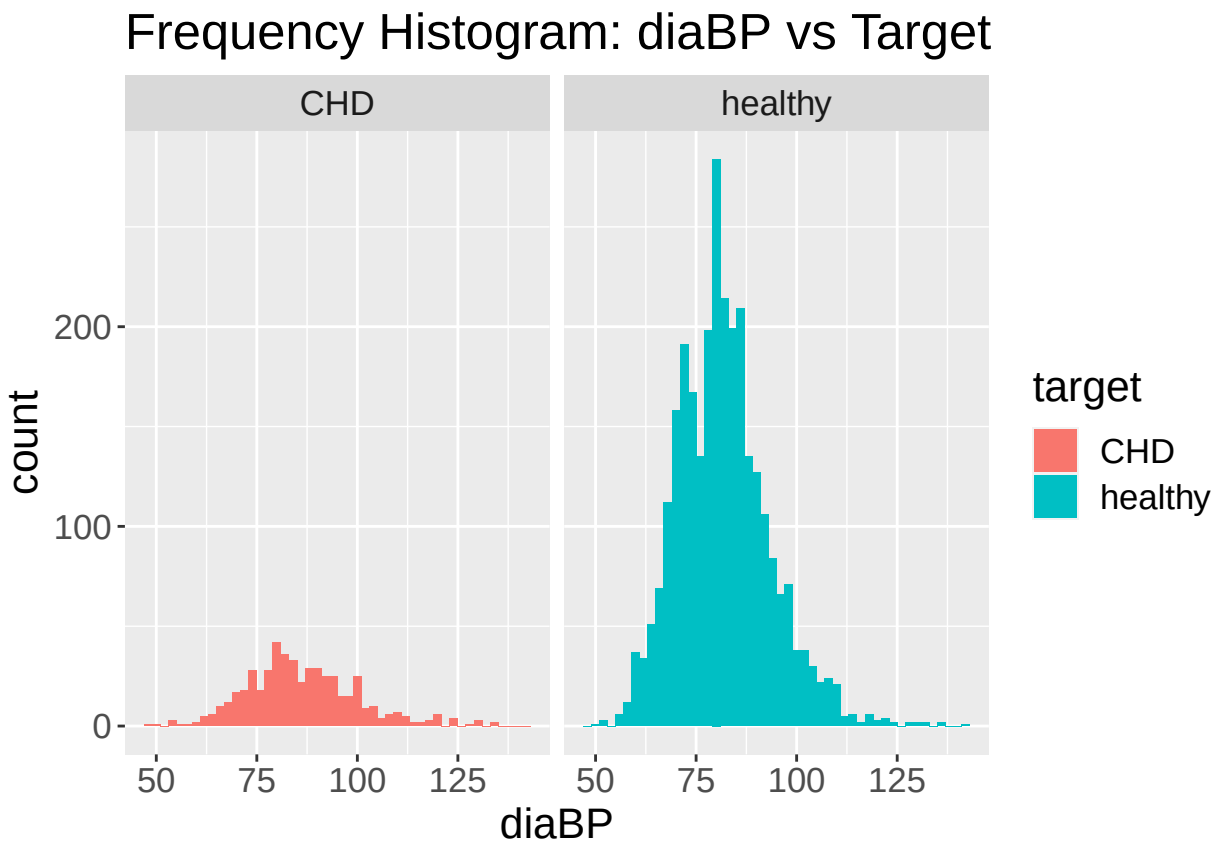

```
t_diaBP_count =round(table(cardio$diaBP)["diaBP"]/nrow(cardio) *100,1)

diaBP_pvalue = t.test (cardio_healthy$diaBP, cardio_chd$diaBP
)$p.value
```

Interpretation:

Die Mehrheit der Teilnehmer hat einen diaBP zwischen 75 und 100 und die meisten Teilnehmer haben diaBP für 80 .

Der P-value liegt bei 8.89527945660407e-12.

```
cardio %>%
  ggplot( aes(x = diaBP, fill = target)) +
    geom_histogram(binwidth =2)+
    facet_wrap(~ target) +
    theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: diaBP vs Target" ) +
          xlab ("diaBP") +
          ylab ("count")
```

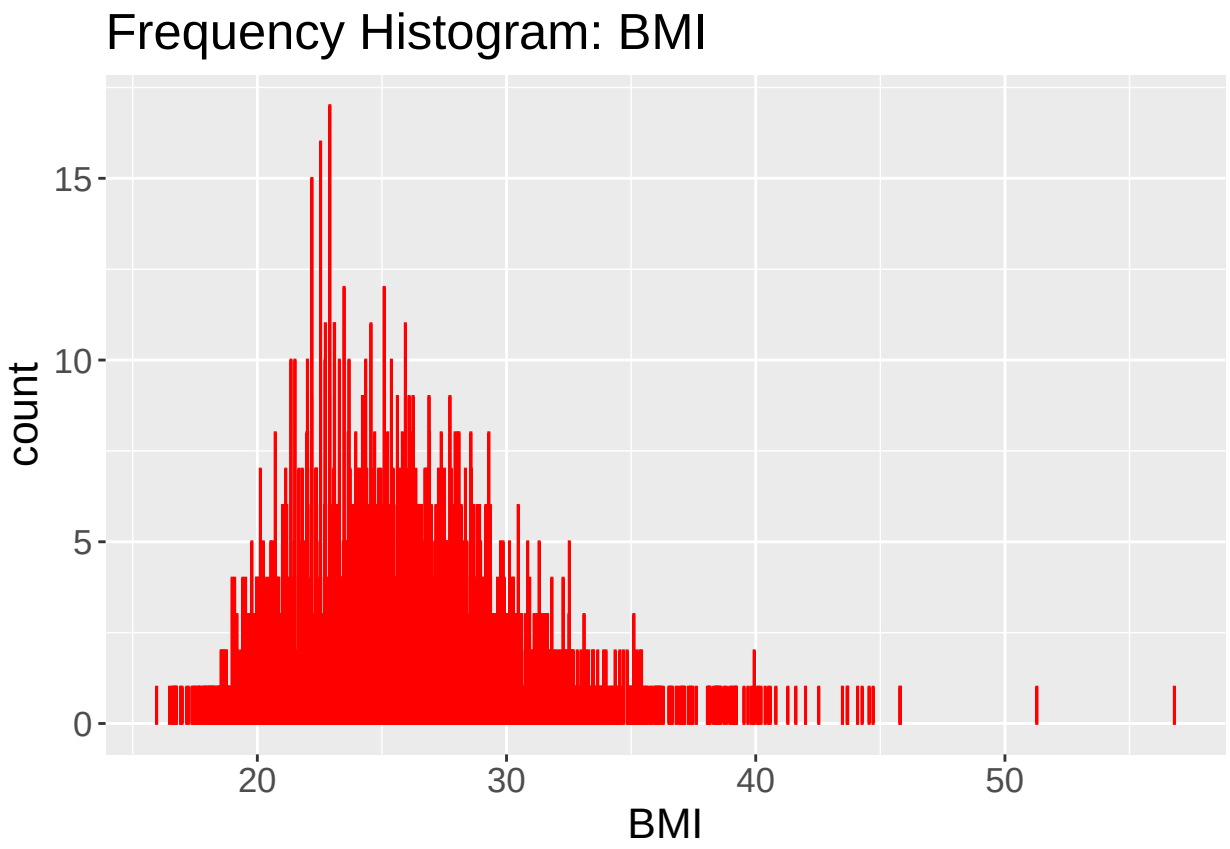



```
diaBP_pvalue = t.test (cardio_healthy$diaBP, cardio_chd$diaBP)$p.value
```

BMI

```
cardio %>%
  ggplot( aes(x =BMI ,fill=BMI))  +
    geom_bar(color="red")+

  theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: BMI" )  +
          xlab ("BMI") +
          ylab  ("count")
```

```
## Warning: Removed 14 rows containing non-finite values (stat_count).
```




```
t_BMI_count =round(table(cardio$BMI)["BMI"]/nrow(cardio) *100,1)

BMI_pvalue = t.test (cardio_healthy$BMI, cardio_chd$BMI)$p.value
```

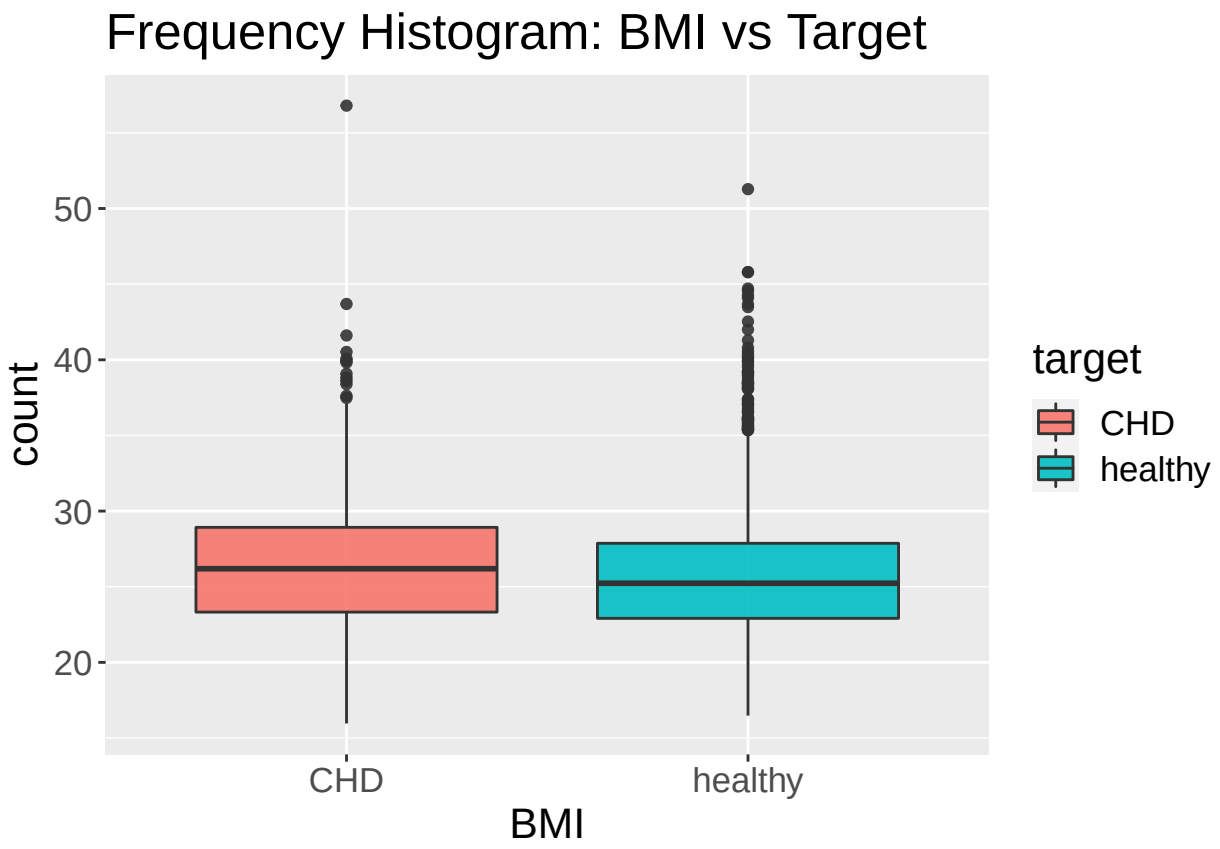
Interpretation:

Die Mehrheit der Teilnehmer hat einen BMI zwischen 20 und 30 und die meisten Teilnehmer haben BMI für 22.91 .

Der P-value liegt bei 0.000414155694522985.

```
cardio %>%
  ggplot( aes(x = target ,y= BMI, fill = target)) +
    geom_boxplot(alpha = 0.9)+
    theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: BMI vs Target" ) +
          xlab ("BMI") +
          ylab ("count")
```

```
## Warning: Removed 14 rows containing non-finite values (stat_boxplot).
```

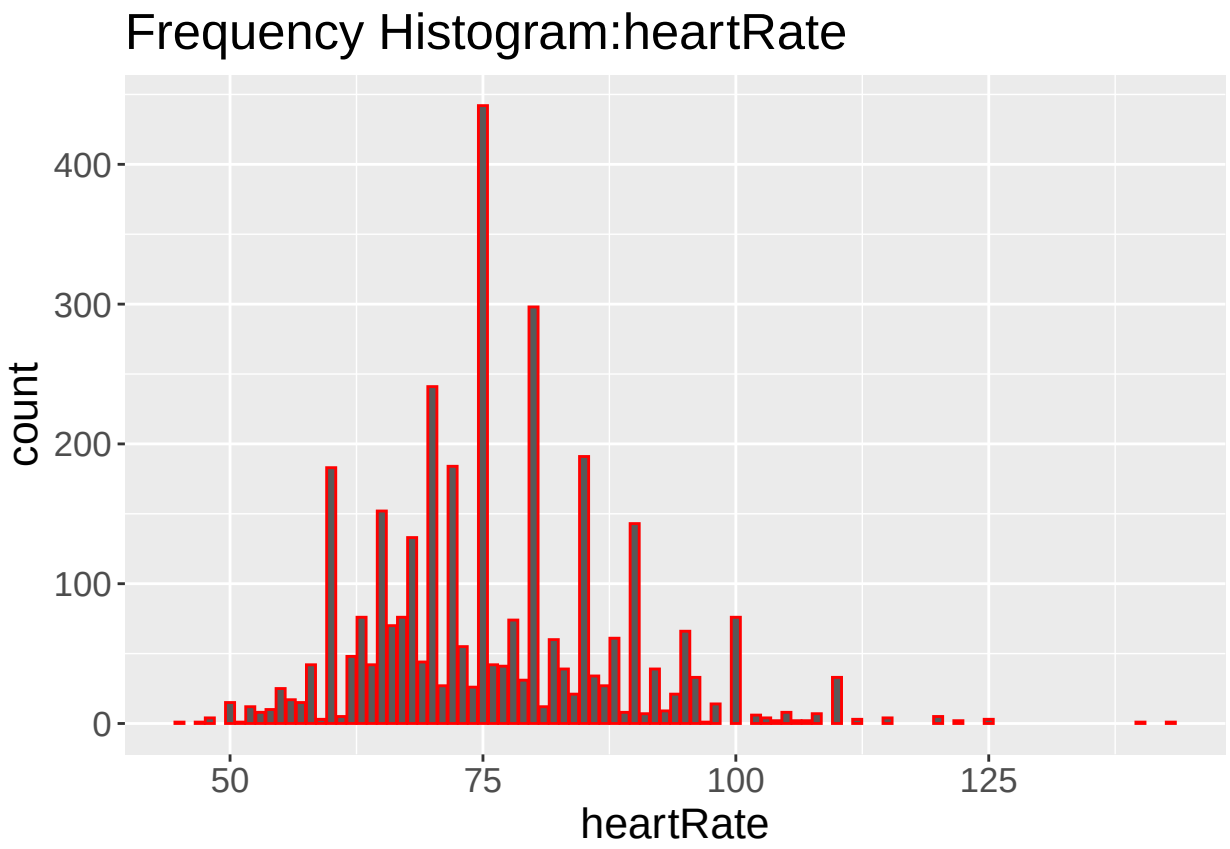



```
BMI_pvalue = t.test (cardio_healthy$BMI, cardio_chd$BMI)$p.value
```

heartRate

```
cardio %>%
  ggplot( aes(x =heartRate ,fill=heartRate))  +
    geom_bar(color="red")+

  theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram:heartRate" )  +
          xlab ("heartRate") +
          ylab  ("count")
```

```
## Warning: Removed 1 rows containing non-finite values (stat_count).
```




```
t_heartRate_count =round(table(cardio$heartRate)["heartRate"]/nrow(cardio) *100,1)

heartRate_pvalue = t.test (cardio_healthy$heartRat, cardio_chd$heartRat)$p.value
```

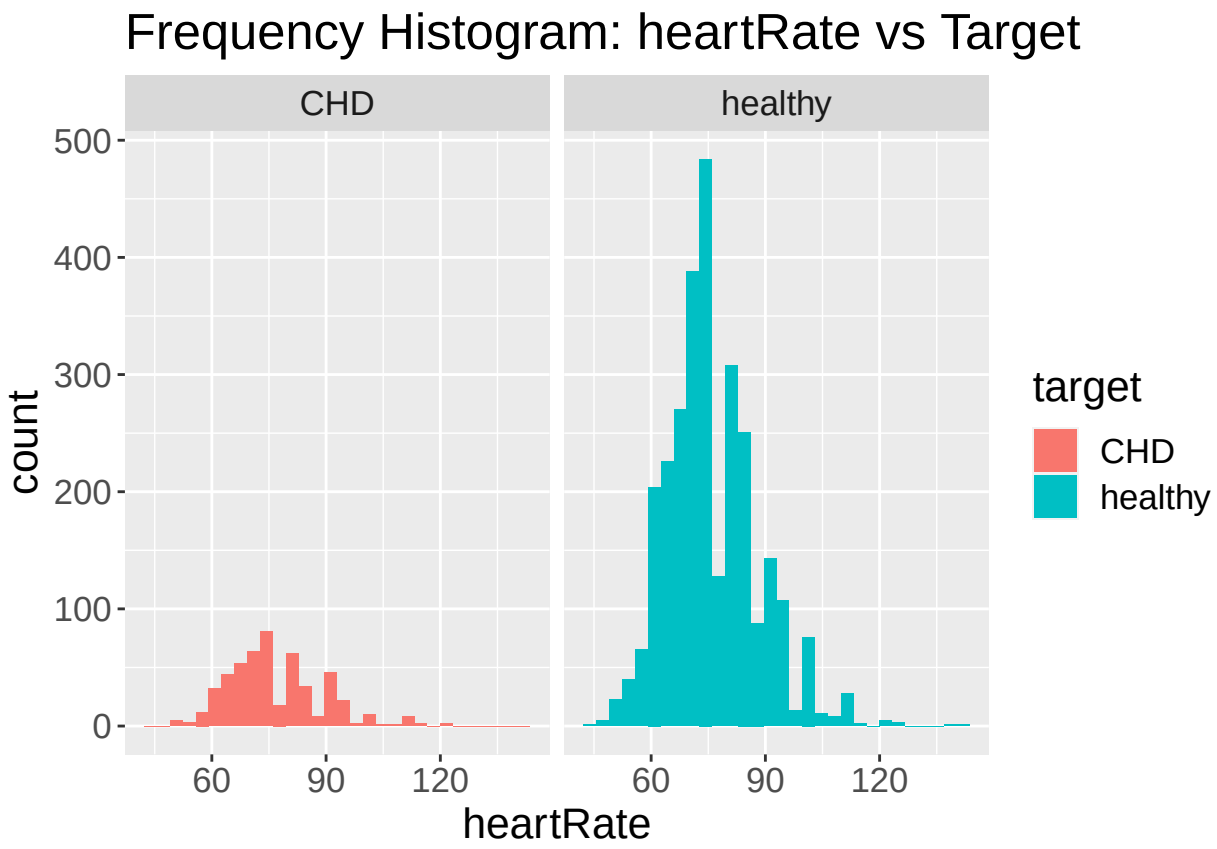
Interpretation:

Die Mehrheit der Teilnehmer hat einen heartRate zwischen 60 und 100 und die meisten Teilnehmer haben heartRate für 75 .

Der P-value liegt bei 0.245536931489089.

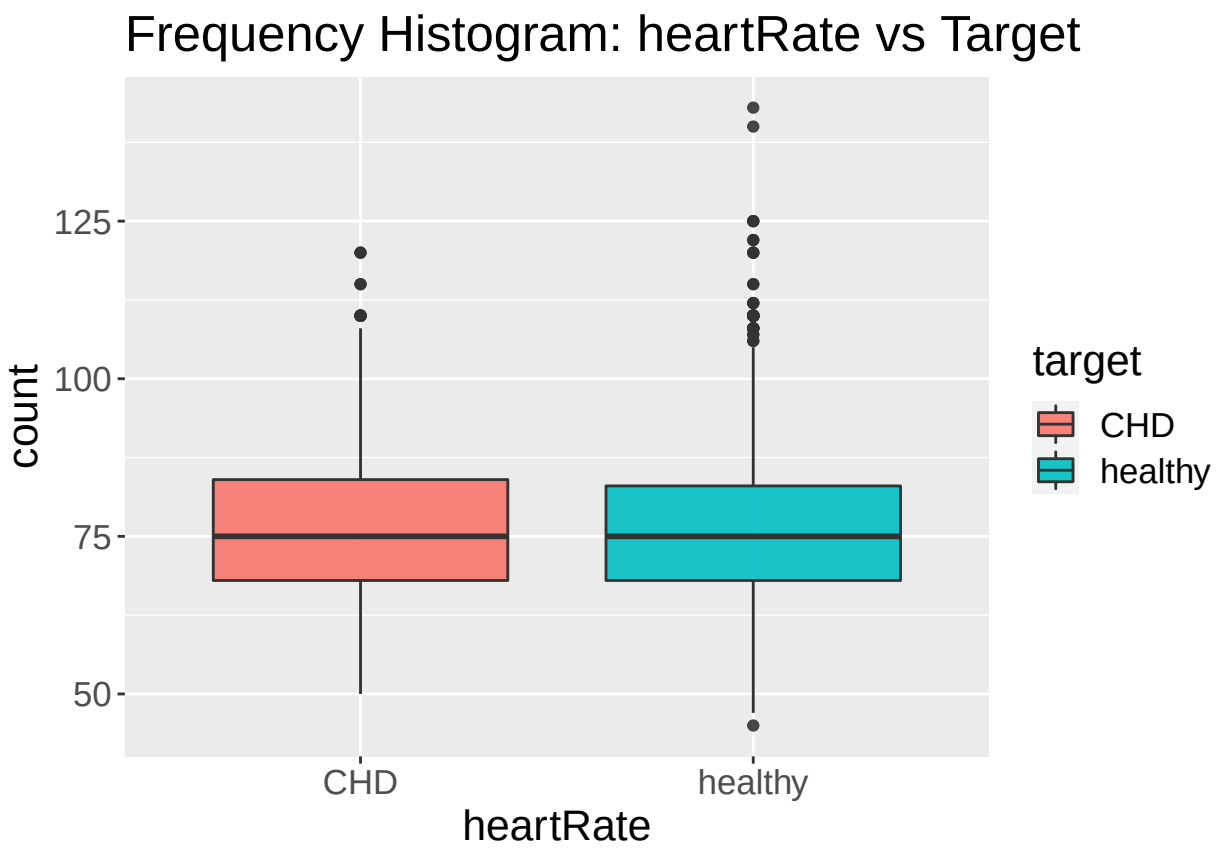
```
cardio %>%
  ggplot( aes(x = heartRate, fill = target)) +
    geom_histogram()+
    facet_wrap(~ target) +
    theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: heartRate vs Target" ) +
          xlab ("heartRate") +
          ylab ("count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 1 rows containing non-finite values (stat_bin).
```




```
cardio %>%
  ggplot( aes(x = target ,y= heartRate, fill = target)) +
    geom_boxplot(alpha = 0.9)+
    theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: heartRate vs Target" ) +
          xlab ("heartRate") +
          ylab ("count")
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```



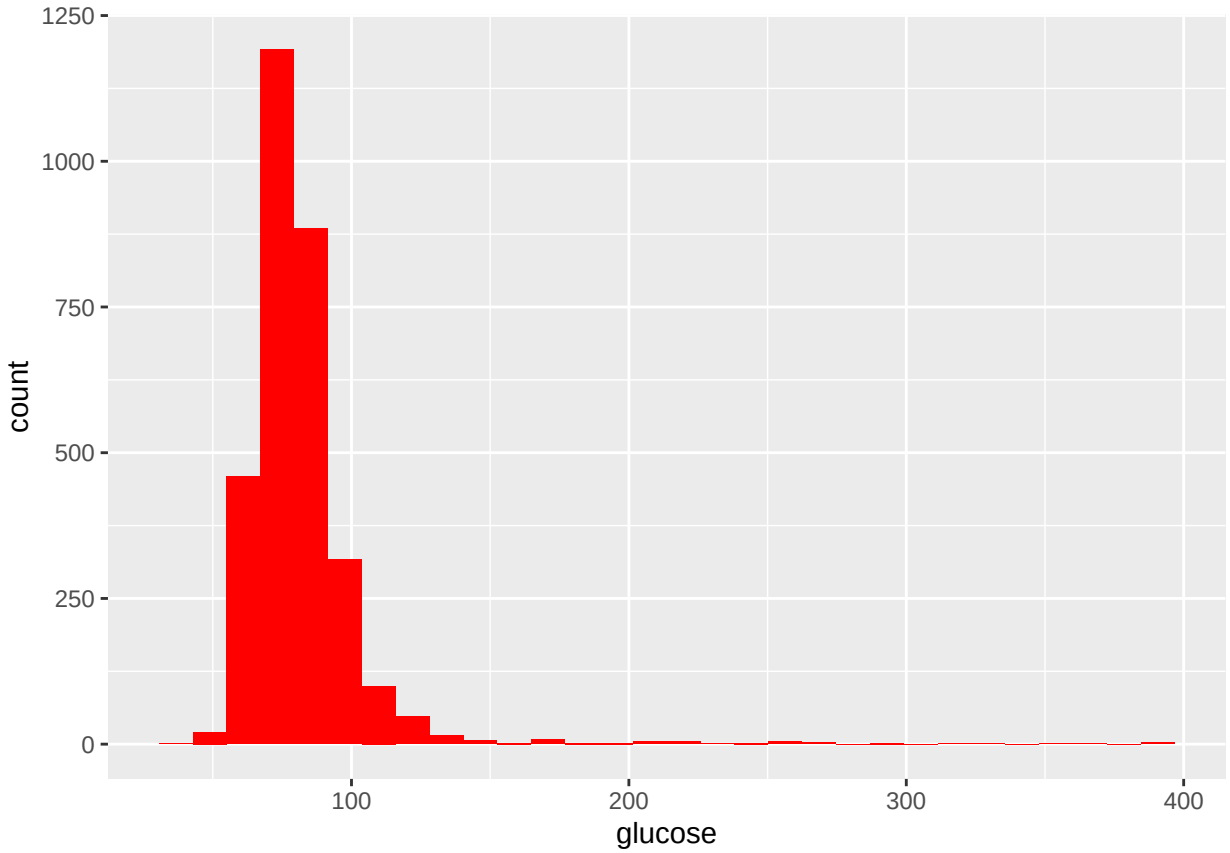

```
heartRate_pvalue = t.test (cardio_healthy$heartRate, cardio_chd$heartRate)$p.value
```

glucose

```
# glucose
cardio %>%
  ggplot( aes(x =glucose ,color=glucose))  +
    geom_histogram(fill="red")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 304 rows containing non-finite values (stat_bin).
```




```
t_glucose_count =round(table(cardio$glucose)["glucose"]/nrow(cardio) *100,1)

glucose_pvalue = t.test (cardio_chd$glucose, cardio_healthy$glucose)$p.value
```

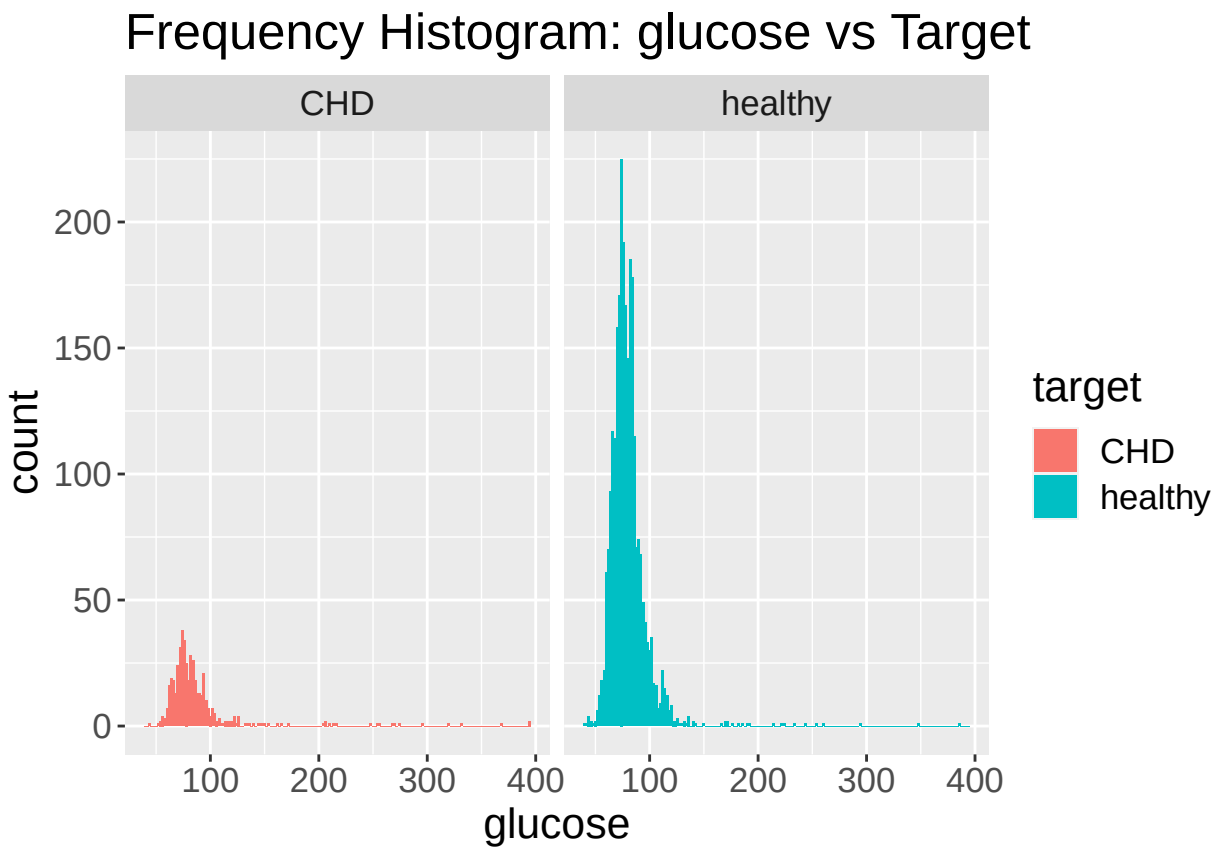
Interpretation:

Die Mehrheit der Teilnehmer hat einen heartRate zwischen 20 und 120 und die meisten Teilnehmer haben heartRate für 75 .

Der P-value liegt bei 3.66688738012339e-06.

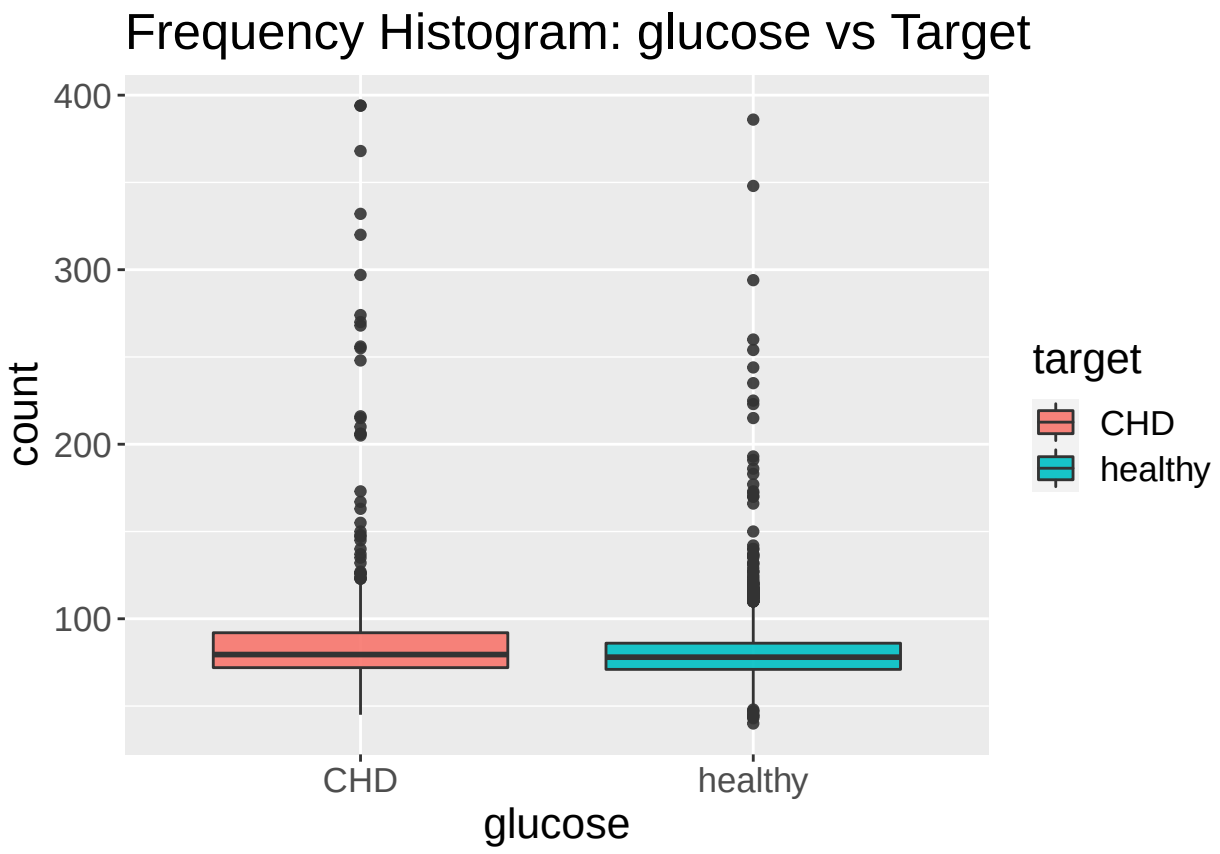
```
cardio %>%
  ggplot( aes(x = glucose, fill = target)) +
    geom_histogram(binwidth =2)+
    facet_wrap(~ target) +
    theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: glucose vs Target" ) +
          xlab ("glucose") +
          ylab ("count")
```

```
## Warning: Removed 304 rows containing non-finite values (stat_bin).
```




```
cardio %>%
  ggplot( aes(x = target ,y= glucose, fill = target)) +
    geom_boxplot(alpha = 0.9)+
    theme(text = element_text(size=16)) +
    labs ( title = "Frequency Histogram: glucose vs Target" ) +
          xlab ("glucose") +
          ylab ("count")
```

```
## Warning: Removed 304 rows containing non-finite values (stat_boxplot).
```




```
glucose_pvalue = t.test (cardio_chd$glucose, cardio_healthy$glucose)$p.value
```

Zusammenfassung

Tabelle mit jedem Feature ist eine Zeile.

Featurename, cardinal, ordinal oder nominal, Effekt, p-value, Anzahl an missing values

```
feature_summary = data.frame(feature = colnames(cardio)[1:ncol(cardio)-1])
feature_summary$art = c("nominal","ordinal","cardinal","cardinal","nominal","cardinal","cardinal","cardi

feature_summary$effekt =c("höheres Alter hat höheres Risiko"
                          , "education_inter"
                          , "Männer erkranken häufiger als Frauen  "
                          ,"der Effekt in beiden fast gleich"
                          , "weniger als 20 Zigaretten,geringes Risiko",
                          "meisten nehmen keine BPMeds ein",
                          "die Wirkung ist nicht effektiv",
                          "mit hypertensive ,höheres Risiko  ",
                         "Diabetes-Patienten haben ein höheres Risiko",
                         "der Effekt in beiden fast gleich",
                         "mit sysBP,höheres Risiko",
                         "höherer diaBP,höheres Risiko",
                         "höherer BMI hat höheres Risiko",
```

```
                          "der Effekt in beiden fast gleich",
                          "Glucose-Patienten haben ein höheres Risiko ")

feature_summary$p_value = c(toString(age_pvalue),toString(education_p_value), sex_pvalue,smoking_pvalue
feature_summary$nbr_missing = ""


for (ir in 1:nrow(feature_summary)) {
    tmp_col <- cardio[,feature_summary$feature[ir] ]
    feature_summary$nbr_missing[ir] <- nrow(cardio) -  table(is.na(tmp_col))["FALSE"]
}

knitr::kable(feature_summary)
```

feature	art	effekt	p_value	nbr_missing
age	nominal	höheres Alter hat höheres Risiko	1.84555411177536e-38	0
education	ordinal	education_inter	6.68203706919536e-05	87
sex	cardinal	Männer erkranken häufiger als Frauen	9.50443889414983e-07	0
smoking	cardinal	der Effekt in beiden fast gleich	0.0490648844448193	0
cigsPerDay	nominal	weniger als 20 Zigaretten,geringes Risiko	0.000373330337109478	22
BloodPresMed	cardinal	meisten nehmen keine BPMeds ein	5.23410965432645e-06	44
stroke	cardinal	die Wirkung ist nicht effektiv	0.000647033881655412	0
hypertensive	cardinal	mit hypertensive ,höheres Risiko	4.91775745429724e-21	0
diabetes	cardinal	Diabetes-Patienten haben ein höheres Risiko	8.89527945660407e-12	0
totChol	nominal	der Effekt in beiden fast gleich	5.18268946012133e-07	38
sysBP	nominal	mit sysBP,höheres Risiko	5.51533444666561e-24	0
diaBP	nominal	höherer diaBP,höheres Risiko	8.89527945660407e-12	0
BMI	nominal	höherer BMI hat höheres Risiko	0.000414155694522985	14
heartRate	nominal	der Effekt in beiden fast gleich	0.245536931489089	1
glucose	nominal	Glucose-Patienten haben ein höheres Risiko	3.66688738012339e-06	304

Literatur und Quellen

- Markdown Tuorial
- Markdown Cheatsheet
- R for Data Science
- R Introduction

https://www.kaggle.com/christofel04/cardiovascular-study-dataset-predict-heartdisea

https://www.kaggle.com/datasets/christofel04/cardiovascular-study-dataset-predict-heart-disea