Total Draft

R-Projekt

Eine Datenanalyse bezüglich kardiovaskulärer Erkrankungen

Einführung in die Wahrscheinlichkeitstheorie und Statistik [MA0009]

Prof. Dr. Silke Rolles

Stephan Haug

München, 12. Februar, 2022

Axha, Frenkli
Grünewald, Mathis A.
Hermann, Friedrich K.

1. Einleitung

Die folgende Analyse des vorliegenden Datensatzes dient der Untersuchung des Einflusses verschiedener Faktoren, im Zusammenhang mit dem individuellen Lebensstil einer Person, auf die Entwicklung einer kardiovaskulären Krankheit.
Dazu werden die Daten des Datensatzes cardio_train untersucht, der Variablen bezüglich des Gesundheitsstandes, des Lebenstils und des Vorliegens einer kardiovaskulären Krankheit von 70000 Patient:innen enthält. Der Datensatz wurde von Kaggle [1] bereitgestellt.
Hier ein Überblick der, mit der Variable BMI erweiterten, Variablen des Datensatzes:

```
##   id      age gender height weight ap_hi ap_lo cholesterol gluc smoke alco
## 1  0 50.39178      2   1.68     62   110    80           1    1     0    0
## 2  1 55.41918      1   1.56     85   140    90           3    1     0    0
## 3  2 51.66301      1   1.65     64   130    70           3    1     0    0
## 4  3 48.28219      2   1.69     82   150   100           1    1     0    0
## 5  4 47.87397      1   1.56     56   100    60           1    1     0    0
## 6  8 60.03836      1   1.51     67   120    80           2    2     0    0
##   active cardio      BMI
## 1      1      0 21.96712
## 2      1      1 34.92768
## 3      0      1 23.50781
## 4      1      1 28.71048
## 5      0      0 23.01118
## 6      0      0 29.38468
```

Einerseits enthält der Datensatz Variablen bezüglich des Gesundheitsstandes der Patient:innen:

-age : Alter der Patient:innen in Jahre.
-height : Größe der Patient:innen in Meter.
-weight : Gewicht in kg.
-ap_hi : systolischer Blutdruck in mmHg.
-ap_lo : diastolischer Blutdruck in mmHg.
-gluc : Glucoselevel im Blut auf einer Skala von 1 bis 3; wobei 1 “normale” Glucosewerte, 2 “mittelgute” Glucosewerte und 3 “schlechte” Glucosewerte beschreibt.
-cholesterol : Cholesterollevel im Blut auf einer Skala von 1 bis 3, analog zu gluc.
-cardio : Vorliegen einer kardiovaskulären Krankheit in binär: 1 für das Vorliegen einer kardiovaskulären; 0 falls keine kardiovaskuläre Krankheit vorliegt. Dies ist die Zielvariable, die näher untersucht werden wird.

Andererseits enthält er Variablen bezüglich des Lebensstils der Patient:innen:

-smoke : Raucheangewohnheit, 1: der Patient:in Raucht; 0: der Patient:in raucht nicht.
-alco : Alkoholkonsum, binär wie bei smoke.
-active : Sportangewohnheit, binär wie bei smoke.

Unsere Absicht ist es zu untersuchen, ob der Gesundheitsstand einen Einfluss auf das Vorliegen einer kardiovaskulären Krankheit hat; und wiederum, ob der Lebensstil einen Einfluss auf den Gesundheitsstand hat.

[1] https://www.kaggle.com/abdallahmahmoud/cardiovascular-disease-prediction-73-59-accuracy

2. Explorative Datenanalyse

Der folgende Abschnitt dient zur Einführung in den vorliegenden Datensatz. Dabei wird dem Leser ein grober Überblick der Gesamtstruktur des Datensatzes gegeben und im späteren Verlauf genauere Charakteristiken bezüglich der Variablen die in der späteren Analyse noch von größerer Bedeutung sind.

Übersicht des Datensatzes und Struktur

```
##        id              age            gender         height
##  Min.   :    0   Min.   :29.58   Min.   :1.00   Min.   :0.550
##  1st Qu.:25007   1st Qu.:48.39   1st Qu.:1.00   1st Qu.:1.590
##  Median :50002   Median :53.98   Median :1.00   Median :1.650
##  Mean   :49972   Mean   :53.34   Mean   :1.35   Mean   :1.644
##  3rd Qu.:74889   3rd Qu.:58.43   3rd Qu.:2.00   3rd Qu.:1.700
##  Max.   :99999   Max.   :64.97   Max.   :2.00   Max.   :2.500
##      weight           ap_hi             ap_lo           cholesterol
##  Min.   : 10.00   Min.   : -150.0   Min.   :  -70.00   Min.   :1.000
##  1st Qu.: 65.00   1st Qu.:  120.0   1st Qu.:   80.00   1st Qu.:1.000
##  Median : 72.00   Median :  120.0   Median :   80.00   Median :1.000
##  Mean   : 74.21   Mean   :  128.8   Mean   :   96.63   Mean   :1.367
##  3rd Qu.: 82.00   3rd Qu.:  140.0   3rd Qu.:   90.00   3rd Qu.:2.000
##  Max.   :200.00   Max.   :16020.0   Max.   :11000.00   Max.   :3.000
##       gluc            smoke              alco             active
##  Min.   :1.000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:1.0000
##  Median :1.000   Median :0.00000   Median :0.00000   Median :1.0000
##  Mean   :1.226   Mean   :0.08813   Mean   :0.05377   Mean   :0.8037
##  3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :3.000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##      cardio            BMI
##  Min.   :0.0000   Min.   :  3.472
##  1st Qu.:0.0000   1st Qu.: 23.875
##  Median :0.0000   Median : 26.374
##  Mean   :0.4997   Mean   : 27.557
##  3rd Qu.:1.0000   3rd Qu.: 30.222
##  Max.   :1.0000   Max.   :298.667
```

mit der Struktur

```
## 'data.frame':    70000 obs. of  14 variables:
##  $ id         : int  0 1 2 3 4 8 9 12 13 14 ...
##  $ age        : num  50.4 55.4 51.7 48.3 47.9 ...
##  $ gender     : int  2 1 1 2 1 1 1 2 1 1 ...
##  $ height     : num  1.68 1.56 1.65 1.69 1.56 1.51 1.57 1.78 1.58 1.64 ...
##  $ 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 ...
##  $ BMI        : num  22 34.9 23.5 28.7 23 ...
```

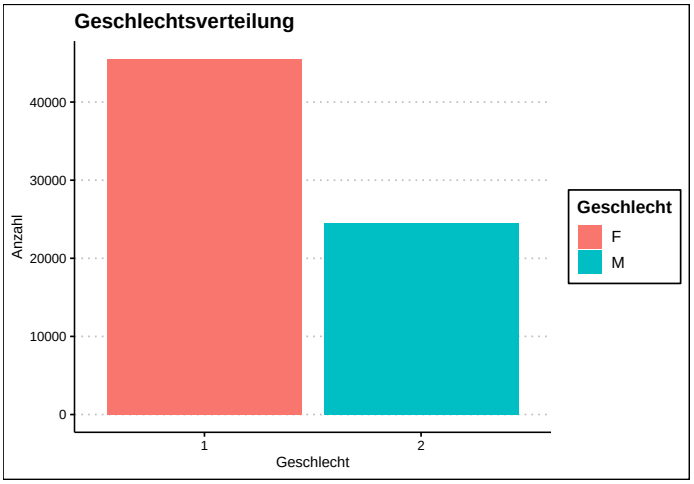
Zuerst untersuchen wir, ob unser Datensatz lückenhaft ist

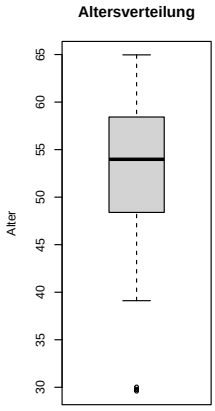
```
number_of_columns <- ncol(df_1)
number_of_lines <- nrow(df_1)
missing_values <- is.na(df_1)

sum(missing_values)
```

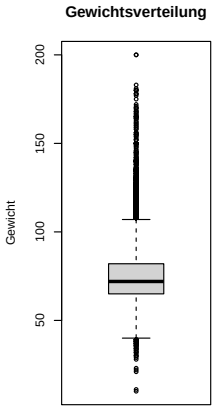
```
## [1] 0
```

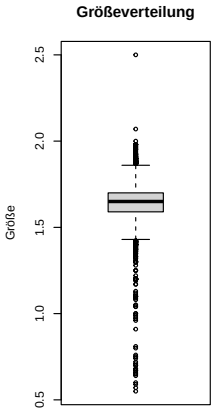
Ein Überblick über die Patient:innen liefert

```
## Scale for 'fill' is already present. Adding another scale for 'fill', which
## will replace the existing scale.
```





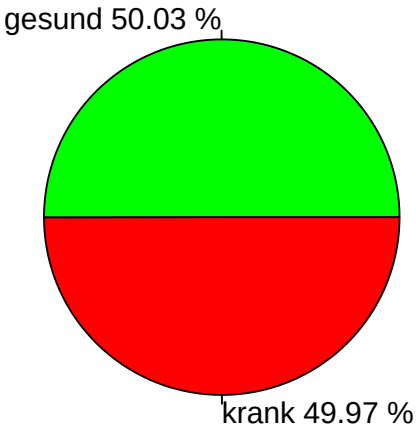




Es handelt sich also um Durchschnittlich 53 jährige, überwiegend weibliche Patient:innen. Dies erklärt auch, weshalb die mittlere Größe “nur” 1,64m ist.

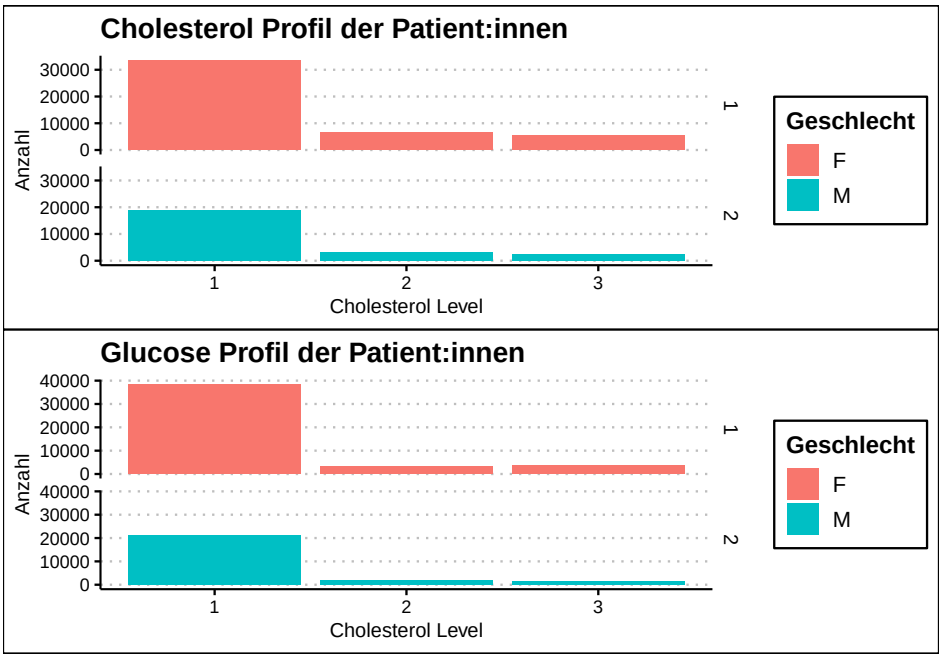
Nun ein Überblick vom Gesundheitsstand der Patient:innen

Kranksheitszustand der Patient:innen

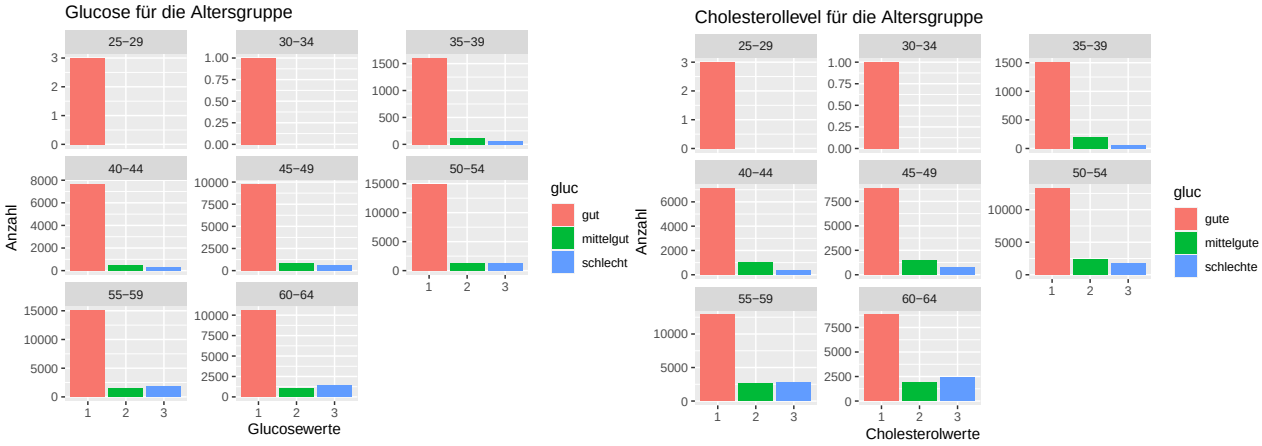


Es gibt in unserem Datensatz also fast genauso viele kranke, wie gesunde Patienten.

Untersuchung der Blutwerte ergibt



Man erkennt an dem Plot, dass die Mehrheit der Patient:innen - sowohl die männlichen als auch weiblichen - gute Glucose- und Cholesterolwerte hat.
Einen genaueren Einblick lieferen beide folgenden Plots.

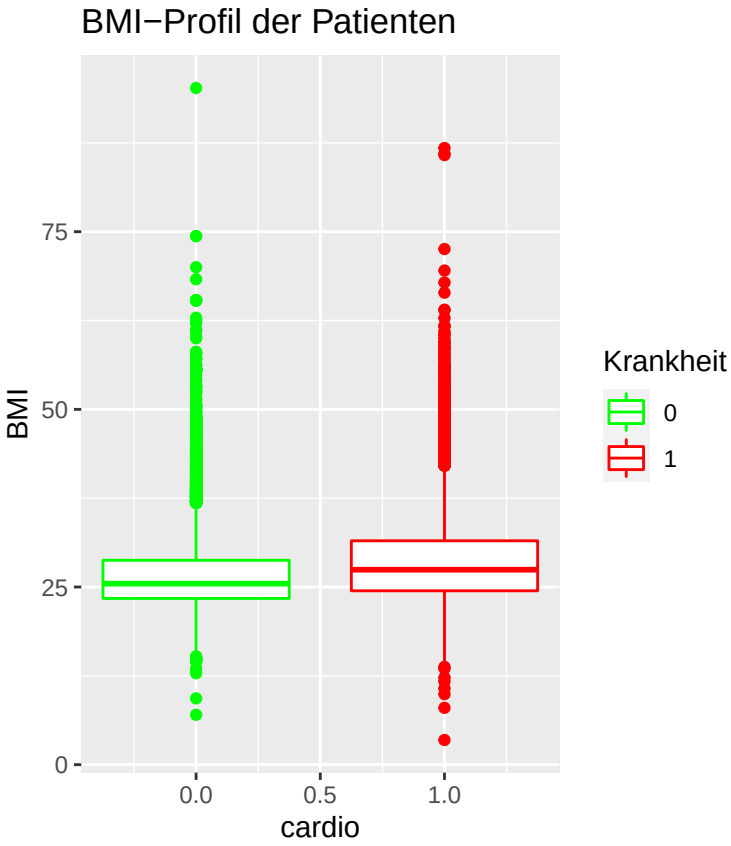


Hier erkennt man ganz klar, dass die Blutwertqualität mit dem alter abnimmt. Je älter die Patient:innen, umso größer ist der Anteil an Patient:innen mit mittelguten und schlechten Glucose bzw. Cholesterolwerten.

Dies gibt uns Anlass zur explorativen Suche nach Einflussvariablen für das Auftreten von kardiovaskulären Krankheiten.

Suche nach Einflussvariablen auf das Auftreten von kardiovaskulären Krankheiten:

Wir filtern die Patient:innen heraus, deren Blutdruckwerte und BMI unrealistisch sind, und erhalten das BMI Profil



BMI Mittelwert der gesunden Patient:innen:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.022  23.389  25.473  26.483  28.764  95.222
```

BMI-Mittelwert der kranken Patient:innen:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.472  24.465  27.435  28.473  31.504  86.777
```

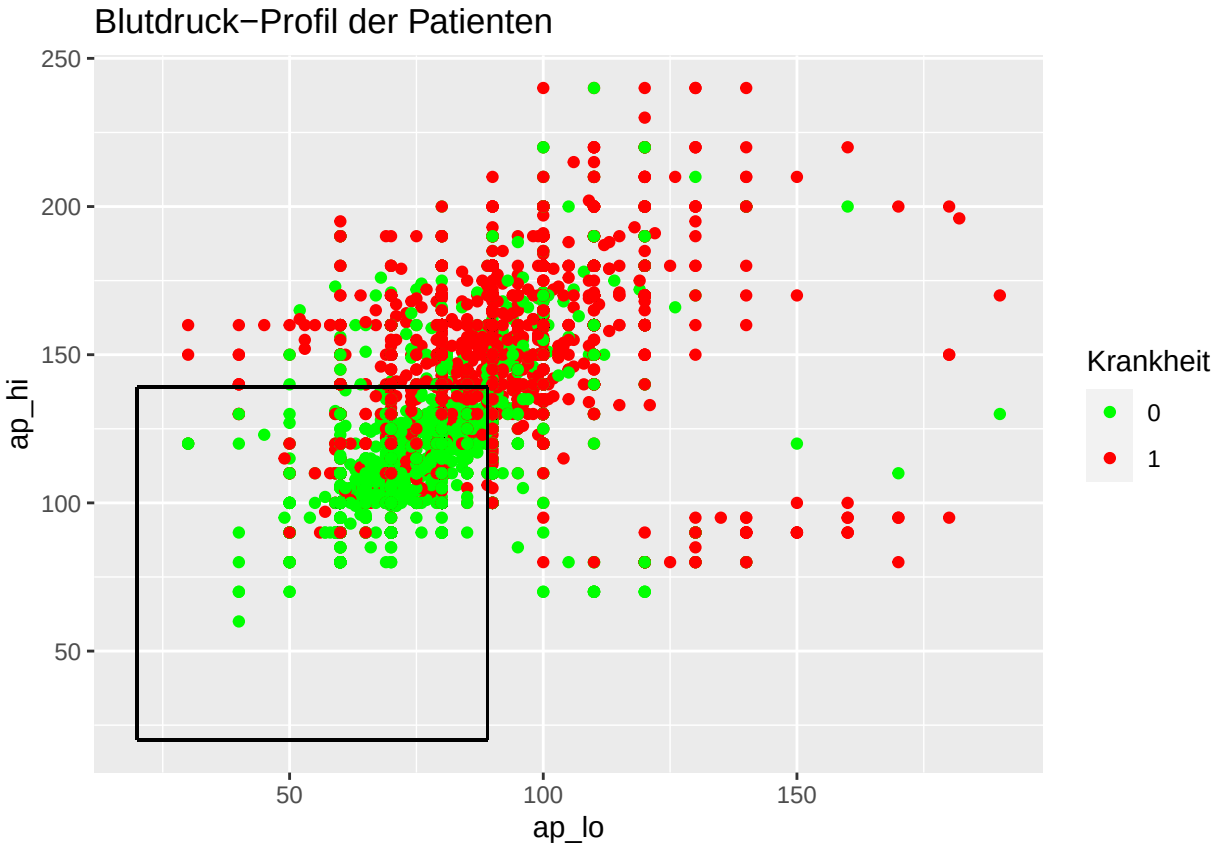
Man sieht anhand des Plot und der Tabellen, dass die kranken Patienten nur einen leicht höheren BMI vorweisen, als die gesunden. Entsprechend 26.483 und 28.473 im Mittel. Vermutlich hat die Variable BMI keinen erheblichen Einfluss, da sich hier keinen klaren Trend ablesen lässt. Genaueres wird im Abschnitt 3 und 4 untersucht. Jedoch gilt ein:e Patient:in ab einem BMI von 25 als übergewichtig, also haben sowohl die gesunden als auch die kranken Patient:innen einen zu hohen BMI.

Untersuchung der Blutdruckwerte:

Wir definieren:

"guter Blutdruck" : ap_hi $\leq$ 139 und ap_lo $\leq$ 89
"schlechter Blutdruck": ap_hi $\geq$ 140 oder ap_lo $\geq$ 90 dabei basieren wir uns auf der schweizerischen Herzstiftung[2].

[2]https://www.swissheart.ch/herzkrankheiten-hirnschlag/risikofaktoren/blutdruck/was-ist-bluthochdruck.html



Krankheitsrate in der Gruppe mit gutem Blutdruck

```
## [1] 0.3458205
```

Krankheitsrate in der Gruppe mit schlechtem Blutdruck

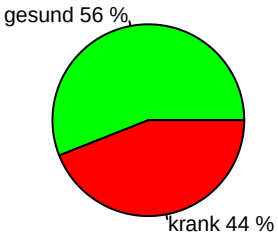
```
## [1] 0.779451
```

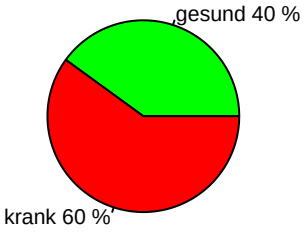
Im Plot befindet sich innerhalb des Quadrats die Gruppe an Patienten:innen, die gute Blutdruckwerte haben und außerhalb des Quadrats die, der Patient:innen mit schlechten Blutdruckwerten. Graphisch erkennt man, dass sich innerhalb des Quadrats eher mehr grüne Punkte befinden als rote; und außerhalb des Quadrats andersherum - was durch die Krankheitsraten in den jeweiligen Gruppen bestätigt wird.

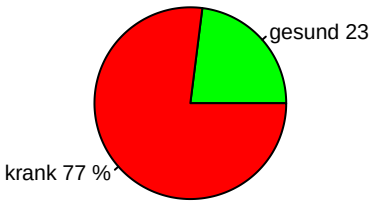
Im folgenden haben wir die Patient:innen, für die Variablen *cholesterol* und *gluc*, in drei Gruppen je nach Blutwerten aufgeteilt und jeweils den Anteil an kranken und gesunden Patient:innen untersucht.

Für die Variable *cholesterol*:

gute Cholesterolwerte **mittelgute Cholesterolwerte** **schlechte Cholesterolwerte**

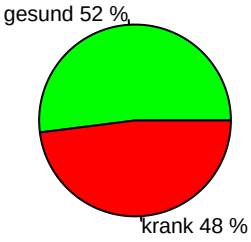


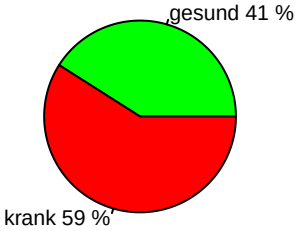


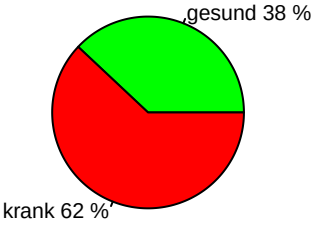


Für die Variable *gluc*:

gute Glucosewerte **mittelgute Glucosewerte** **schlechte Glucosewerte**







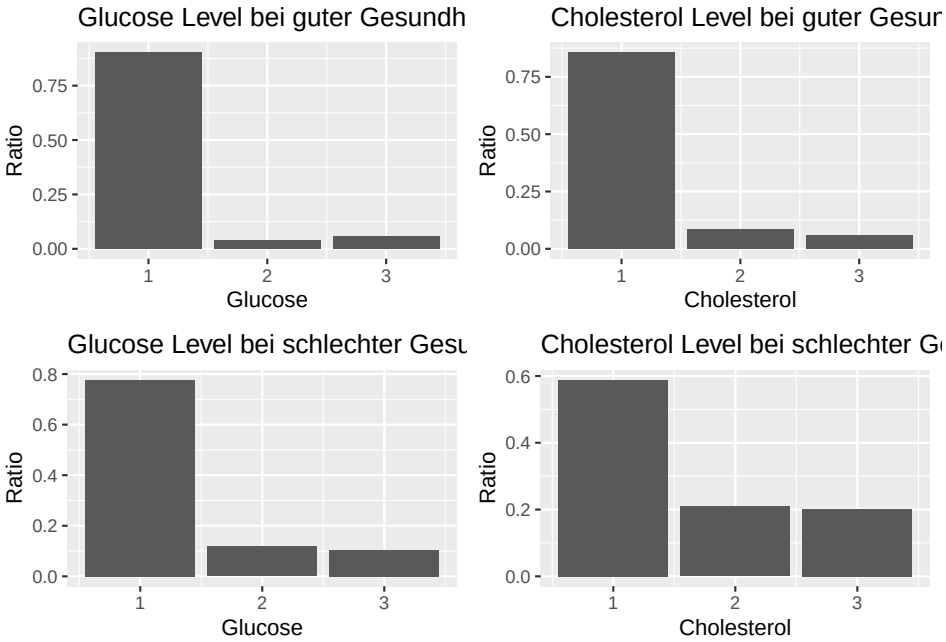
Klar erkennbar ist, dass in beiden Fällen der Anteil an kranken Patient:innen steigt, umso schlechter der Blutwert ist. Es lässt sich also vermuten, dass eine große Korrelation zwischen den Variablen *gluc* und *cholesterol* einerseits, und *cardio* andererseits vorhanden ist.

Schließlich wird der Lebensstil der Patient:innen untersucht:

Wir definieren dazu:

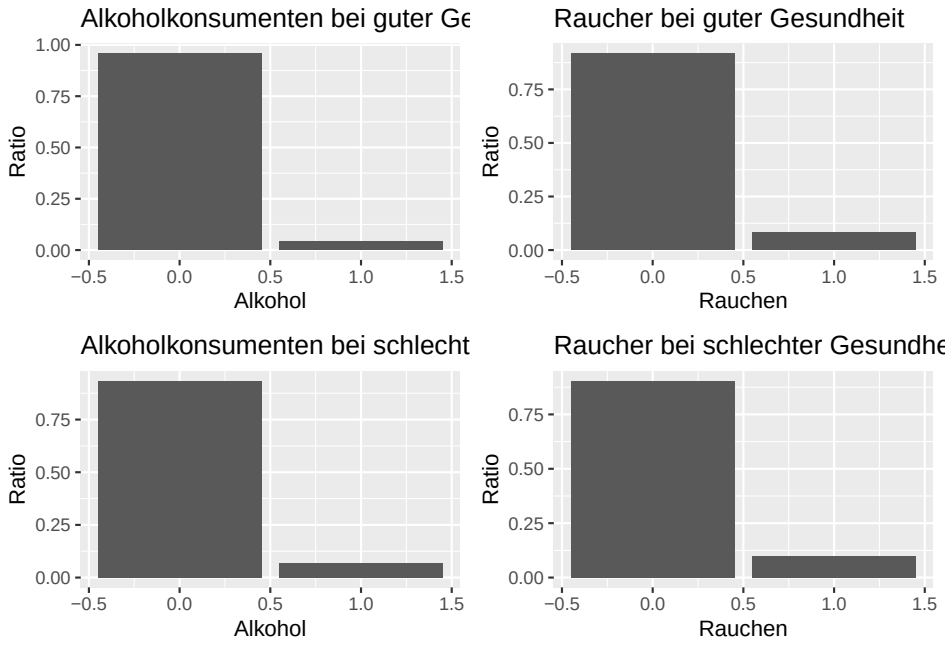
gute Gesundheit:
-BMI 18 bis 25
-ap_lo < 90
-ap_hi < 140

schlechte Gesundheit:
- BMI > 25 oder BMI < 18 (über- und untergewicht)
- ap_lo > 90
- ap_hi > 140



Hier stellt sich heraus, dass der Anteil an Patient:innen mit mittelguten und schlechten Blutwerten in der Gruppe der Patient:innen mit schlechter Gesundheit leicht höher ist, als in der Gruppe der Patient:innen mit guter Gesundheit.

Überraschenderweise stellt sich beim Alkoholkonsum und beim Rauchen nur ein geringfügiger Unterschied heraus:


Alkoholkonsumenten bei guter Ge
Raucher bei guter Gesundheit
Alkoholkonsumenten bei schlecht
Raucher bei schlechter Gesundhe
Ratio
Alkohol
Rauchen
1.00
0.75
0.50
0.25
0.00
−0.5
0.0
0.5
1.0
1.5

3. Methoden

In den Abschnitten 3. und 4. der Analyse werden nun statistische Modelle zur exploration versteckter Statistischer Prozesse genutzt. Diese werden im folgenden Vorgestellt.

Hypothesentest

Wir betrachten das statistische Modell

$\Omega = \{1, \ldots, 11753\},\ \mathcal{F} = \mathcal{P}(\otimes),\ P_\theta = Binomial(11753, \theta)_{\theta \in \Theta}$

zum Testproblem: $H_0 : \theta \leq 0.4997,\ H_1 : \theta > 0.4997$

Ein gleichmäßig bester Test ist von der Form : $D(x) = 1_K(x)$ *mit* $K = \{c, \ldots, 11753\}$ *der Ablehnungsbereich* [3].

Der R-Befehl

```
qbinom(0.01, size = 11753, prob = 0.4997, lower.tail = FALSE)
```

```
## [1] 5999
```

Damit ist $K = \{5999, ..., 11753\}$ der Ablehnungsbereich des Test zum Signifikanzniveau 1%.

Die Anzahl der Herzkranken in der Gruppe mit schlechter Gesundheit beträgt

```
## [1] 9945
```

Damit ist die Nullhypothese (bei weitem) verworfen.

Probit Regression

Aufgrund der vorliegenden Erkenntnisse aus der Explorativen Datenanalyse, erscheint es sinnvoll die versteckte Beziehung zwischen den Variablen genauer zu untersuchen und über eine Regression in linearen Zusammenhang zu der Variable *Cardio* zu stellen.
Die binäre Variable *Cardio* liefert nicht die gewünschte Feinheit, durch welche Änderungen in der abhängigen Größe auf die unabhängigen Einflussvariablen zurückgespielt werden können. Ohne Einschränkung der Aussagekraft wird daher statt der *Cardio* Variable, ihr, von den Einflussvariablen, abhängiger Erwartungswert beschrieben.
Jedoch besteht für die klassische lineare Regression weiterhin große Fehleranfälligkeit. Zum einen führt der beschriebene Lineare Zusammenhang zu erwarteten Wahrscheinlichkeiten die über 1 liegen und zum anderen sind aus der Vorlesung bekannte Tests zur Genauigkeit der Regression wenig Aussagekräftig.
Um das erstere Problem zu umgehen findet die Probit Regression Anwendung. Hierbei wird der bedingte Erwartungswert der abhängigen Variable über die Verteilungsfunktion der Normalverteilung berechnet. Die Regression modeliert dabei den z-Wert, welcher schließlich über den *qnorm* Befehl zu der erhaltenen Vorhersage $\Phi(z)$ führt. Die hierfür verwendete Formel lautet [4]:

$$P(Y = 1|X_1, X_2, ..., X_k) = \Phi(\beta_0 + \beta_1 * X_1 + ... + \beta_k * X_k)$$

für das Modell

$$Y = \beta_0 + \beta_1 * X_1 + ... + \beta_k * X_k + u$$

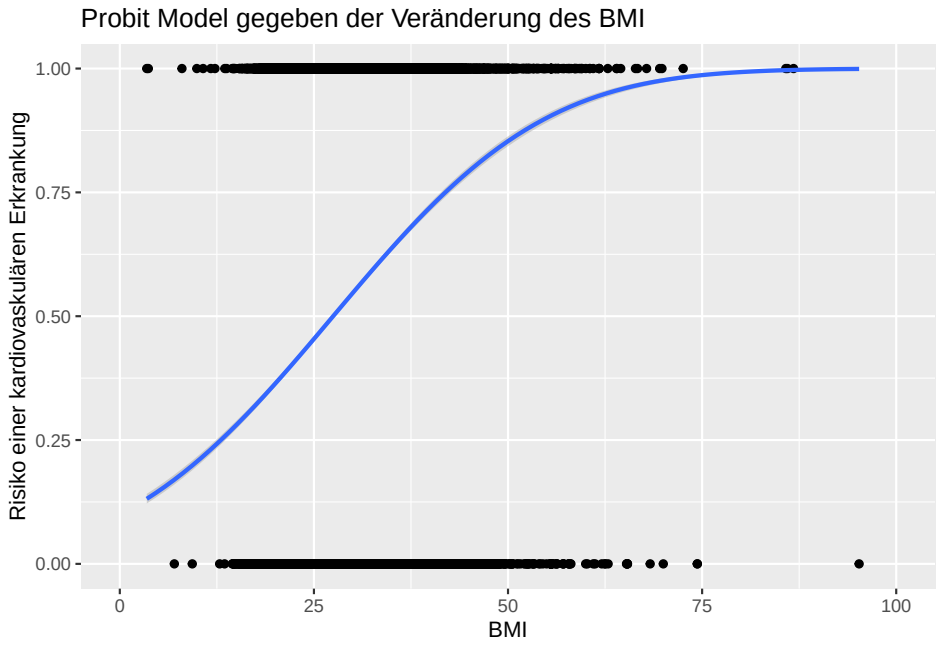
[3] Vorlesungsskript
[4] https://www.econometrics-with-r.org/11.2-palr.html

Dabei wird angenommen, dass die bedingten Erwartungswerte Normalverteilt sind. Die Unabhängigkeit der Variablen ist durch die Zufällige Probe der Personen gegeben. Die Deutung der berechneten Koeffizienten wird intuitiv wenn man die Veränderung des z-Wertes bei isolierter Veränderung der jeweiligen unabhängigen Variable betrachtet.

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

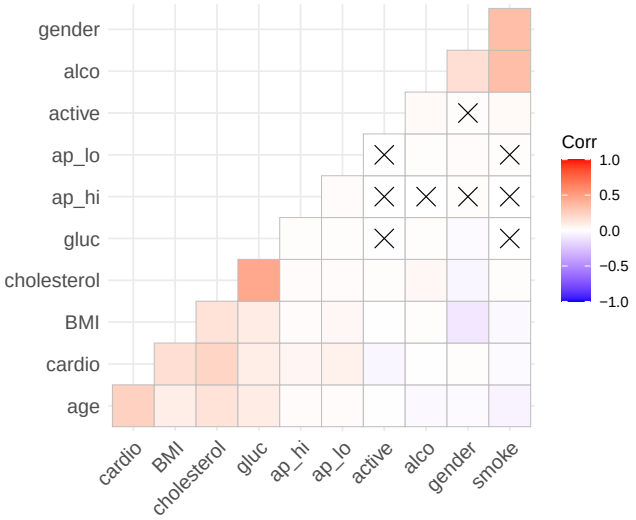
```
##
## Call:
## glm(formula = cardio ~ age + gender + BMI + ap_hi + ap_lo + cholesterol +
##     gluc + smoke + alco + active, family = binomial(link = "probit"),
##     data = df_1, control = list(maxit = 100))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.4904  -0.9697  -0.0331   0.9997   7.1317
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.621e+00  6.122e-02 -91.819  < 2e-16 ***
## age          3.290e-02  7.754e-04  42.434  < 2e-16 ***
## gender       4.377e-02  1.136e-02   3.851 0.000117 ***
## BMI          1.748e-02  9.919e-04  17.618  < 2e-16 ***
## ap_hi        2.441e-02  3.519e-04  69.354  < 2e-16 ***
## ap_lo        1.490e-04  3.502e-05   4.254 2.10e-05 ***
## cholesterol  3.127e-01  8.859e-03  35.295  < 2e-16 ***
## gluc        -6.769e-02  1.015e-02  -6.671 2.53e-11 ***
## smoke       -7.803e-02  1.995e-02  -3.911 9.20e-05 ***
## alco        -9.817e-02  2.413e-02  -4.069 4.72e-05 ***
## active      -1.281e-01  1.272e-02 -10.073  < 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: 97041  on 69999  degrees of freedom
## Residual deviance: 81585  on 69989  degrees of freedom
## AIC: 81607
##
## Number of Fisher Scoring iterations: 13
```

Die Zusammenfassung der Probit Regression wird mit der Funktion *summary* erzeugt. Besonders wichtig sind hierzu die werte der jeweiligen Koeffizienten sowie der p-Werte der Hypothesentest entgegen der Nullhypothese "der Koeffizient ist gleich 0". Dies lässt auf die signifikanz des Koeffizienten schließen.
Die Visualisierung der Prognosen bezüglich der abhängigen Variabel BMI verdeutlichen die unterschiede welche eine Probit Regression im Vergleich zu einer Linearen Regression haben.

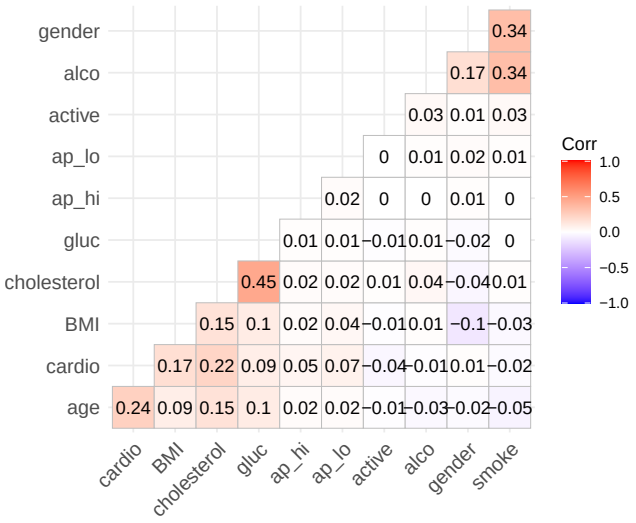


Korrelationspyramide

Die Korrelationspyramide zeigt die jeweiligen Korrelationen der untersuchten Variablen auf. Dabei wird besonderes Augenmerk auf die Korrelation der unabhängigen Variablen mit der abhängigen Variablen gesetzt, da diese die Aussagekraft des Models untermalen. Die berechnung und visualisierung übernimmt hierbei das ggcorrplot package [5] welches zur Berechnung der Relevanten Werte auf die Funktion *cor.test* aus dem Packet *stats* zurückgreift. Die aufgezeigten Werte gleichen entsprechend dem empirischen Korrelationskoeffizienten [6]. Mittels eines p.Test werden die Werte in einem weiteren Plot gegen die Nullhypothese $H.0 := Der\,Korrelationskoeffizient\,ist\,gleich\,0$ getestet. Die Werte die hierbei nicht verworfen werden können, sind dabei mit einem X gekennzeichnet.



[5] "https://cran.r-project.org/web/packages/corrplot/index.html"
[6] Definition 6.38 aus dem Vorlesungsskript



Die folgende Tabelle veranschaulicht die Abweichung der Prognose von den reellen Daten. Aus der Vorlesung ist bekannt, dass der empirische Mittelwert annähernd Normalverteilt ist, entsprechend läuft der Vergleich nach dem Prinzip des mittleren quadratischen Fehlers mit $n = 1$.
Es wird ersichtlich, dass das Model die Daten gut erklärt. Lediglich bei der Anwendung des Models auf die Gruppe mit Guter Gesundheit prognostiziert das Model ein grundsätzlich höheres Risiko.

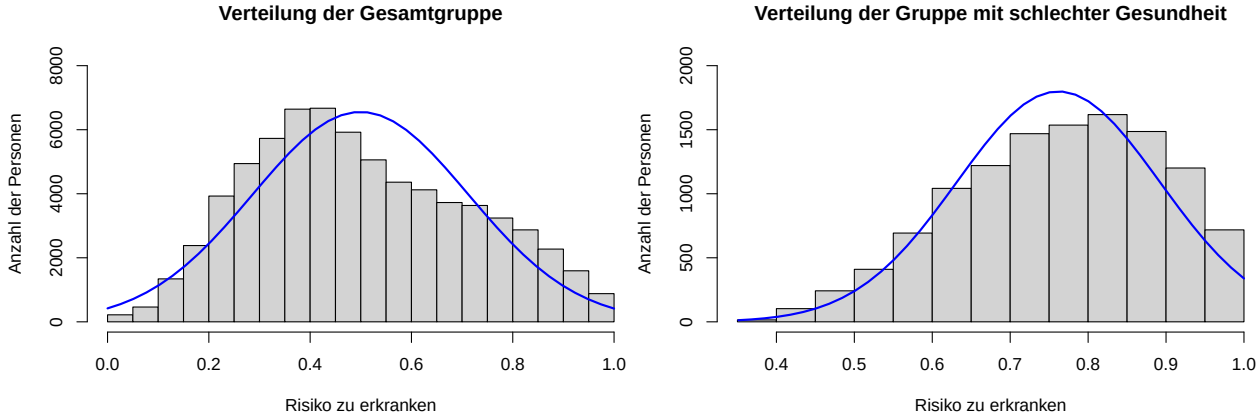
```
##   Betrachtete Gruppe Wahrer Median Model Median Prozentuale Abweichung
## 1              Total     0.4997000    0.4993872             0.06259775
## 2                Gut     0.2930277    0.3507572            19.70104667
## 3           Schlecht     0.8461669    0.7618140             9.96882501
```

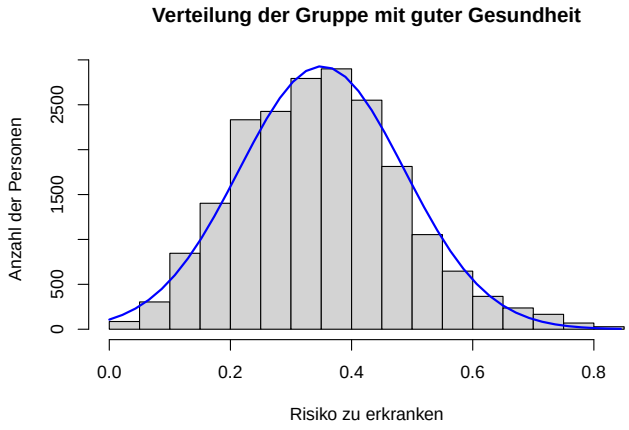
4. Ergebnisse und Schlussfolgerung

Die p-Werte liefern, wie bereits erwähnt, eine Aussage bezüglich der Signifikanz der berechneten β-Koeffizienten. Intuitiv folgt daraus, dass eine Veränderung in der unabhängigen Variable über die betrachtete unabhängige Variable erklärt werden kann.
Aus der summary des Models wird sofort ersichtlich, dass jede der gelisteten Variable signifikant zu einem Nivea sind welches kleiner als $\alpha = 0,1\%$ ist. Es lässt sich also schließen, dass zu jeder abhängigen Variable, die Bewegung zur Veränderung des z-Wertes führt.
Ähnliches lässt sich auch von der Korrelationspyramide ableiten. Hier weist keine Variable eine insignifikante korrelation mit der unabhängigen Variable auf. Auffallend ist dabei, dass die Variablen: BMI und $Cholesterol$ am stärksten mit $Cardio$ korrelieren. Interessant ist auch, dass das Geschlecht ebenfalls einen Einfluss auf das Risiko hat an kardiovaskulären Krankheiten zu erleiden. Dies deckt sich mit der Datenanalyse aus Teil 2 (überprüfen).

```
##                 Estimate   Std. Error    z value      Pr(>|z|)
## (Intercept) -5.620703942 6.121536e-02 -91.818525  0.000000e+00
## age          0.032903590 7.753996e-04  42.434363  0.000000e+00
## gender       0.043765709 1.136361e-02   3.851390  1.174494e-04
## BMI          0.017475063 9.918799e-04  17.618124  1.788390e-69
## ap_hi        0.024405941 3.519049e-04  69.353804  0.000000e+00
## ap_lo        0.000149002 3.502412e-05   4.254267  2.097345e-05
## cholesterol  0.312691195 8.859331e-03  35.295125 6.975957e-273
## gluc        -0.067687731 1.014598e-02  -6.671381  2.534072e-11
## smoke       -0.078034595 1.995354e-02  -3.910815  9.198509e-05
## alco        -0.098173086 2.412572e-02  -4.069229  4.716904e-05
## active      -0.128113952 1.271860e-02 -10.072961  7.275419e-24
```

Um Aussagen über die Genauigkeit des Tests treffen zu können, werden die Erkentnisse der explorativen Datenanalyse genommen und mit den Charakteristiken der geschätzten Werte verglichen.
Die Histogramme wurden entsprechen der geschilderten Gruppenaufteilung des Datensets angefertigt und darüber eine ideale Normalverteilung (in blau) gelegt, mit jeweiliger verschiedener Varianzen und Erwartungswerten. Es lässt sich deutlich erkennen, dass die Personen mit schlechter Gesundheit ein höheres Risiko und die mit guter Gesundheit ein niedrigere Risiko aufweisen, zu erkranken. Auch interessant ist, dass eine höhere Menge an Personen in unserem Gesamt Datenset ein niedrigeres Risiko aufweisen. Der Erwartungswert liegt bei $\sim 50\%$.




Verteilung der Gruppe mit guter Gesundheit
Anzahl der Personen
0
500
1500
2500
0.0
0.2
0.4
0.6
0.8
Risiko zu erkranken