# Vertiefende statistische Verfahren

3. Übungsblatt SS 2024

## Allgemeine Information

Alle Aufgaben sind mit R zu lössen, wenn nicht explizit anders angegeben. Die Berechnungen sollen nachvollziehbar und dokumentiert sein. Um die vollständige Punktezahl zu erreichen, müssen alle Ergebnisse und Fragen entsprechend interpretiert bzw. beantwortet werden. Code alleine ist nicht ausreichend! Die Abgabe erfolgt über Moodle entsprechend der Abgaberichtlinien als pdf und Rmd File. Bitte inkludieren Sie namentlich alle beteiligten Gruppenmitglieder sowohl im Bericht als auch im Source Code. Die jeweiligen Datensätze die für diese Übung relevant sind finden Sie ebenfalls in Moodle.

## 1 Einfaktorielle ANOVA (händisch) [2P]

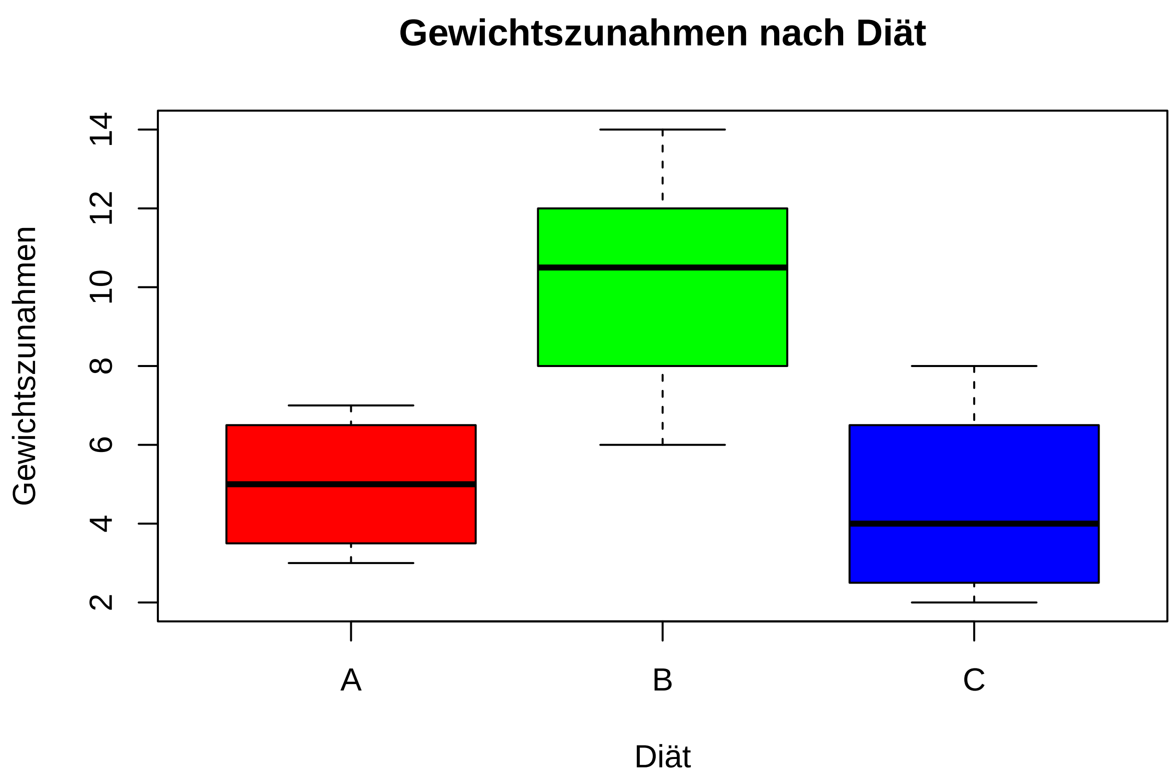
Ein Sportwissenschafter möchte sehen, ob es einen Unterschied in der Gewichtszunahme von Sportlern gibt, die einer von drei speziellen Diäten folgen. Die Athleten werden nach dem Zufallsprinzip drei Gruppen zugewiesen und unterziehen sich für 6 Wochen der jeweiligen Diät. Die Gewichtszunahmen (in Pfund) sind angegeben. Nehmen Sie an, dass die Gewichtszunahmen normalverteilt sind und die Varianzen gleich sind. Kann der Sportwissenschafter bei einem Signifikanzniveau von 0,05 schlussfolgern, dass es einen Unterschied zwischen den Diäten gibt? Führen Sie die ANOVA händisch durch und überprüfen Sie das Ergebnis mit `summary(aov(...))`.

| Diät | Messwerte |
|---|---|
| A | 3, 6, 7, 4 |
| B | 10, 12, 11, 14, 8, 6 |
| C | 8, 3, 2, 5 |

```r
# ANOVA händische durchführung
gains <- list(
  A = c(3, 6, 7, 4),
  B = c(10, 12, 11, 14, 8, 6),
  C = c(8, 3, 2, 5))

# Überblick über die Daten / Farben nach Diet
str(gains)
```

```
## List of 3
##  $ A: num [1:4] 3 6 7 4
##  $ B: num [1:6] 10 12 11 14 8 6
##  $ C: num [1:4] 8 3 2 5
```

```r
boxplot(gains, col = c("red", "green", "blue"),
        ylab = "Gewichtszunahmen",
        xlab = "Diät",
        main = "Gewichtszunahmen nach Diät")
```




```r
# Hypothesen
# H0: mu_a = mu_b = mu_c
# H1: mu_i != mu_j (für mindestens ein Paar i,j)

# Berechnung der Mittelwerte
mean_total = mean(unlist(gains))

# Berechnung der Quadratsummen
qs_between <- 0
for (group in gains) {
  group_mean <- mean(group)
  n <- length(group)
  qs_between <- qs_between + n * (group_mean - mean_total)^2
}

paste("Die Quadratsumme zwischen den Gruppen beträgt: ", round(qs_between,2))
```

```
## [1] "Die Quadratsumme zwischen den Gruppen beträgt:  101.1"
```

```r
qs_within <- 0
for (group in gains) {
  group_mean <- mean(group)
  for (value in group) {
    qs_within <- qs_within + (value - group_mean)^2
  }
}

paste("Die Quadratsumme innerhalb der Gruppen beträgt: ", round(qs_within,2))
```

```
## [1] "Die Quadratsumme innerhalb der Gruppen beträgt:  71.83"
```

```r
qs_total <- sum((unlist(gains) - mean_total)^2)

paste("Die Quadratsumme insgesamt beträgt: ", round(qs_total,2))
```

```
## [1] "Die Quadratsumme insgesamt beträgt:  172.93"
```

```r
# Prüfung der Quadratsummen
round(qs_total, 2)==round(qs_between + qs_within, 2)
```

```
## [1] TRUE
```

```r
# Überprüfen der Vorraussetzungen

# Normalverteilung
for (group in gains) {
  print(shapiro.test(group))
}
```

```
##
##  Shapiro-Wilk normality test
##
## data:  group
## W = 0.94971, p-value = 0.7143
##
##
##  Shapiro-Wilk normality test
##
## data:  group
## W = 0.98901, p-value = 0.9866
##
##
##  Shapiro-Wilk normality test
##
## data:  group
## W = 0.94563, p-value = 0.6889
```

```r
# Varianzhomogenität (Barlet-Test)
bartlett.test(gains)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  gains
## Bartlett's K-squared = 0.61192, df = 2, p-value = 0.7364
```

```r
# Freiheitsgrade
df_between <- length(gains) - 1
df_within <- length(unlist(gains)) - length(gains)
df_total <- length(unlist(gains)) - 1

# Berechnung der Varianzen (Mittlere Quadratsummen)
ms_total <- qs_total / df_total
ms_between <- qs_between / df_between
ms_within <- qs_within / df_within

# Berechnung des F-Wertes
f_value <- ms_between / ms_within
paste("Der F-Wert beträgt: ", round(f_value, 2))
```

```
## [1] "Der F-Wert beträgt:  7.74"
```

```r
# kritischer F-Wert aus Tabelle
alpha <- 0.05
f_critical <- qf(alpha, df_between, df_within,lower.tail = F)
paste("Der kritische F-Wert beträgt: ", round(f_critical, 2))
```

```
## [1] "Der kritische F-Wert beträgt:  3.98"
```

```r
# Ergebnis
if (f_value > f_critical) {
  print("H0 wird verworfen")
} else {
  print("H0 wird nicht verworfen")
}
```

```
## [1] "H0 wird verworfen"
```

```r
# p-Wert
p_value <- pf(f_value, df_between, df_within, lower.tail = F)
paste("Der berechnete p-Wert beträgt: ", round(p_value,5))
```

```
## [1] "Der berechnete p-Wert beträgt:  0.00797"
```

```r
# Überprüfung mit R
diet <- factor(rep(names(gains), times = sapply(gains, length)))
all_gains <- unlist(gains)

aov_result <- aov(all_gains ~ diet)
summary(aov_result)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)   
## diet         2 101.10   50.55    7.74 0.00797 **
## Residuals   11  71.83    6.53                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation: Es gibt einen signifikanten Einfluss des Faktors Diät, $F(2,11) = 7.74$; $p < 0.05$. Als Effektgröße wurde ein Eta-Quadrat von 0.58 berechnet. Das Eta-Quadrat schätzt den Anteil der Varianz, der durch den eine Variable erklärt wird. Jedoch ist der Literatur zu entnehmen, dass der Einfluss in der Varianz durch diese Maßzahl meist überschätzt wird.

# 2 Einfaktorielle ANOVA [2P]

Analysieren Sie die Daten zu niedrigem Geburtsgewicht von Neugeborenen in `birthwt_aov.xlsx`. Untersuchen Sie ob die ethnische Zugehörigkeit der Mutter (Variable `ethnic`, wobei `1 = "white"`, `2 = "black"`, `3 = "other"` ) einen Einfluss auf das Geburtsgewicht von Neugeborenen hat (Variable `bwt`). Beachten Sie dabei folgende Punkte:

i) Überprüfen Sie die Vorraussetzungen.
ii) Verwenden Sie eine geeignete grafische Darstellung der Daten / Ergebnisse.
iii) Gibt es signifikante Unterschiede zwischen den Gruppen, wenn ja zwischen welchen Gruppen?
iv) Achten Sie auf eine "statistisch korrekte" Formulierung des Ergebnisses.

```r
# Daten einlesen
library(readxl)
birthwt <- read_excel("UE3 Daten/birthwt_aov.xlsx")

# Faktor für ethnische Zugehörigkeit
birthwt$ethnic <- factor(birthwt$ethnic, levels = c(1,2,3), labels = c("white", "black", "other"))

# Überblick über die Daten
str(birthwt)
```

```
## tibble [189 x 10] (S3: tbl_df/tbl/data.frame)
##  $ low   : num [1:189] 0 0 0 0 0 0 0 0 0 0 ...
##  $ age   : num [1:189] 19 33 20 21 18 21 22 17 29 26 ...
##  $ lwt   : num [1:189] 182 155 105 108 107 124 118 103 123 113 ...
##  $ ethnic: Factor w/ 3 levels "white","black",..: 2 3 1 1 1 3 1 3 1 1 ...
##  $ smoke : num [1:189] 0 0 1 1 1 0 0 0 1 1 ...
##  $ ptl   : num [1:189] 0 0 0 0 0 0 0 0 0 0 ...
##  $ ht    : num [1:189] 0 0 0 0 0 0 0 0 0 0 ...
##  $ ui    : num [1:189] 1 0 0 1 1 0 0 0 0 0 ...
##  $ ftv   : num [1:189] 0 3 1 2 0 0 1 1 1 0 ...
##  $ bwt   : num [1:189] 2523 2551 2557 2594 2600 ...
```

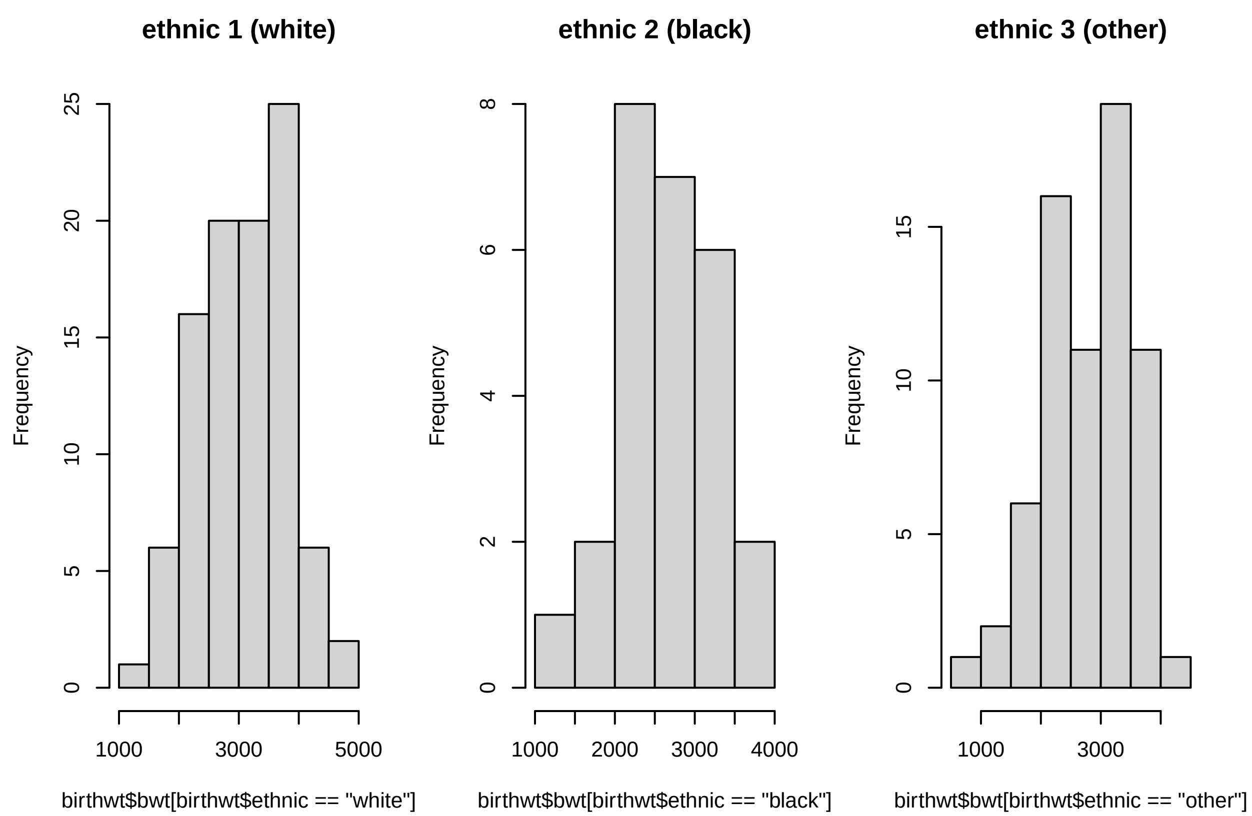
```r
table(birthwt$ethnic)
```

```
## 
## white black other 
##    96    26    67
```

```r
summary(birthwt$bwt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     709    2414    2977    2945    3487    4990
```

```r
# Hypothesen
# H0: mu_white = mu_black = mu_other
# H1: mu_i != mu_j (für mindestens ein Paar i,j)

# Überprüfen der Vorraussetzungen

# Normalverteilung (für jede Gruppe)
par(mfrow = c(1,3))
hist(birthwt$bwt[birthwt$ethnic == 'white'], main = "ethnic 1 (white)")
hist(birthwt$bwt[birthwt$ethnic == 'black'], main = "ethnic 2 (black)")
hist(birthwt$bwt[birthwt$ethnic == 'other'], main = "ethnic 3 (other)")
```




```r
ethics <- unique(birthwt$ethnic)
# Shapiro-Wilk-Test
for (ethic in ethics) {
  print(ethic)
  print(shapiro.test(birthwt$bwt[birthwt$ethnic == ethic]))
}
```

```
## [1] "black"
```

```
##
##  Shapiro-Wilk normality test
##
## data:  birthwt$bwt[birthwt$ethnic == ethic]
## W = 0.97696, p-value = 0.8038
##
## [1] "other"
##
##  Shapiro-Wilk normality test
##
## data:  birthwt$bwt[birthwt$ethnic == ethic]
## W = 0.97537, p-value = 0.2046
##
## [1] "white"
##
##  Shapiro-Wilk normality test
##
## data:  birthwt$bwt[birthwt$ethnic == ethic]
## W = 0.98727, p-value = 0.4861
```

Ein visuelle Überprüfung der Normalverteilung zeigt, dass diese für alle Gruppen gegeben ist. Der Shapiro-Wilk-Test bestätigt diese Annahme, weshalb die Annahme der Normalverteilung erfüllt ist.

```r
# Varianzhomogenität (Heteroskedastizität)

# Levene-Test
library(car)
leveneTest(bwt ~ ethnic, data = birthwt)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.4684 0.6267
##       186
```

```r
# Bartlett-Test
bartlett.test(birthwt$bwt ~ birthwt$ethnic)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  birthwt$bwt by birthwt$ethnic
## Bartlett's K-squared = 0.65952, df = 2, p-value = 0.7191
```

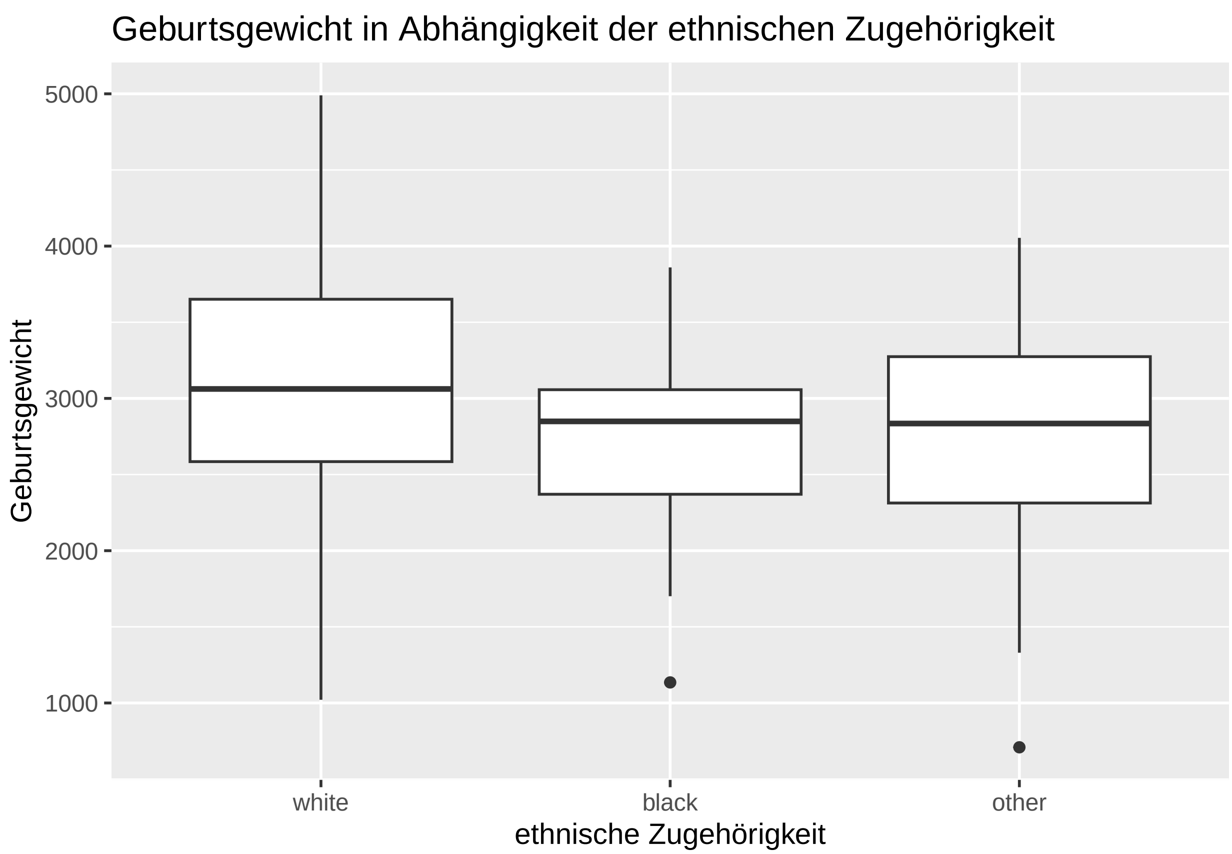
Sowohl der Levene- als auch der Bartlett-Test zeigen, dass die Varianzhomogenität gegeben ist.

Alle wesentlichen Vorraussetzungen für die ANOVA sind somit gegeben und sie kann ohne Bedenken durchgeführt werden.

```r
# ANOVA
aov_result <- aov(bwt ~ ethnic, data = birthwt)
summary(aov_result)
```

```
##               Df   Sum Sq Mean Sq F value  Pr(>F)
## ethnic         2  5015725 2507863   4.913 0.00834 **
## Residuals    186 94953931  510505
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Darstellung der Ergebnisse
library(ggplot2)
ggplot(birthwt, aes(x = ethnic, y = bwt)) +
  geom_boxplot() +
  labs(title = "Geburtsgewicht in Abhängigkeit der ethnischen Zugehörigkeit",
       x = "ethnische Zugehörigkeit",
       y = "Geburtsgewicht")
```




Die ethnische Zugehörigkeit hat einen signifikanten Einfluss auf das Geburtsgewicht von Neugeborenen, F(2,186) = 4.91, p = 0.008.

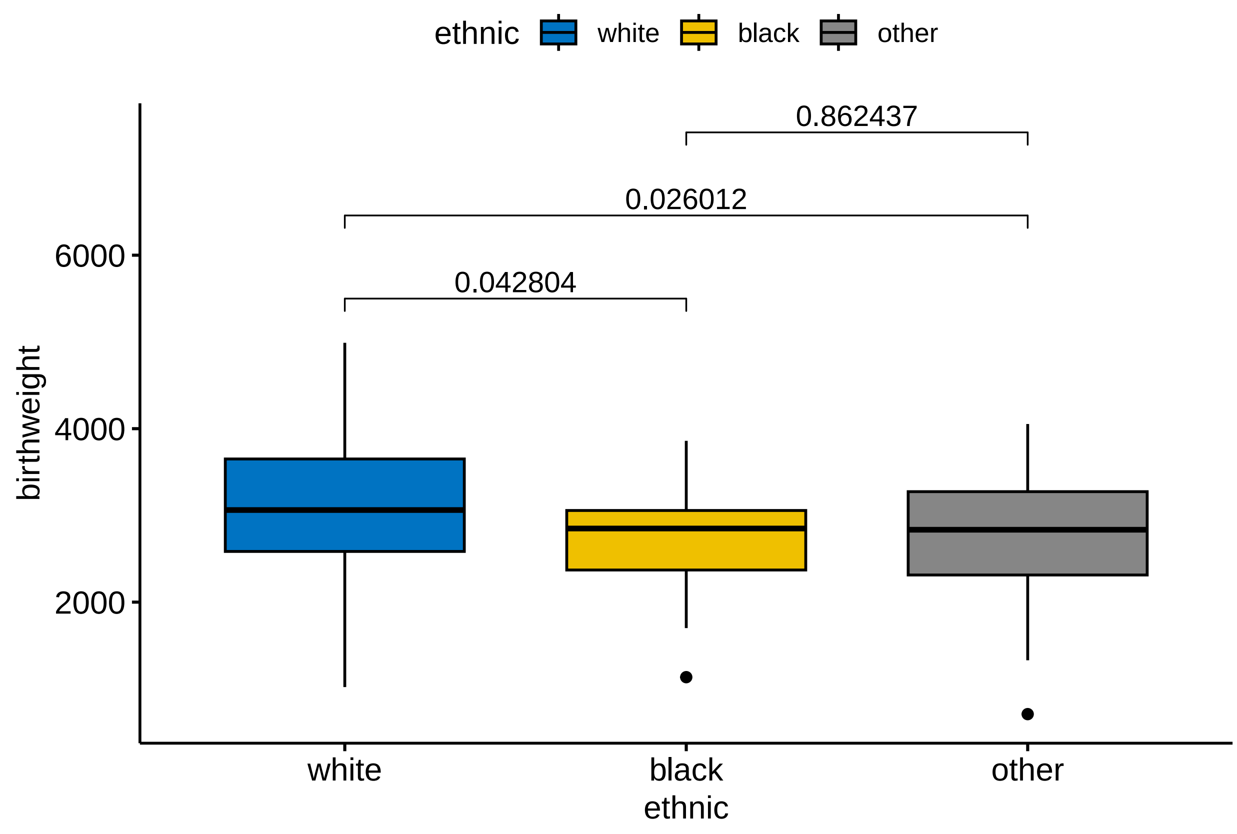
Um herauszufinden zwischen welchen Gruppen signifikante Unterschiede bestehen, wird nun eine Post-Hoc-Analyse durchgeführt. Als geeigneter Test wird der Tukey HSD-Test verwendet.

```r
# Post-Hoc-Test
pht <- TukeyHSD(aov_result)
print(pht)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
```

```
##
## Fit: aov(formula = bwt ~ ethnic, data = birthwt)
##
## $ethnic
##                   diff       lwr        upr     p adj
## black-white -383.02644 -756.2363  -9.816581 0.0428037
## other-white -297.43517 -566.1652 -28.705095 0.0260124
## other-black   85.59127 -304.4521 475.634630 0.8624372
```

```r
# Darstellung der Ergebnisse
library(ggpubr)

stat.test <- tibble::tribble(
  ~group1, ~group2,   ~p.adj,
  "white",     "black", round(pht$ethnic[1,4],6),
  "white",     "other", round(pht$ethnic[2,4],6),
  "black",     "other", round(pht$ethnic[3,4],6))

# Boxplots (inkl. p-Werte)
ggboxplot(birthwt, x = "ethnic", y = "bwt",
          fill = "ethnic",
          palette = "jco",
          ylab = "birthweight") +
  stat_pvalue_manual(stat.test,
                     y.position = 5500, step.increase = 0.2,
                     label = "p.adj")
```



Der Tukey HSD-Test zeigt, dass es signifikante Unterschiede zwischen den ethnischen Gruppen "white" und "black" ($p = 0.043$) sowie "white" und "other" ($p = 0.026$) gibt. Zwischen den Gruppen "black" und "other" besteht hingegen kein signifikanter Unterschied ($p = 0.862$). Konkret wiegen schwarze Neugeborene durchschnittlich 383,03 Gramm weniger als weiße. Neugeborene aus anderen ethnischen Gruppen weisen ebenfalls ein signifikant geringeres Geburtsgewicht auf als weiße Neugeborene, mit einem durchschnittlichen Unterschied von 297,44 Gramm. Im Vergleich zwischen schwarzen und anderen ethnischen Gruppen gibt es jedoch keinen signifikanten Unterschied im Geburtsgewicht, wie die Analyse mit einem p-Wert von 0,8624 zeigt.

# 3 Einfaktorielle ANOVA [2P]

Verwenden Sie den bereits bekannten `Framingham.sav` Datensatz. Analysieren Sie ob es Unterschiede im BMI in Abhängigkeit von der Schulbildung gibt. Achten Sie auf Ausreißer und fehlende Daten (`NaN, NA's`).

i) Überprüfen Sie die Vorraussetzungen.
ii) Verwenden Sie eine geeignete grafische Darstellung der Daten / Ergebnisse.
iii) Gibt es signifikante Unterschiede zwischen den Bildungsstufen, wenn ja zwischen welchen Stufen?
iv) Achten Sie auf eine "statistisch korrekte" Formulierung des Ergebnisses.

# 4 Mehrfaktorielle ANOVA [2P]

Sie führen eine Studie bezüglich des Einflusses unterschiedlicher Diäten und Aktivitätslevels auf den Erfolg bei der Gewichtsabnahme durch. Jedem Probanden wird eine Diät und ein Aktivitätslevel zugewiesen und die Differenz zum Ausgangsgewicht nach 2 Monaten gemessen (in kg). Die Daten der Studie sind in `weightloss.sav` zusammengefasst. Analysieren Sie ob die beiden Faktoren einen Einfluss auf das Gewicht haben.

i) Überprüfen Sie die Vorraussetzungen.
ii) Verwenden Sie eine geeignete grafische Darstellung der Daten / Ergebnisse.
iii) Gibt es signifikante Haupteffekte sowie Interaktionseffekte (inkl. Interaktions-Plot)?
iv) Achten Sie auf eine "statistisch korrekte" Formulierung des Ergebnisses.

```r
# Laden der Daten
library(foreign)
```

```
## Warning: Paket 'foreign' wurde unter R Version 4.3.2 erstellt
```

```r
weightloss <- read.spss("UE3 Daten/weightloss.sav", to.data.frame = TRUE)

str(weightloss)
```

```
## 'data.frame':    180 obs. of  4 variables:
##  $ id      : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ diet    : Factor w/ 3 levels "None","Atkins",..: 3 1 1 2 3 1 1 1 1 1 ...
##  $ exercise: Factor w/ 3 levels "None","30 minutes per day",..: 3 2 3 2 3 2 1 2 2 1 ...
##  $ wloss   : num  13.96 -2.12 6.92 0.68 4.68 ...
##  - attr(*, "variable.labels")= Named chr [1:4] "" "Diet assigned to participant" "Exercise level ass
##   ..- attr(*, "names")= chr [1:4] "id" "diet" "exercise" "wloss"
##  - attr(*, "codepage")= int 65001
```

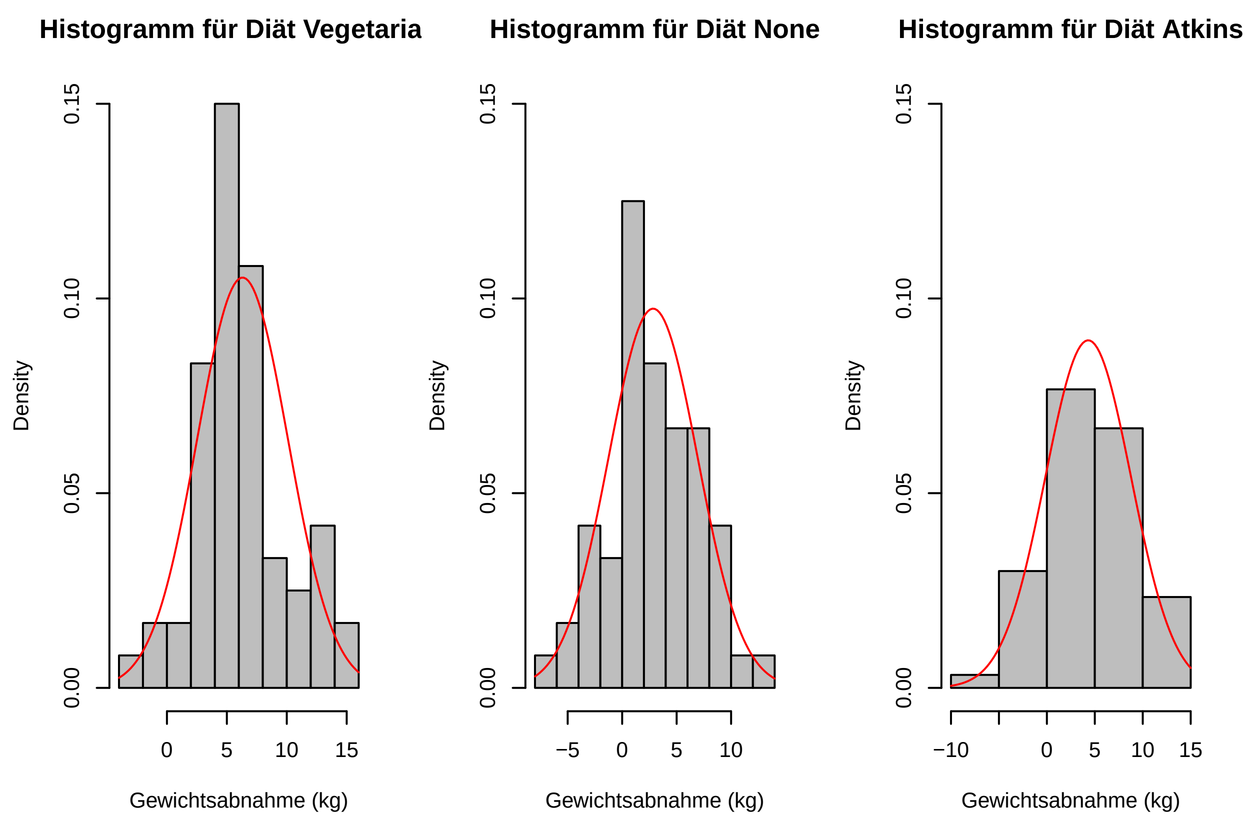
```r
summary(weightloss)
```

```
##        id                 diet                    exercise      wloss
##  Min.   :  1.00   None      :60   None              :60   Min.   :-6.440
##  1st Qu.: 45.75   Atkins    :60   30 minutes per day:60   1st Qu.: 1.220
##  Median : 90.50   Vegetarian:60   60 minutes per day:60   Median : 4.640
##  Mean   : 90.50                                           Mean   : 4.488
##  3rd Qu.:135.25                                           3rd Qu.: 7.320
##  Max.   :180.00                                           Max.   :14.760
```

```r
# Hypothesen

# H0 (Diät): mu_noDiet = mu_vegetarian = mu_atkins
# H0 (Activität): mu_none = mu_30 = mu_60
# H0 (Interaktion): Alle Mittelwertunterschiede werden durch die Haupteffekte erklärt.
```

Nun werden die Vorraussetzungen für die ANOVA überprüft.

```r
# Überprüfen der Vorraussetzungen

# -- für Diet-Gruppen
unique_diets <- unique(weightloss$diet)

# Normalverteilung
par(mfrow = c(1, 3))
for(diet in unique_diets) {
  # Daten für die spezifische Diät
  diet_data <- weightloss$wloss[weightloss$diet == diet]

  # Histogramm erstellen
  hist(diet_data, probability = TRUE,
       main = paste("Histogramm für Diät", diet),
       xlab = "Gewichtsabnahme (kg)",
       col = "gray", border = "black",
       ylim = c(0, 0.15))

  # Normalverteilungskurve hinzufügen
  curve(dnorm(x, mean = mean(diet_data), sd = sd(diet_data)), add = TRUE, col = "red")
}
```

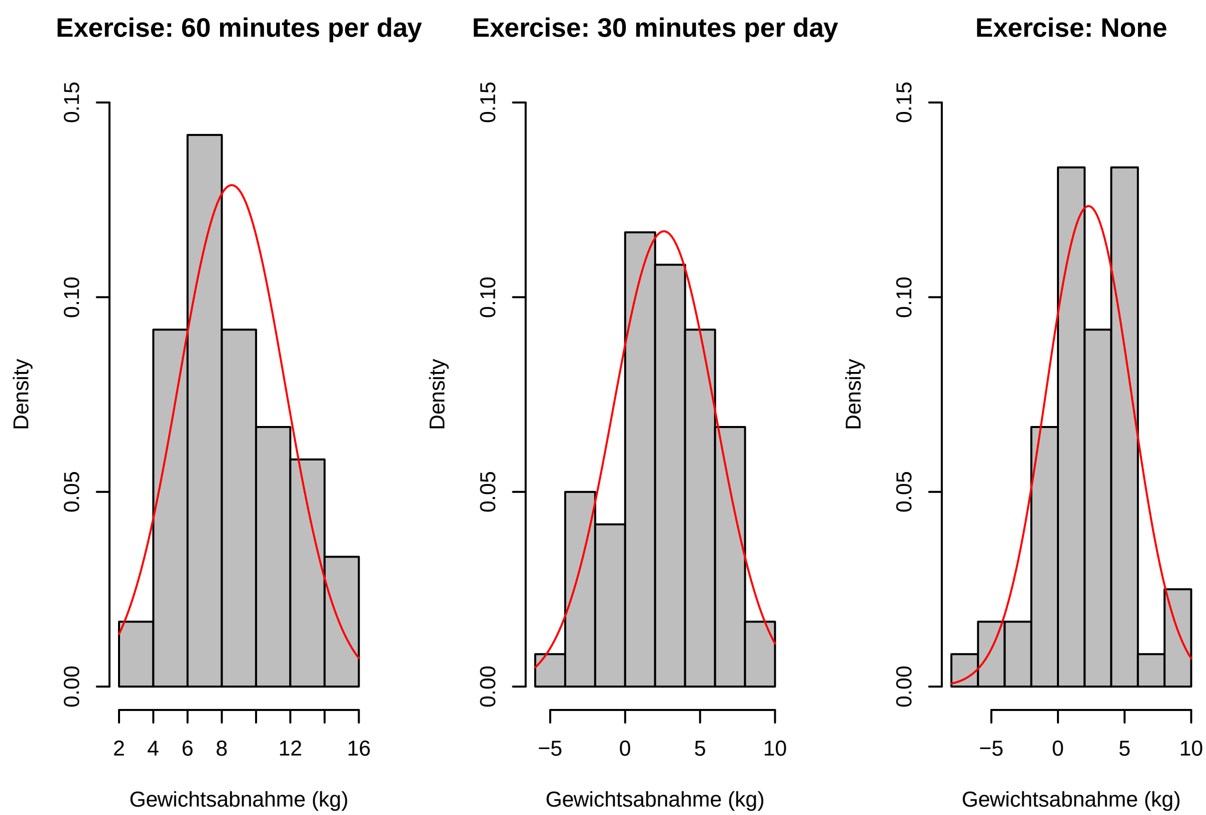



```r
# Shapiro-Wilk-Test
for (diet in unique_diets) {
  print(shapiro.test(weightloss$wloss[weightloss$diet == diet]))
}
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weightloss$wloss[weightloss$diet == diet]
## W = 0.96877, p-value = 0.1271
##
##
##  Shapiro-Wilk normality test
##
## data:  weightloss$wloss[weightloss$diet == diet]
## W = 0.98733, p-value = 0.7893
##
##
##  Shapiro-Wilk normality test
##
## data:  weightloss$wloss[weightloss$diet == diet]
## W = 0.98252, p-value = 0.5436
```

```r
# -- für Activity-Gruppen
unique_activities <- unique(weightloss$exercise)
```

```r
# Normalverteilung
par(mfrow = c(1, 3))
for(activity in unique_activities) {
  # Daten für die spezifische Aktivität
  activity_data <- weightloss$wloss[weightloss$exercise == activity]

  # Histogramm erstellen
  hist(activity_data, probability = TRUE,
       main = paste("Exercise:", activity),
       xlab = "Gewichtsabnahme (kg)",
       col = "gray", border = "black",
       ylim = c(0, 0.15))

  # Normalverteilungskurve hinzufügen
  curve(dnorm(x, mean = mean(activity_data), sd = sd(activity_data)), add = TRUE, col = "red")
}
```




```r
# Shapiro-Wilk-Test
for (activity in unique_activities) {
  print(shapiro.test(weightloss$wloss[weightloss$exercise == activity]))
}
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  weightloss$wloss[weightloss$exercise == activity]
## W = 0.96325, p-value = 0.06792
##
##
##  Shapiro-Wilk normality test
##
## data:  weightloss$wloss[weightloss$exercise == activity]
## W = 0.98855, p-value = 0.8468
##
##
##  Shapiro-Wilk normality test
##
## data:  weightloss$wloss[weightloss$exercise == activity]
## W = 0.98445, p-value = 0.6417
```

```r
# Varianzhomogenität (Heteroskedastizität)

# Levene-Test
leveneTest(wloss ~ diet, data = weightloss)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  1.4859 0.2291
##       177
```

```r
leveneTest(wloss ~ exercise, data = weightloss)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.3179 0.7281
##       177
```

```r
leveneTest(wloss ~ diet * exercise, data = weightloss)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   8  0.7178 0.6756
##       171
```

```r
# Bartlett-Test
bartlett.test(wloss ~ diet, data = weightloss)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  wloss by diet
## Bartlett's K-squared = 1.613, df = 2, p-value = 0.4464
```

```r
bartlett.test(wloss ~ exercise, data = weightloss)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  wloss by exercise
## Bartlett's K-squared = 0.54989, df = 2, p-value = 0.7596
```

Die Überprüfung der Varianzhomogenität mittels Levene-Test und Bartlett-Test ergab keine signifikanten Unterschiede in den Varianzen zwischen den Diät- und Aktivitätsgruppen. Somit erfüllen die Daten die Voraussetzung der Varianzhomogenität für die Durchführung einer mehrfaktoriellen ANOVA. Es ist somit statistisch angemessen, die Haupteffekte sowie den Interaktionseffekt der beiden Variablen zu analysieren.

```r
# Durchführung der ANOVA
summary(aov(wloss~diet+exercise,data=weightloss))
```

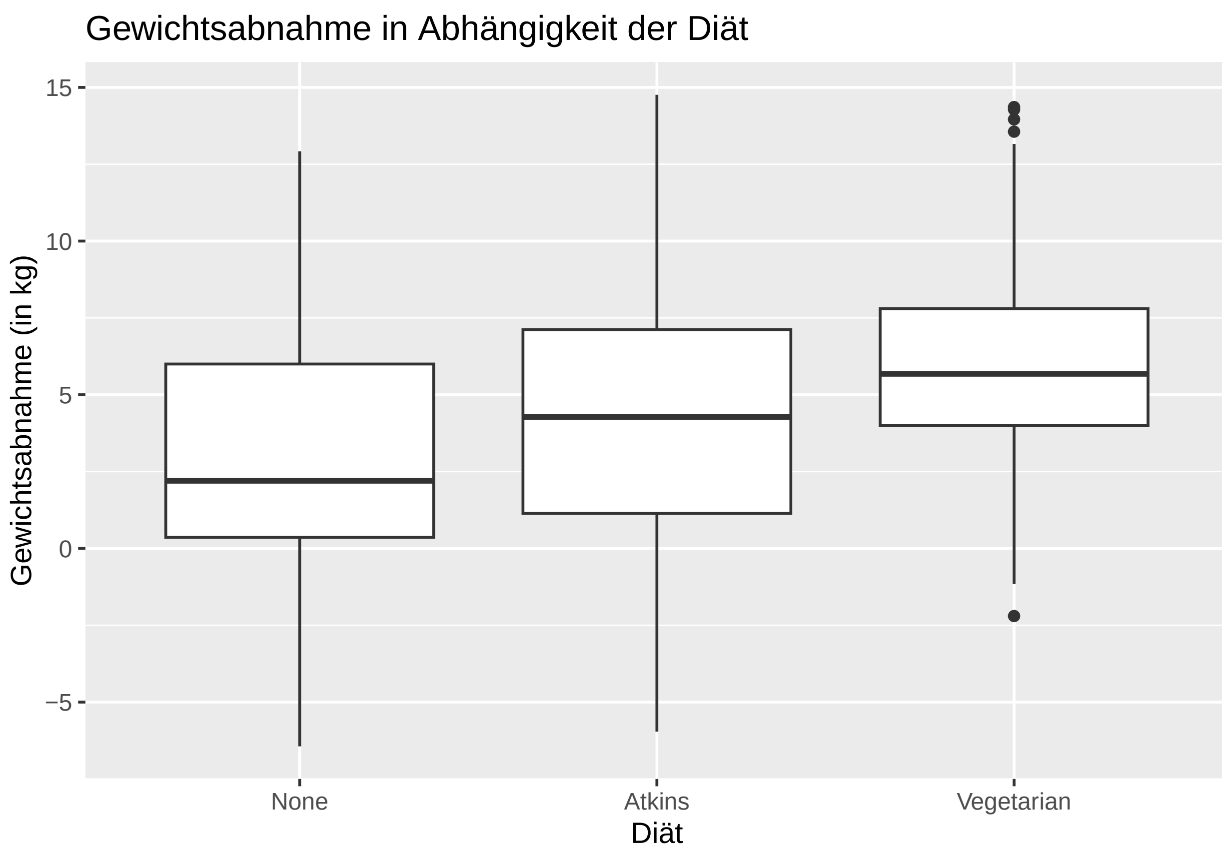
```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## diet          2  363.9   182.0   21.16 5.91e-09 ***
## exercise      2 1508.9   754.4   87.72  < 2e-16 ***
## Residuals   175 1505.2     8.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#mit Interaktionseffekte
summary(aov(wloss~diet*exercise,data=weightloss))
```

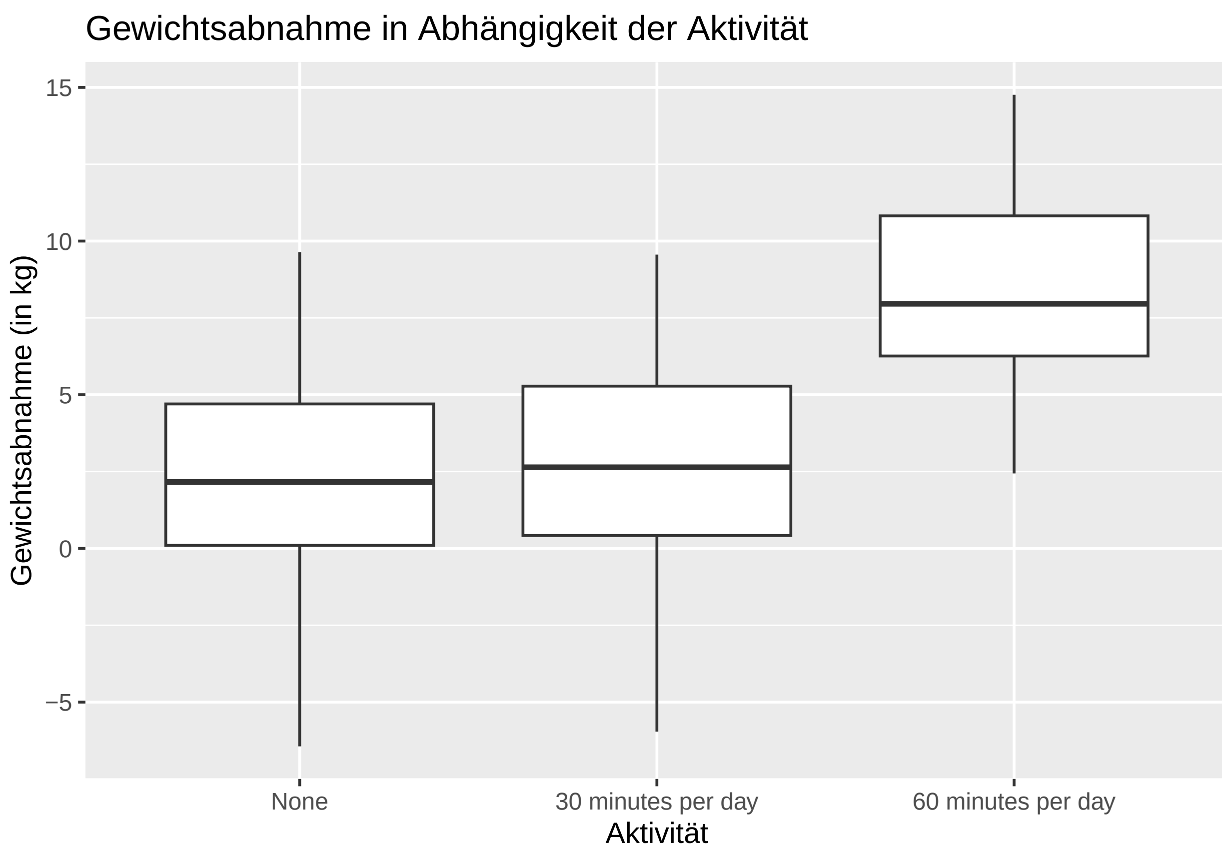
```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## diet            2  363.9   182.0  21.124 6.33e-09 ***
## exercise        2 1508.9   754.4  87.589  < 2e-16 ***
## diet:exercise   4   32.3     8.1   0.937    0.444
## Residuals     171 1472.9     8.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### Ergebnisse der ANOVA

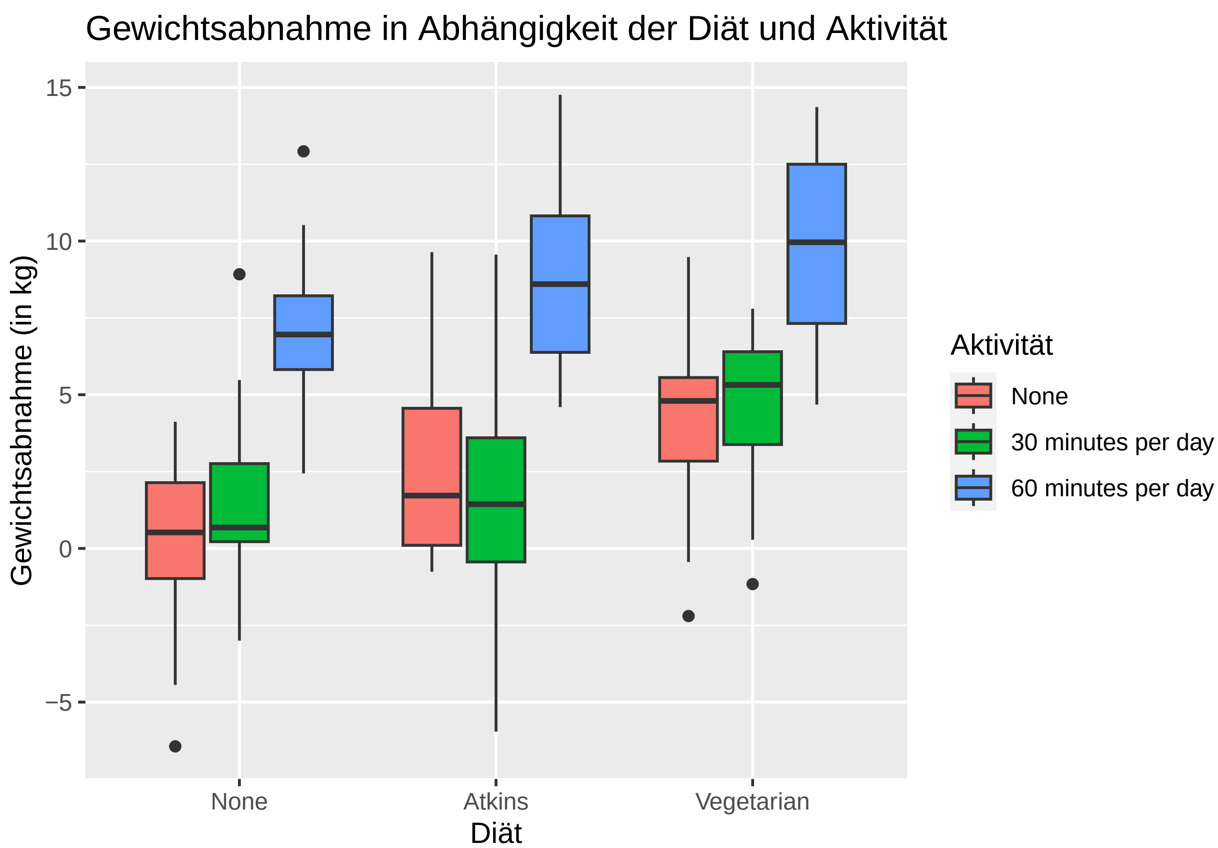
```r
# Boxplot (diets)
ggplot(weightloss, aes(x = diet, y = wloss)) +
  geom_boxplot() +
  labs(title = "Gewichtsabnahme in Abhängigkeit der Diät",
       x = "Diät",
       y = "Gewichtsabnahme (in kg)")
```




```r
# Boxplot (activities)
ggplot(weightloss, aes(x = exercise, y = wloss)) +
  geom_boxplot() +
  labs(title = "Gewichtsabnahme in Abhängigkeit der Aktivität",
       x = "Aktivität",
       y = "Gewichtsabnahme (in kg)")
```




```r
# Boxplot (diets & activities)
ggplot(weightloss, aes(x = diet, y = wloss, fill = exercise)) +
  geom_boxplot() +
  labs(title = "Gewichtsabnahme in Abhängigkeit der Diät und Aktivität",
       x = "Diät",
       y = "Gewichtsabnahme (in kg)",
       fill = "Aktivität")
```

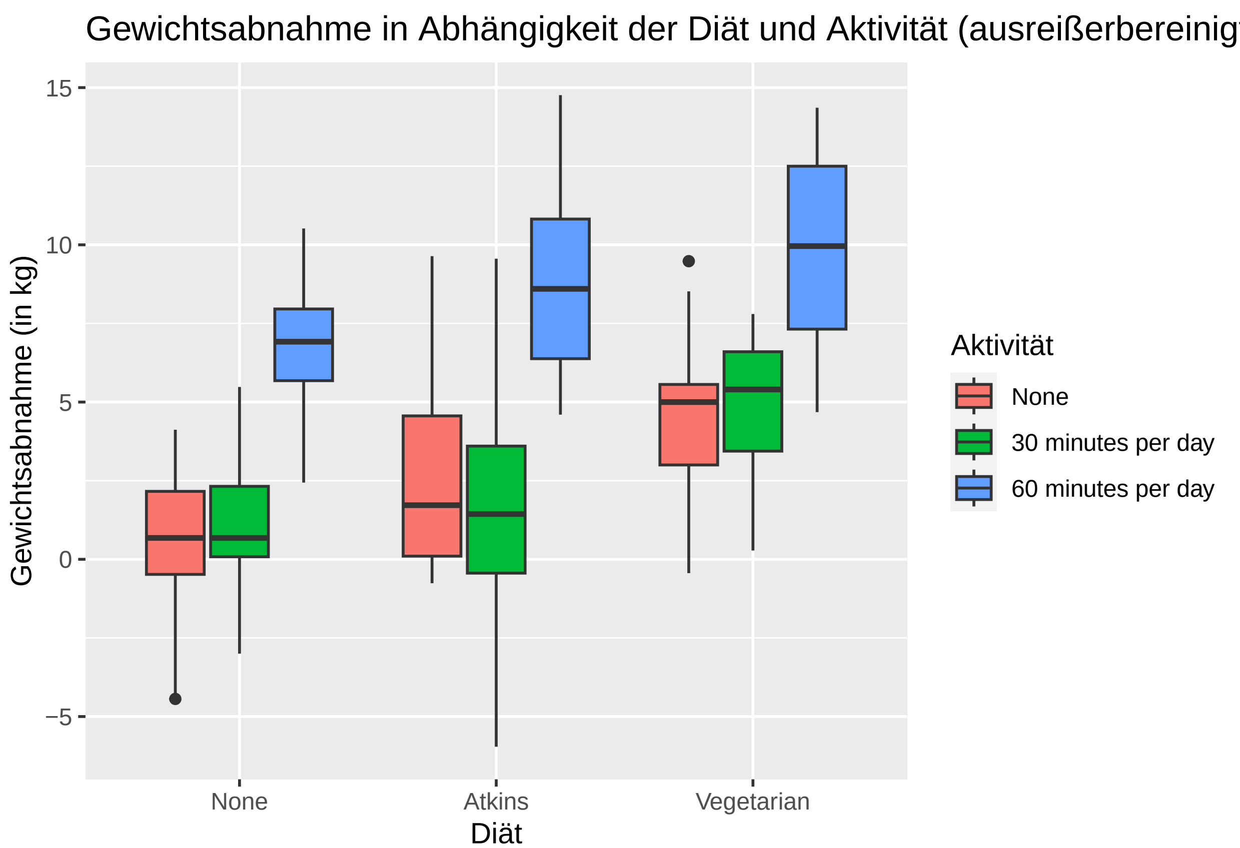



```r
# Ausreißer bereinigen (median +- 1.5 IQR)
library(dplyr)
```

```
##
## Attache Paket: 'dplyr'

## Das folgende Objekt ist maskiert 'package:car':
##
##     recode

## Die folgenden Objekte sind maskiert von 'package:stats':
##
##     filter, lag

## Die folgenden Objekte sind maskiert von 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
weightloss_clean <- weightloss %>%
  group_by(diet, exercise) %>%
  mutate(lb = quantile(wloss, 0.25) - 1.5 * IQR(wloss),
         ub = quantile(wloss, 0.75) + 1.5 * IQR(wloss)) %>%
  filter(wloss >= lb & wloss <= ub)
```

```r
# Boxplot (diets & activities)
ggplot(weightloss_clean, aes(x = diet, y = wloss, fill = exercise)) +
  geom_boxplot() +
  labs(title = "Gewichtsabnahme in Abhängigkeit der Diät und Aktivität (ausreißerbereinigt)",
       x = "Diät",
       y = "Gewichtsabnahme (in kg)",
       fill = "Aktivität")
```



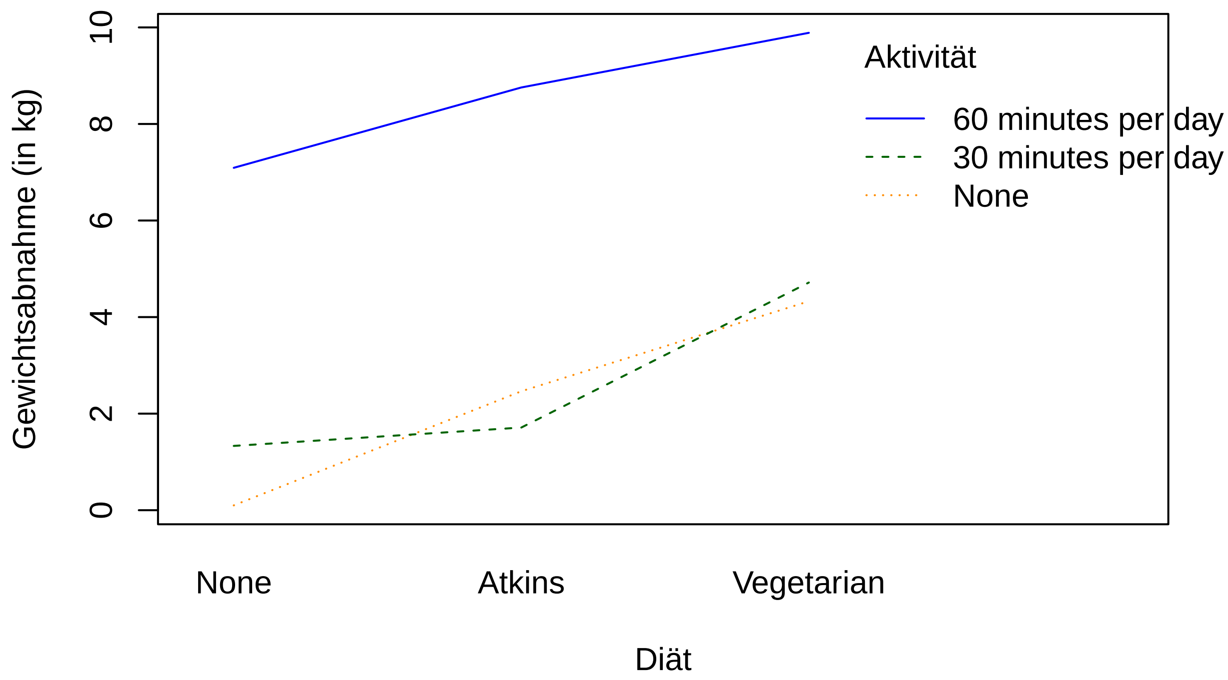

i) Gibt es signifikante Haupteffekte sowie Interaktionseffekte (inkl. Interaktions-Plot)?

Der Faktor Diet (Df=2) hatte einen signifikanten Einfluss auf die Gewichtsabnahme mit einem F-Wert von 21.124 und einem p-Wert von weniger als 0.00001 (6.33e-09).

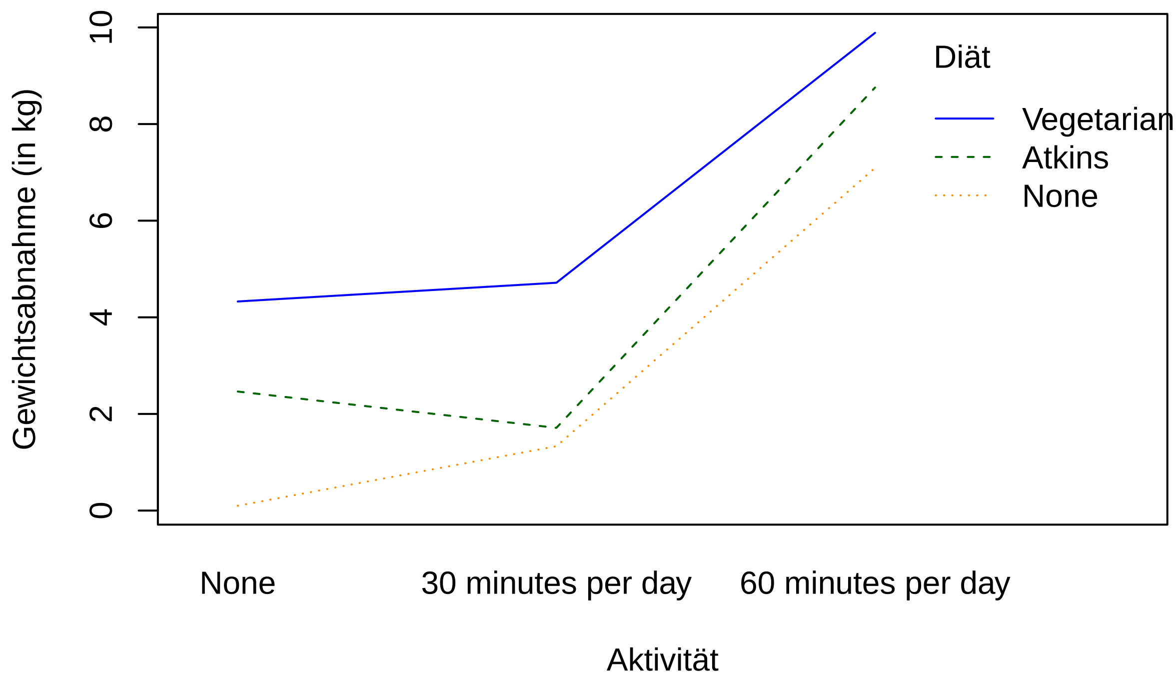
Das Bewegungsausmaß (Df=2) hatte einen noch stärkeren Einfluss auf die Gewichtsabnahme mit einem F-Wert von 87.589 und einem p-Wert von kleiner als 0.00001 (< 2e-16).

Die Interaktion zwischen Diät und Bewegung (Df = 4) war nicht signifikant mit einem F-Wert von 0.937 und einem p-Wert von 0.444. Dies bedeutet, dass die Wechselwirkung zwischen Diät und Bewegung keinen signifikanten Beitrag zur Erklärung der Gewichtsabnahme liefert.

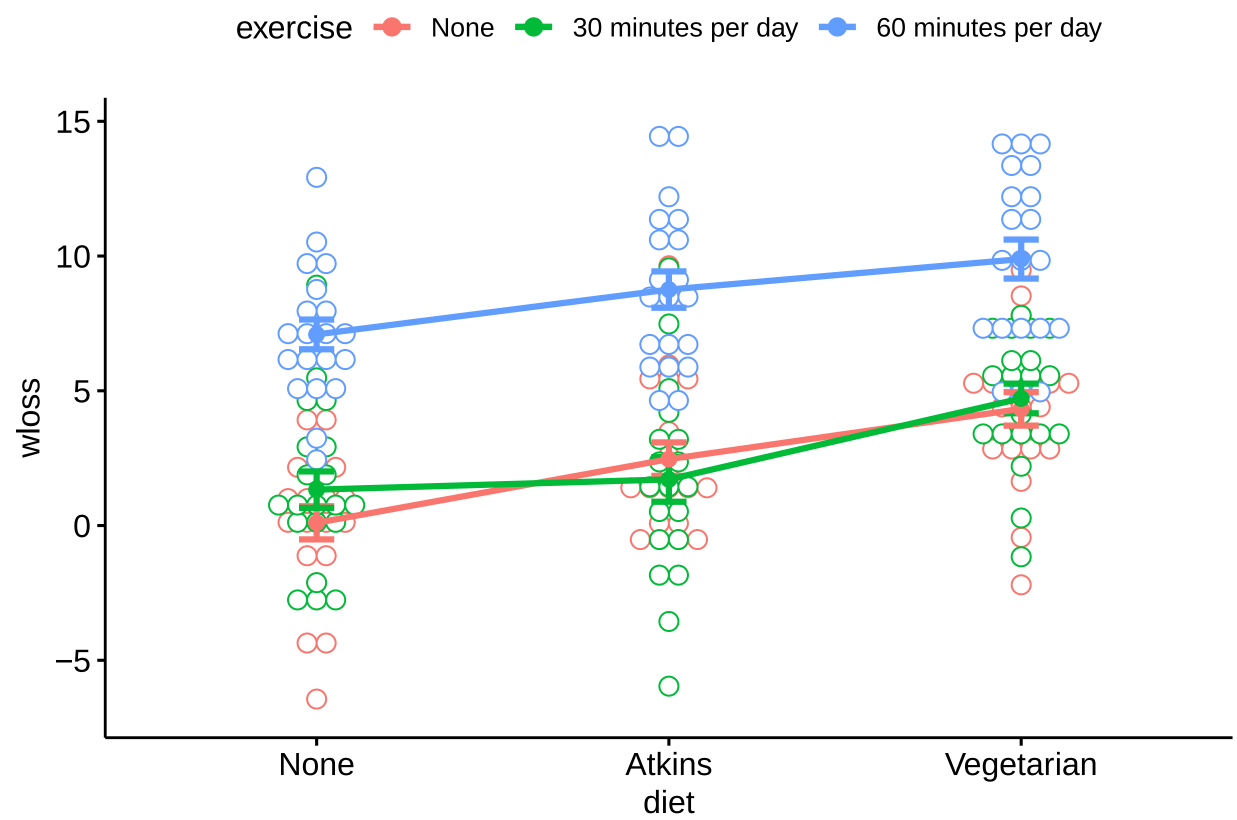
```r
# Interaktionsplots (simple)
ip_act <- interaction.plot(weightloss$diet, weightloss$exercise, weightloss$wloss,
                 col = c("darkorange", "darkgreen", "blue"),
                 pch = c(1, 2, 3), fixed = FALSE, xlab = "Diät",
                 ylab = "Gewichtsabnahme (in kg)", trace.label = "Aktivität")
```



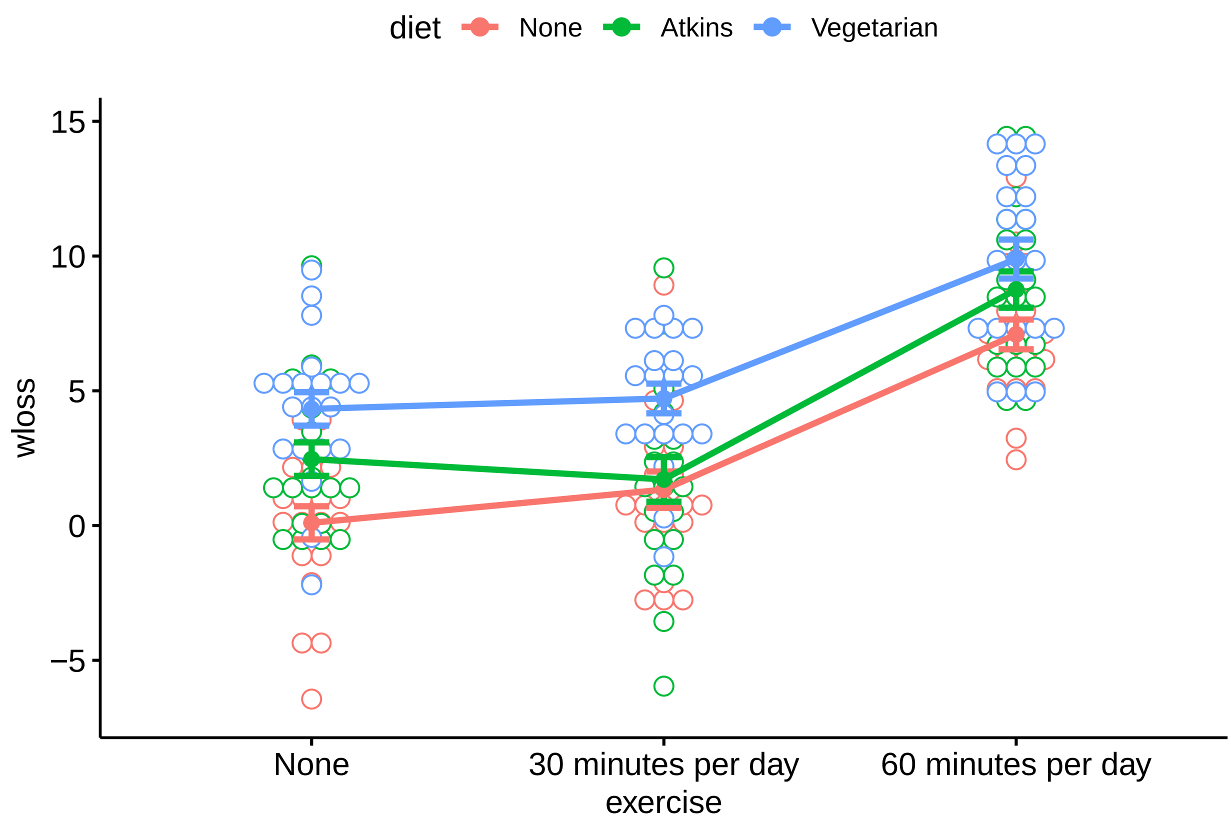

```r
ip_diet <- interaction.plot(weightloss$exercise, weightloss$diet, weightloss$wloss,
                col = c("darkorange", "darkgreen", "blue"),
                pch = c(1, 2, 3), fixed = FALSE, xlab = "Aktivität",
                ylab = "Gewichtsabnahme (in kg)", trace.label = "Diät")
```




```r
# Interaktionsplot (ggpubr)
library(ggpubr)
ggline(weightloss, x = "diet", y = "wloss", color = "exercise", size = 1.1,
       add = c("mean_se","dotplot"))
```




```
ggline(weightloss, x = "exercise", y = "wloss", color = "diet", size = 1.1,
       add = c("mean_se","dotplot"))
```




Interaktionsplot 1: Dieser Plot zeigt die Gewichtsabnahme aufgeteilt nach Diättyp und gruppiert nach Aktivitätsniveau. Die drei Linien repräsentieren die unterschiedliche Aktivitätsniveaus. Auffallend ist hier, dass sich die Linien für keine Aktivität und die des mittleren Atktivitätsniveaus bei der Atkins-Diet kreuzen. Dies könnte auf eine leichte semidisordinale Interaktion, in diesem Fall hindeuten. Allerdings war die Interaktion in der ANOVA statistisch nicht signifikant ($p = 0.444$), somit ist diese visuell wahrgenommene Interaktion möglicherweise nicht stark genug ist, um statistisch statistische Signifikanz zu erreichen.

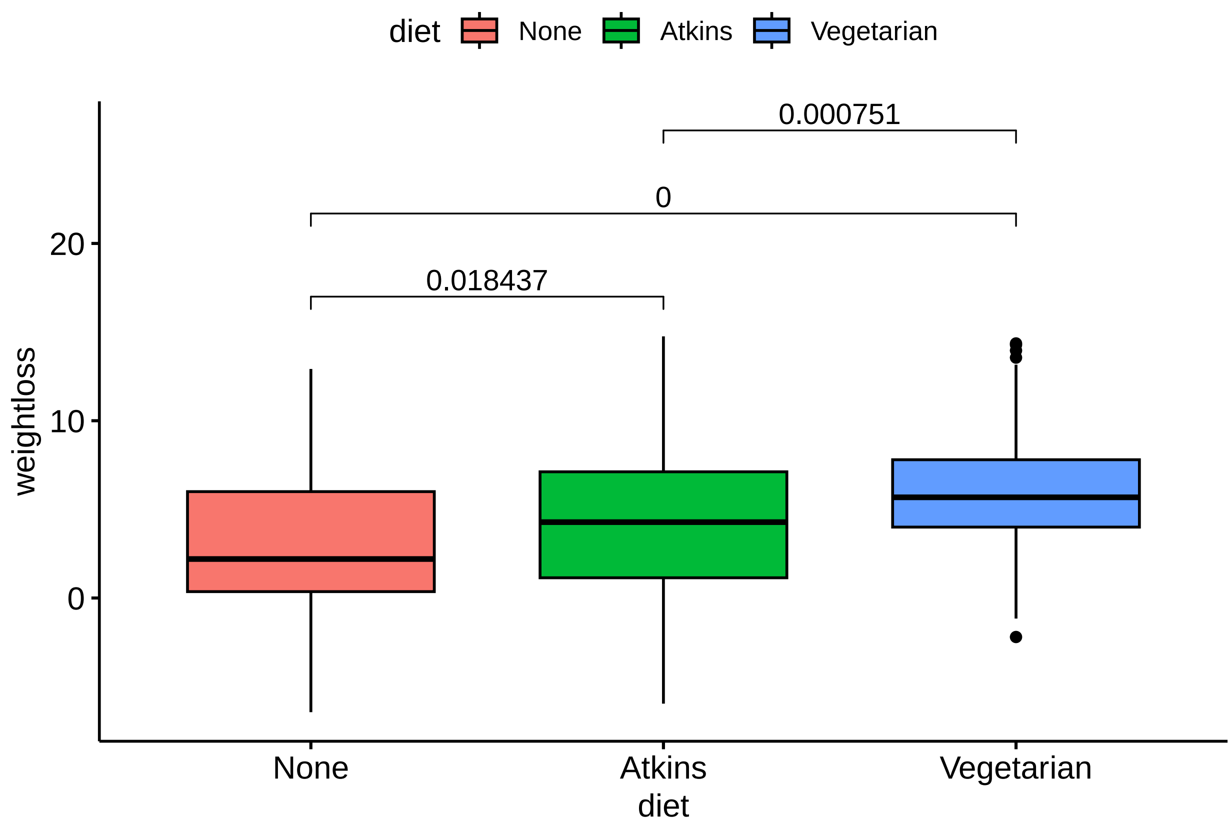
Interaktionsplot 2: Auf diesem Plot ist die Gewichtsabnahme aufgeteilt nach Aktivitätsniveau und gruppiert nach Diättyp. Die drei Linien repräsentieren die unterschiedlichen Diättypen. Die Linien zeigen sehr ähnliche Trends auf, und kreuzen sich nicht. Dies unterstützt das ANOVA-Ergebnis, dass keine signifikante Interaktion vorliegt. Die Linie für "60 minutes per day" zeigt durchwegs die höchsten Werte, was die starke Wirkung eines erhöhten Aktivitätsniveau auf die Gewichtsabnahme unterstreicht, unabhängig von der gewählten Diät.

Um die Unterschiede zwischen den Gruppen genauer zu untersuchen, werden folglich Post-hoc-Tests durchgeführt. Hierfür eignen sich beispielsweise der Tukey's HSD-Test.

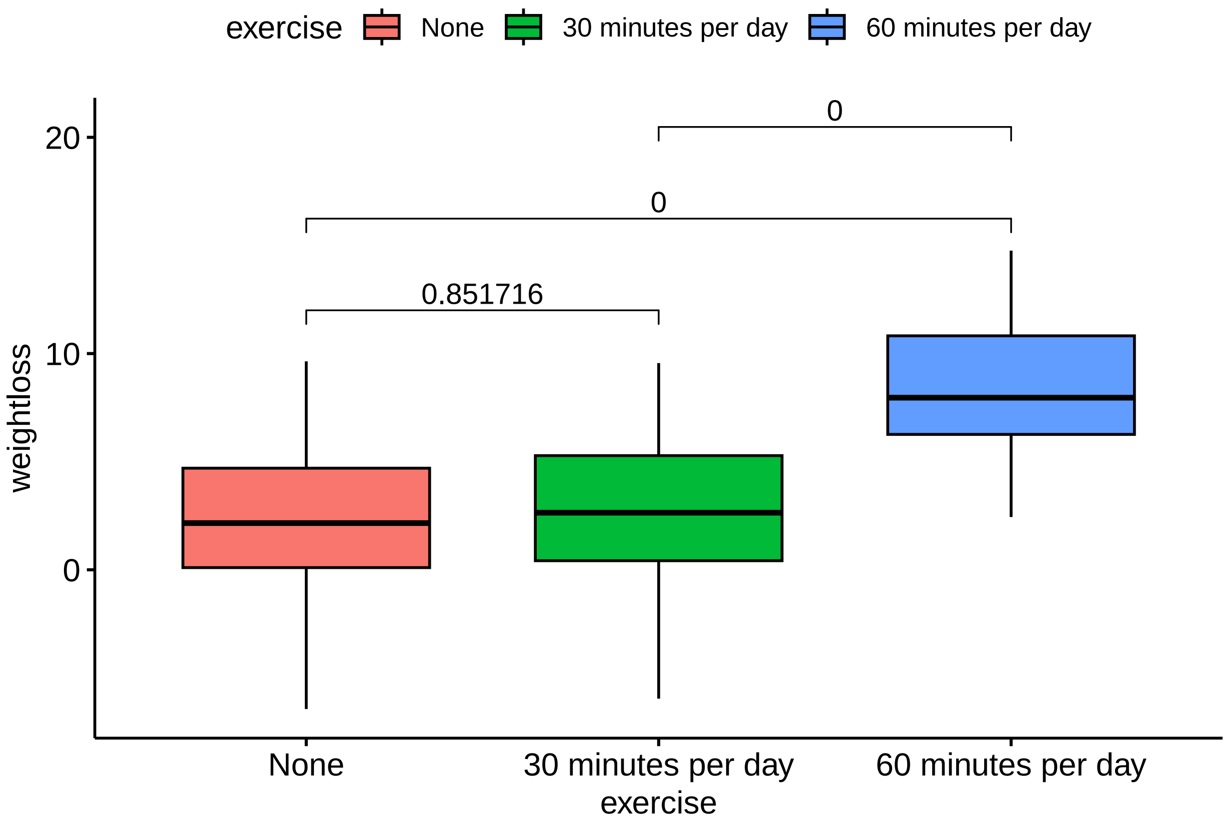
```
# Post-hoc Tests
pht_diet <- TukeyHSD(aov(wloss~diet*exercise,data=weightloss),which="diet")
pht_act <- TukeyHSD(aov(wloss~diet*exercise,data=weightloss),which="exercise")
```

```
# Darsellung der Post-hoc Test Ergebnisse
stat.test_diet <- tibble::tribble(
  ~group1, ~group2,   ~p.adj,
  "Atkins",      "None", round(pht_diet$diet[1,4],6),
  "Vegetarian",      "None", round(pht_diet$diet[2,4],6),
  "Vegetarian",      "Atkins", round(pht_diet$diet[3,4],6))

# Boxplot
```

```r
ggboxplot(weightloss, x = "diet", y = "wloss",
          fill = "diet",
          palette = "Okabe-Ito",
          ylab = "weightloss") +
  stat_pvalue_manual(stat.test_diet,
                     y.position = 17, step.increase = 0.2,
                     label = "p.adj")
```




```r
stat.test_act <- tibble::tribble(
  ~group1, ~group2,   ~p.adj,
  "30 minutes per day",     "None", round(pht_act$exercise[1,4],6),
  "60 minutes per day",     "None", round(pht_act$exercise[2,4],6),
  "60 minutes per day",     "30 minutes per day", round(pht_act$exercise[3,4],6))

# Boxplot
ggboxplot(weightloss, x = "exercise", y = "wloss",
          fill = "exercise",
          palette = "Okabe-Ito",
          ylab = "weightloss") +
  stat_pvalue_manual(stat.test_act,
                     y.position = 12, step.increase = 0.2,
                     label = "p.adj")
```




Interpretation: Eine 2x3 ANOVA wurde durchgeführt, um den Einfluss unterschiedlicher Diäten (Keine, Atkins, Vegetarisch) und Aktivitätsniveaus (Keine, 30 min/Tag, 60 min/Tag) auf die Gewichtsabnahme zu untersuchen. Es wurde ein signifikanter Haupteffekt des Faktors Diät [F(2,171)=21.12,p $<$ 0.01] oder das Aktivitätsniveau [F(2,171)=87.59,p $<$ 0.01] gefunden. Die Interaktion zwischen Diät und Aktivitätsniveau war nicht signifikant (p = 0.444).

Post-hoc-Tests Analysen zeigten signifikanten Unterschiede zwischen einzelnen Diätgruppen und Aktivitätsniveaus.

Personen, die eine vegetarische Diät befolgten, verloren durchschnittlich 3,47 kg mehr als jene ohne spezielle Diät und um 2 kg mehr als Personen, die der Atkins-Diät folgen. Diese Unterschiede sind statistisch signifikant, was die vegetarische Diät als besonders effektiv herausstellt. Die Atkins-Diät zeigt ebenfalls eine überlegene Wirkung im Vergleich zu keiner Diät mit einer signifikant höheren mittleren Gewichtsabnahme von 1,47 kg.

Die tägliche Dauer der körperlichen Aktivität spielte ebenfalls eine entscheidende Rolle bei der Gewichtsreduktion. Die Analyse ergab, dass 60 Minuten tägliche Bewegung zu einer signifikant höheren Gewichtsabnahme von 6,28 kg im Vergleich zu keiner Bewegung führte. Selbst im Vergleich zu 30 Minuten täglicher Aktivität ist das längere Training mit einer zusätzlichen Gewichtsabnahme von fast 6 kg deutlich effektiver. Im Gegensatz dazu zeigt die geringere Dauer von 30 Minuten keine signifikante Wirkung gegenüber keiner Bewegung.

# 5 Mehrfaktorielle ANOVA [2P]

Verwenden Sie den erneut den `Framingham.sav` Datensatz. Analysieren Sie den systolischen Blutdruck `sysbp` abhängig von Geschlecht und Bildungsstufe. Achten Sie auf Ausreißer und fehlende Daten (`NaN, NA's`).

i) Überprüfen Sie die Vorraussetzungen.
ii) Verwenden Sie eine geeignete grafische Darstellung der Daten / Ergebnisse.
iii) Gibt es signifikante Haupteffekte sowie Interaktionseffekte (inkl. Interaktions-Plot)?
iv) Achten Sie auf eine "statistisch korrekte" Formulierung des Ergebnisses.