WDDA FS 2025: Leitfaden für Aufgabenserie 3

2025-03-26

Einleitung

Dieser Leitfaden bietet detaillierte Erklärungen für die Übungen in WDDA FS 2025 Aufgabenserie 3. Für jede Aufgabe werden wir den Denkprozess (und die notwendigen Schritte in R) durchgehen und erklären, wie man zur richtigen Lösung gelangt.

Daten laden:
Laden Sie den Datensatz aus der Excel-Datei in R mit der `readxl`-Bibliothek.

```
library(readxl)
data <- read_excel("../data/WDDA_03.xlsx", sheet = "BFH")
```

Aufgabe 1: Mittelwerte und Vergleich der Geschlechtsgruppen

Aufgabenstellung:
1. Ermitteln Sie die mittlere Körpergrösse der Männer und der Frauen.
(a) Wie viele Frauen sind grösser als die durchschnittliche Körpergrösse der Männer?
(b) Ist die Anzahl der Männer, die kleiner sind als die mittlere Frauengrösse, ähnlich?

Schritt-für-Schritt-Erklärung:

1. **Daten filtern und Mittelwerte berechnen:**
   Erstellen Sie Teildatensätze für Männer und Frauen und berechnen Sie jeweils den Mittelwert der Körpergrösse.

```
# Beispiel: Annahme, dass 'data' Ihren Datensatz enthält
male_height <- data$height[data$gender == "Male"]
female_height <- data$height[data$gender == "Female"]

mean_male <- mean(male_height, na.rm = TRUE)
mean_female <- mean(female_height, na.rm = TRUE)

cat("Mittlere Körpergrösse Männer:", mean_male, "cm\n")
```

```
## Mittlere Körpergrösse Männer: 181.1308 cm
```

```
cat("Mittlere Körpergrösse Frauen:", mean_female, "cm\n")
```

```
## Mittlere Körpergrösse Frauen: 167.186 cm
```

2. **Vergleiche durchführen:**

(a) Zählen Sie, wie viele Frauen grösser sind als der Mittelwert der Männer.

(b) Zählen Sie, wie viele Männer kleiner sind als der Mittelwert der Frauen.

```
count_female_above_male_mean <- sum(female_height > mean_male, na.rm = TRUE)
count_male_below_female_mean <- sum(male_height < mean_female, na.rm = TRUE)

cat("Anzahl Frauen grösser als mittlere Männergrösse:", count_female_above_male_mean, "\n")
```

```
## Anzahl Frauen grösser als mittlere Männergrösse: 1
```

```
cat("Anzahl Männer kleiner als mittlere Frauengrösse:", count_male_below_female_mean, "\n")
```

```
## Anzahl Männer kleiner als mittlere Frauengrösse: 3
```

3. **Ergebnisvergleich:**
   Die Lösungen lauten:
   - Mittlere Männergrösse: 181.1 cm
   - Mittlere Frauengrösse: 167.2 cm
   - (a) 1 Frau grösser als der Mittelwert der Männer
   - (b) 3 Männer kleiner als der Mittelwert der Frauen

Aufgabe 2: Schwarmintelligenz – Analyse der Variable `jar`

Aufgabenstellung:
2. Die Schwarmintelligenz besagt, dass das kombinierte Wissen einer Menge besser ist als das Wissen einer einzelnen Person. Anhand der Daten von `jar` (es waren tatsächlich 405 M&Ms im Glas) sollen folgende Aufgaben gelöst werden:
(a) Bestimmen Sie den Mittelwert von `jar`.
(b) Finden Sie das Geburtsdatum der Person, die am besten geraten hat.
(c) Wie viele Individuen sind näher an der tatsächlichen Anzahl M&Ms als am Mittelwert des Datensatzes?

Schritt-für-Schritt-Erklärung:

1. **Mittelwert berechnen:**
   Nutzen Sie die Funktion `mean()`, um den Mittelwert der Schätzwerte zu bestimmen.

```
mean_jar <- mean(data$jar, na.rm = TRUE)
cat("Mittelwert von jar:", mean_jar, "\n")
```

```
## Mittelwert von jar: 299.6636
```

2. **Bestes Schätzer-Identifizieren:**
   Finden Sie den Datensatz-Eintrag, dessen Schätzwert am nächsten an 405 liegt.

```r
diff_to_actual <- abs(data$jar - 405)
best_index <- which.min(diff_to_actual)
best_birthdate <- data$dob[best_index]  # Annahme: Geburtsdatum in 'dob'
cat("Geburtsdatum des besten Schätzers:", best_birthdate, "\n")
```

```
## Geburtsdatum des besten Schätzers: 1997-04-10
```

3. **Vergleich der Abstände:**
   Ermitteln Sie, wie viele Individuen näher an 405 liegen als am Mittelwert (mean_jar).

```r
closer_than_mean <- sum(abs(data$jar - 405) < abs(data$jar - mean_jar), na.rm = TRUE)
cat("Anzahl Individuen, die näher an 405 liegen als am Mittelwert:", closer_than_mean, "\n")
```

```
## Anzahl Individuen, die näher an 405 liegen als am Mittelwert: 61
```

4. **Lösungshinweise:**
   Die erwarteten Lösungen sind:
   - (a) 299.7
   - (b) 1997-04-10
   - (c) 61

Aufgabe 3: Vergleich von Mittelwert und Median

Aufgabenstellung:
3. Vergleichen Sie für die folgenden Variablen den Mittelwert mit dem Median. Falls sie nicht ungefähr gleich sind, geben Sie eine Erklärung, warum dies nicht der Fall ist.
(a) `jar`
(b) `siblings`
(c) `distance`

Schritt-für-Schritt-Erklärung:

1. **Berechnungen durchführen:**
   Verwenden Sie `mean()` und `median()` für jede Variable.

```r
# Für jar:
mean_jar <- mean(data$jar, na.rm = TRUE)
median_jar <- median(data$jar, na.rm = TRUE)
cat("jar - Mittelwert:", mean_jar, "Median:", median_jar, "\n")
```

```
## jar - Mittelwert: 299.6636 Median: 287
```

```r
# Für siblings:
mean_siblings <- mean(data$siblings, na.rm = TRUE)
median_siblings <- median(data$siblings, na.rm = TRUE)
cat("siblings - Mittelwert:", mean_siblings, "Median:", median_siblings, "\n")
```

```
## siblings - Mittelwert: 1.543779 Median: 1
```

```r
# Für distance:
mean_distance <- mean(data$distance, na.rm = TRUE)
median_distance <- median(data$distance, na.rm = TRUE)
cat("distance - Mittelwert:", mean_distance, "Median:", median_distance, "\n")
```

```
## distance - Mittelwert: 25.19631 Median: 24
```

2. **Interpretation:**
   - Bei `jar` liegen die Werte z.B. bei 299.7 (Mittelwert) und 287 (Median).
   - Bei `siblings` ist der Unterschied (1.5 vs. 1) klein, bei `distance` (25.2 km vs. 24 km) könnte eine Rechtsschiefe vorliegen, weil einige Personen sehr weit entfernt wohnen.

Aufgabe 4: Schätzung des zentralen Trends von `cash`

Aufgabenstellung:
4. Schätzen Sie vor der Berechnung, welcher Wert grösser ist: der Mittelwert oder der Median von `cash`?

Schritt-für-Schritt-Erklärung:

1. **Überlegungen anstellen:**
   Bei rechtsschiefen Daten – etwa wenn einzelne Personen extrem hohe Bargeldbeträge besitzen – wird der Mittelwert tendenziell durch Ausreisser nach oben verzerrt, während der Median robuster ist.
2. **Erwartete Lösung:**
   Aufgrund einiger sehr hoher Werte im Datensatz wird der Mittelwert höher sein als der Median.

Aufgabe 5: Durchschnittsausgabe für Geschwister-Geschenk

Aufgabenstellung:
5. Wie viel geben die Studierenden im Datensatz durchschnittlich für ein Geschwister-Geschenk aus?

Schritt-für-Schritt-Erklärung:

1. **Mittelwert berechnen:**
   Verwenden Sie `mean()` auf der entsprechenden Variable (Spalte `present` im `BFH` Datenblatt).

```r
mean_gift <- weighted.mean(data$present, data$siblings)
cat("Durchschnittliche Ausgabe für ein Geschwister-Geschenk:", mean_gift, "CHF\n")
```

```
## Durchschnittliche Ausgabe für ein Geschwister-Geschenk: 85.10448 CHF
```

2. **Lösungshinweis:**
   Die Lösung lautet 85.1 CHF.

Aufgabe 6: Bestimmung der Spannweite (Range)

Aufgabenstellung:
6. Finden Sie den Bereich (Spannweite) der folgenden Daten:
(a) `height`
(b) `maths`
(c) `cash`

Schritt-für-Schritt-Erklärung:

1. **Berechnung:**
   Die Spannweite berechnet sich als Differenz zwischen dem Maximum und Minimum.

```
range_height <- diff(range(data$height, na.rm = TRUE))
range_maths <- diff(range(data$maths, na.rm = TRUE))
range_cash <- diff(range(data$cash, na.rm = TRUE))

cat("Spannweite height:", range_height, "cm\n")
```

```
## Spannweite height: 43 cm
```

```
cat("Spannweite maths:", range_maths, "\n")
```

```
## Spannweite maths: 5
```

```
cat("Spannweite cash:", range_cash, "CHF\n")
```

```
## Spannweite cash: 749.85 CHF
```

2. **Ergebnis:**
   Erwartete Werte:
   - (a) 43 cm
   - (b) 5
   - (c) 749.85 CHF

Aufgabe 7: Interquartilsabstand (IQR)

Aufgabenstellung:
7. Ermitteln Sie den Interquartilsabstand der folgenden Listen:
(a) `height`
(b) `maths`
(c) `cash`

Schritt-für-Schritt-Erklärung:

1. **IQR berechnen:**
   Nutzen Sie die Funktion `IQR()`.

```r
iqr_height <- IQR(data$height, na.rm = TRUE)
iqr_maths <- IQR(data$maths, na.rm = TRUE)
iqr_cash <- IQR(data$cash, na.rm = TRUE)

cat("IQR height:", iqr_height, "cm\n")
```

```
## IQR height: 14 cm
```

```r
cat("IQR maths:", iqr_maths, "\n")
```

```
## IQR maths: 1.5
```

```r
cat("IQR cash:", iqr_cash, "CHF\n")
```

```
## IQR cash: 60 CHF
```

2. **Ergebnis:**
   Erwartete Werte:
   - (a) 14 cm
   - (b) 1.5
   - (c) 60 CHF

Aufgabe 8: Nebeneinander liegende Boxplots

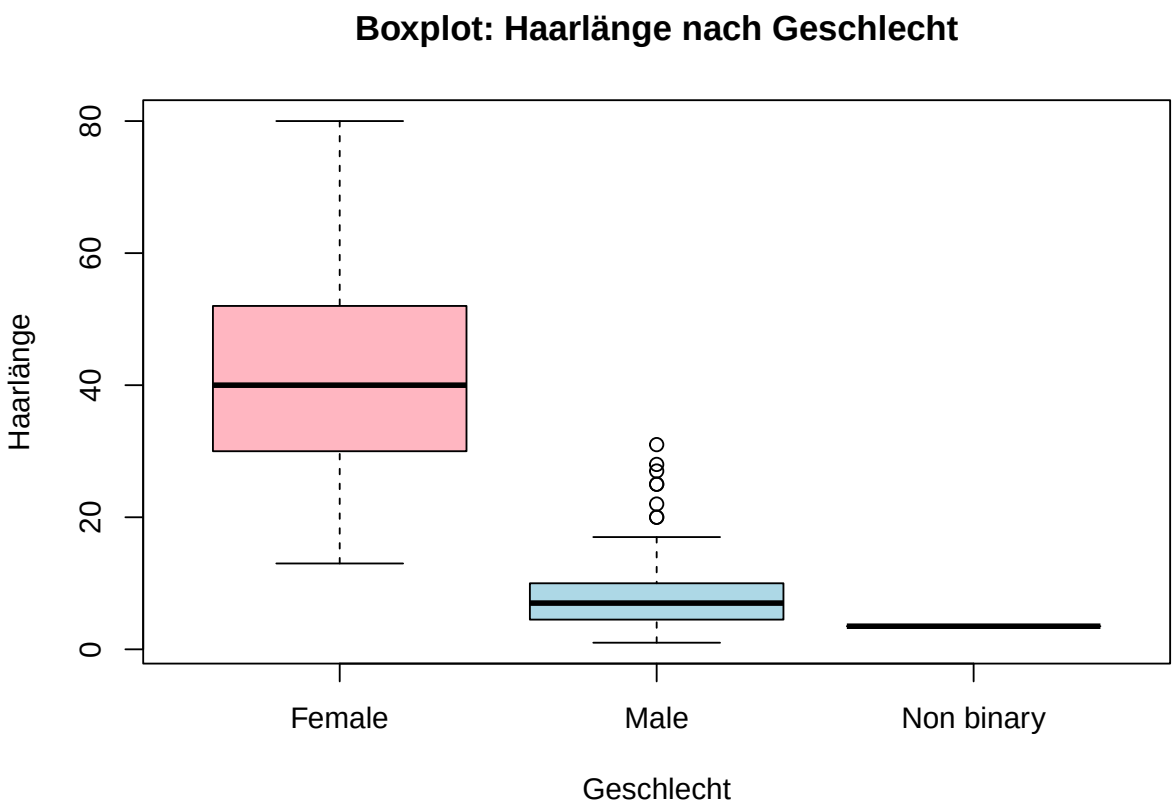
Aufgabenstellung:
8. Erstellen Sie nebeneinander liegende Boxplots, um die folgenden Datenpaare zu vergleichen:
(a) `hair` und `gender`
(b) `distance` und `transport`
(c) `foot` und `eye`

Schritt-für-Schritt-Erklärung:

1. **Boxplots erstellen:**
   Nutzen Sie die Funktion `boxplot()` mit einer Gruppierungsformel.

   Beispiel (a) – Vergleich der Haarlänge zwischen den Geschlechtern:

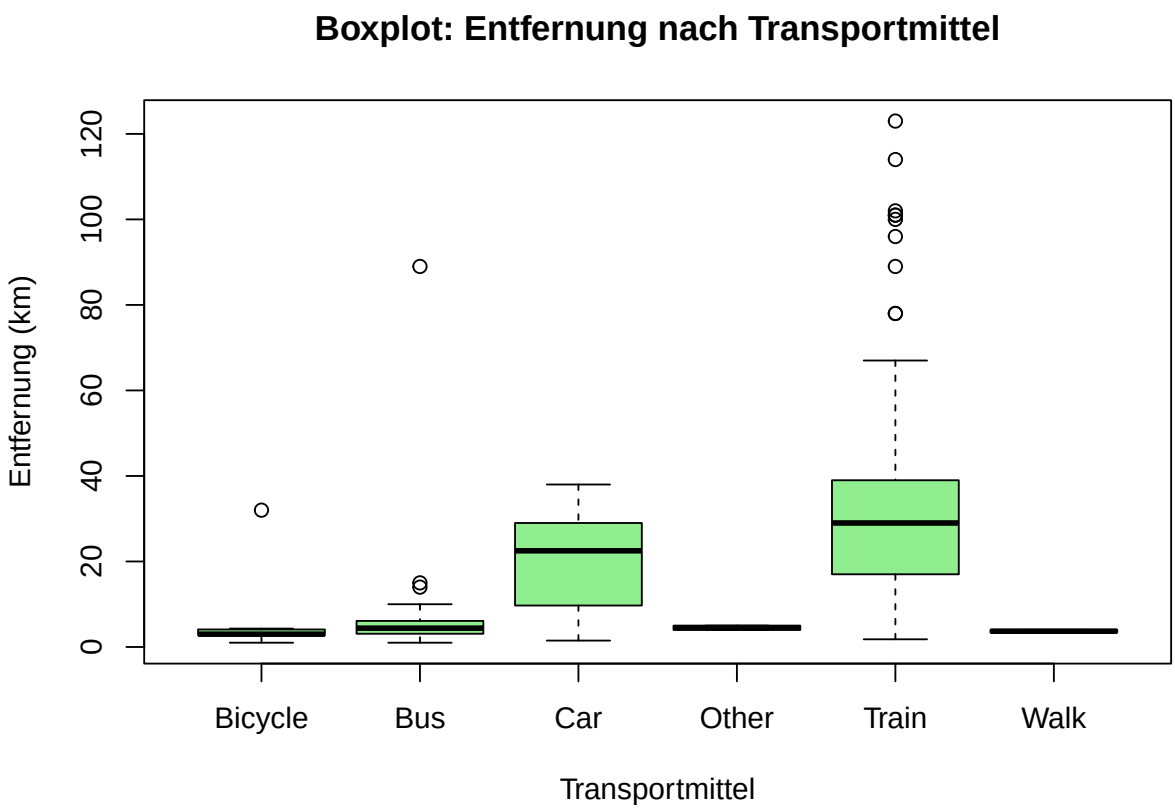
```r
boxplot(data$hair ~ data$gender,
        main = "Boxplot: Haarlänge nach Geschlecht",
        xlab = "Geschlecht", ylab = "Haarlänge",
        col = c( "lightpink","lightblue"))
```



2. **Weitere Vergleiche:**
   Für (b) und (c) gehen Sie analog vor – falls die Gruppierungsvariable numerisch ist, müssen Sie diese ggf. in einen Faktor umwandeln.

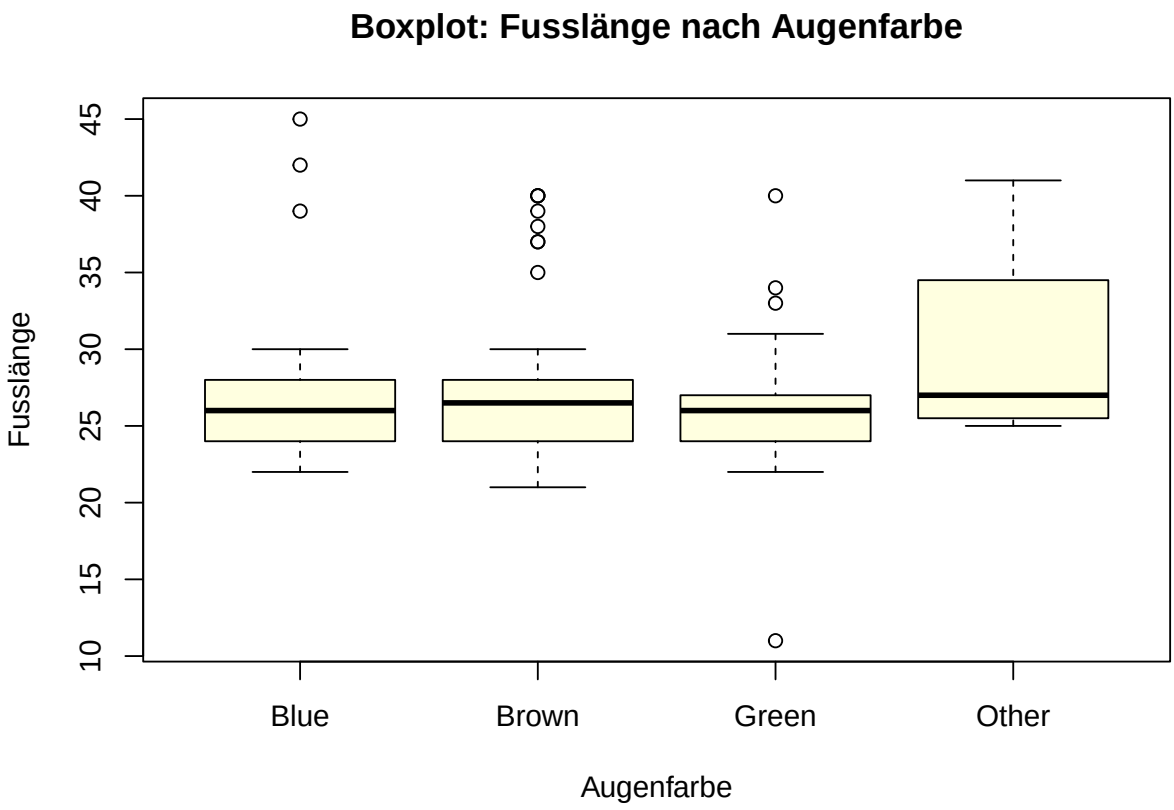
```
# Beispiel (b): distance nach transport
# Sicherere Implementierung mit Fehlerbehandlung
transport_data <- data.frame(
  distance = data$distance,
  transport = data$transport
)
transport_data <- transport_data[complete.cases(transport_data), ]

if(nrow(transport_data) > 0 && length(unique(transport_data$transport)) > 0) {
  boxplot(distance ~ transport, data = transport_data,
          main = "Boxplot: Entfernung nach Transportmittel",
          xlab = "Transportmittel", ylab = "Entfernung (km)",
          col = "lightgreen")
} else {
  cat("Nicht genügend gültige Daten für den Transport-Boxplot\n")
}
```




```r
# Beispiel (c): foot nach eye
# Sicherere Implementierung mit Fehlerbehandlung
eye_foot_data <- data.frame(
  foot = data$foot,
  eye = data$eye
)
eye_foot_data <- eye_foot_data[complete.cases(eye_foot_data), ]

if(nrow(eye_foot_data) > 0 && length(unique(eye_foot_data$eye)) > 0) {
  boxplot(foot ~ eye, data = eye_foot_data,
          main = "Boxplot: Fusslänge nach Augenfarbe",
          xlab = "Augenfarbe", ylab = "Fusslänge",
          col = "lightyellow")
} else {
  cat("Nicht genügend gültige Daten für den Augenfarbe-Boxplot\n")
}
```



3. **Interpretation:**
   Diskutieren Sie anhand der Boxplots, ob und wie sich die Gruppen in ihren Verteilungen unterscheiden.

Aufgabe 9: Standardabweichung (als Stichprobe)

Aufgabenstellung:
9. Ermitteln Sie die Standardabweichung (als Stichprobe) der folgenden Daten:
(a) `height`
(b) `maths`
(c) `cash`

Schritt-für-Schritt-Erklärung:

1. **Berechnung mit `sd()`:**

```
sd_height <- sd(data$height, na.rm = TRUE)
sd_maths <- sd(data$maths, na.rm = TRUE)
sd_cash <- sd(data$cash, na.rm = TRUE)

cat("Standardabweichung height:", sd_height, "cm\n")
```

```
## Standardabweichung height: 9.292882 cm
```

```r
cat("Standardabweichung maths:", sd_maths, "\n")
```

```
## Standardabweichung maths: 0.878979
```

```r
cat("Standardabweichung cash:", sd_cash, "CHF\n")
```

```
## Standardabweichung cash: 83.79841 CHF
```

2. **Ergebnis:**
   Erwartete Lösungen:
   - (a) 9.3 cm
   - (b) 0.9
   - (c) 83.8 CHF

Aufgabe 10: Überprüfung der empirischen Regel

Aufgabenstellung:
10. Überprüfen Sie, ob die empirische Regel für die folgenden Datensätze gilt:
(a) `height`
(b) `foot`
(c) `distance`
(d) `reaction1`

Schritt-für-Schritt-Erklärung:

1. **Empirische Regel:**
   Die empirische Regel (68-95-99.7-Regel) besagt, dass etwa 68% der Werte innerhalb einer Standardabweichung, 95% innerhalb von zwei und 99.7% innerhalb von drei Standardabweichungen vom Mittelwert liegen.
2. **Berechnungen:**
   Für jede Variable:
   - Berechnen Sie Mittelwert und Standardabweichung.
   - Ermitteln Sie den Anteil der Daten, der in den Intervallen [Mittelwert ± 1*SD], [± 2*SD] und [± 3*SD] liegt.

   Beispiel für `height`:

```r
mean_h <- mean(data$height, na.rm = TRUE)
sd_h <- sd(data$height, na.rm = TRUE)
lower1 <- mean_h - sd_h; upper1 <- mean_h + sd_h
proportion1 <- mean(data$height >= lower1 & data$height <= upper1, na.rm = TRUE)
cat("Anteil innerhalb 1 SD für height:", proportion1, "\n")
```

```
## Anteil innerhalb 1 SD für height: 0.6267281
```

3. **Erwartete Ergebnisse:**
   Die Lösungen werden als Mengen von Anteilen angegeben, z.B. für `height`: {0.63, 0.97, 1} usw.

Aufgabe 11: Analyse der IHG Kundenbewertungen

Aufgabenstellung:
11. Die Intercontinental Hotels Group (IHG) zeigt Kundenbewertungen (1–5) für Holiday Inn Hotels in 25 grossen europaeischen Städten.
(a) Berechnen Sie den Mittelwert und den Median.
(b) Erläutern Sie, ob es besser ist, den Mittelwert oder den Median als Mass für die zentrale Tendenz zu verwenden.
(c) Berechnen Sie das erste und dritte Quartil sowie den IQR.
(d) Berechnen Sie das 85. Perzentil.

Schritt-für-Schritt-Erklärung:

Vorbereitung: Daten laden

```
data <- read_excel("../data/WDDA_03.xlsx", sheet = "IHG")
```

1. **Berechnungen:**

```
mean_ratings <- mean(data$`Customer ratings`, na.rm = TRUE)
median_ratings <- median(data$`Customer ratings`, na.rm = TRUE)
quantiles <- quantile(data$`Customer ratings`, probs = c(0.25, 0.75, 0.85), na.rm = TRUE)
iqr_ratings <- IQR(data$`Customer ratings`, na.rm = TRUE)

cat("Mittelwert IHG:", mean_ratings, "\n")
```

```
## Mittelwert IHG: 4.1612
```

```
cat("Median IHG:", median_ratings, "\n")
```

```
## Median IHG: 4.2
```

```
cat("Q1:", quantiles[1], " Q3:", quantiles[2], " IQR:", iqr_ratings, "\n")
```

```
## Q1: 4  Q3: 4.3  IQR: 0.3
```

```
cat("85. Perzentil:", quantiles[3], "\n")
```

```
## 85. Perzentil: 4.4
```

2. **Interpretation:**
   - (a) Erwartete Werte: Mittelwert ≈ 4.16, Median ≈ 4.2.
   - (b) Da keine offensichtlichen Ausreisser vorhanden sind, ist der Mittelwert ein gutes Mass.
   - (c) Q1 = 3.95, Q3 = 4.3, IQR = 0.35.
   - (d) 85. Perzentil: 4.4

Aufgabe 12: Polizeiaufzeichnungen – Winter vs. Sommer

Aufgabenstellung:
12. Polizeiaufzeichnungen zeigen die täglich gemeldeten Verbrechen in zwei Zeitperioden.
Winter: 18, 20, 15, 16, 21, 20, 12, 16, 19, 20
Sommer: 28, 18, 24, 32, 18, 29, 23, 38, 28, 18
(a) Berechnen Sie die Spannweite und den IQR für jede Periode.
(b) Berechnen Sie die Varianz und Standardabweichung.
(c) Berechnen Sie den Variationskoeffizienten.
(d) Vergleichen Sie die Variabilität der beiden Zeiträume.

Schritt-für-Schritt-Erklärung:

1. **Daten definieren:**

```
winter <- c(18, 20, 15, 16, 21, 20, 12, 16, 19, 20)
sommer <- c(28, 18, 24, 32, 18, 29, 23, 38, 28, 18)
```

2. **Spannweite und IQR berechnen:**

```
range_winter <- diff(range(winter))
iqr_winter <- IQR(winter)
range_sommer <- diff(range(sommer))
iqr_sommer <- IQR(sommer)

cat("Winter - Spannweite:", range_winter, ", IQR:", iqr_winter, "\n")
```

```
## Winter - Spannweite: 9 , IQR: 4
```

```
cat("Sommer - Spannweite:", range_sommer, ", IQR:", iqr_sommer, "\n")
```

```
## Sommer - Spannweite: 20 , IQR: 9.5
```

3. **Varianz, Standardabweichung und Variationskoeffizient:**

```
var_winter <- var(winter)
sd_winter <- sd(winter)
vc_winter <- sd_winter / mean(winter) * 100  # in Prozent

var_sommer <- var(sommer)
sd_sommer <- sd(sommer)
vc_sommer <- sd_sommer / mean(sommer) * 100

cat("Winter - Varianz:", var_winter, ", SD:", sd_winter, ", VK:", vc_winter, "%\n")
```

```
## Winter - Varianz: 8.233333 , SD: 2.869379 , VK: 16.21118 %
```

```
cat("Sommer - Varianz:", var_sommer, ", SD:", sd_sommer, ", VK:", vc_sommer, "%\n")
```

```
## Sommer - Varianz: 44.48889 , SD: 6.669999 , VK: 26.05468 %
```

4. **Erwartete Ergebnisse:**
   - Winter: Spannweite = 2, IQR = 4, Varianz ≈ 8.23, SD ≈ 2.87, VK ≈ 16.21%
   - Sommer: Spannweite = 20, IQR = 9 (Lösungsblatt hatte hier einen Tippfehler), Varianz ≈ 44.49, SD ≈ 6.67, VK ≈ 26.05%
   - (d) Die Sommerperiode weist eine grössere Variabilität auf.

Aufgabe 13: Standardisierung – z-Werte

Aufgabenstellung:
13. Bestimmen Sie den Mittelwert und die Standardabweichung (als Stichprobe) der z-Werte für die folgenden Daten:
(a) `height`
(b) `foot`
(c) `house`

Schritt-für-Schritt-Erklärung:

1. **Z-Transformation:**
   Standardisieren Sie die Daten mit der Formel $z = \frac{x - \text{Mittelwert}}{\text{SD}}$.

```
# Daten laden
 data <- read_excel("../data/WDDA_03.xlsx", sheet = "BFH")


z_height <- scale(data$height)
z_foot   <- scale(data$foot)
z_house  <- scale(data$house)

cat("Mittelwert z-height:", mean(z_height, na.rm = TRUE), "\n")
```

```
## Mittelwert z-height: -1.163098e-15
```

```
cat("SD z-height:", sd(z_height, na.rm = TRUE), "\n")
```

```
## SD z-height: 1
```

```
# Analog für foot und house
```

2. **Ergebnis:**
   Alle standardisierten Listen haben einen Mittelwert von 0 und eine Standardabweichung von 1.

Aufgabe 14: Schiefe berechnen

Aufgabenstellung:
14. Berechnen Sie die Schiefe für die folgenden Daten:
(a) `height`

(b) `distance`
(c) `cash`
(d) `hair` – getrennt nach Männern und Frauen

Schritt-für-Schritt-Erklärung:

1. **Berechnung der Schiefe:**
   Nutzen Sie z.B. die Funktion `skewness()` aus dem Package `e1071` oder `moments`.

```
library(e1071)

skew_height <- skewness(data$height, na.rm = TRUE)
skew_distance <- skewness(data$distance, na.rm = TRUE)
skew_cash <- skewness(data$cash, na.rm = TRUE)
skew_hair_m <- skewness(data$hair[data$gender == "Male"], na.rm = TRUE)
skew_hair_f <- skewness(data$hair[data$gender == "Female"], na.rm = TRUE)

cat("Schiefe height:", skew_height, "\n")
```

```
## Schiefe height: -0.1537161
```

```
cat("Schiefe distance:", skew_distance, "\n")
```

```
## Schiefe distance: 1.714979
```

```
cat("Schiefe cash:", skew_cash, "\n")
```

```
## Schiefe cash: 5.416389
```

```
cat("Schiefe hair (Männer):", skew_hair_m, "\n")
```

```
## Schiefe hair (Männer): 1.702794
```

```
cat("Schiefe hair (Frauen):", skew_hair_f, "\n")
```

```
## Schiefe hair (Frauen): 0.302638
```

2. **Erwartete Werte:**
   - (a) $\approx$ -0.2
   - (b) $\approx$ 1.7
   - (c) $\approx$ 5.5
   - (d) Männer: ca. 1.74, Frauen: ca. 0.31

Aufgabe 15: Analyse der Ebola-Symptomentwicklung

Aufgabenstellung:
15. Eine Grafik (The Economist) zeigt die Verteilung der Anzahl Tage von der Infektion bis zum Auftreten von Ebola-Symptomen.
(a) Schätzen Sie die mittlere Anzahl von Tagen.
(b) Wie gross ist der Modus (häufigster Wert)?
(c) Diskutieren Sie die Schiefe des Diagramms.
(d) Wie kann das Diagramm verwendet werden, um die Gesamtzahl der Fälle in der Studie anzugeben?

Schritt-für-Schritt-Erklärung:

1. **Schätzung der zentralen Tendenzen:**
   Anhand der Grafik können Sie visuell abschätzen:
   - (a) Der Mittelwert wird etwa bei 11 Tagen liegen.
   - (b) Der Modus entspricht dem höchsten Punkt der Kurve und liegt bei ca. 7 Tagen.
2. **Diskussion der Schiefe:**
   Da die x-Achse gegen Null begrenzt ist und längere Wartezeiten vorkommen, ist die Verteilung rechtsschief.
3. **Fläche unter der Kurve:**
   (d) Die Gesamtzahl der Fälle entspricht der Fläche unter der Kurve, was zur Abschätzung der Gesamtzahl herangezogen werden kann.

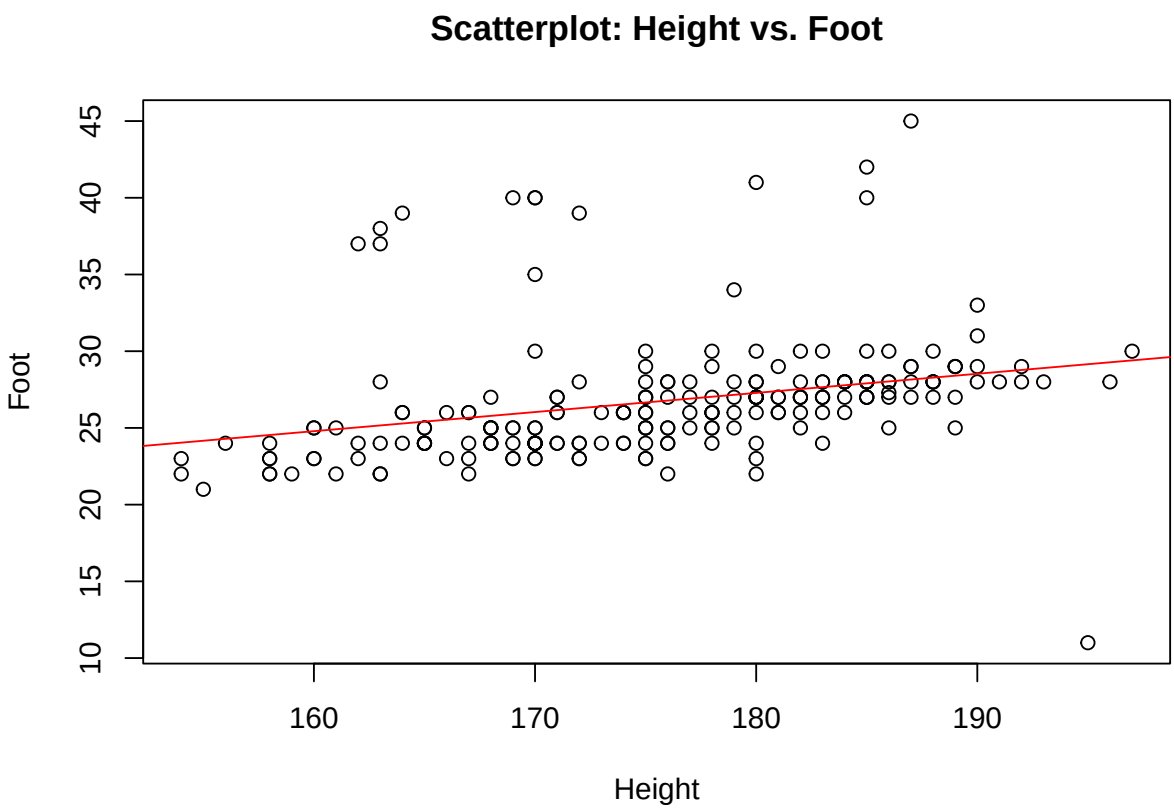
Aufgabe 16: Streudiagramme, Regressionsgerade und Korrelation

Aufgabenstellung:
16. Vergleichen Sie die folgenden Datensätze mit Hilfe eines Streudiagramms, einer Anpassungsgeraden und des Korrelationskoeffizienten:
(a) `height` und `foot`
(b) `hair` und `foot`
(c) `distance` und `siblings`

Schritt-für-Schritt-Erklärung:

1. **Streudiagramm erstellen und lineare Regression berechnen:**
   Beispiel (a):

```
plot(data$height, data$foot,
     main = "Scatterplot: Height vs. Foot",
     xlab = "Height", ylab = "Foot")
lm_model_a <- lm(foot ~ height, data = data)
abline(lm_model_a, col = "red")
```

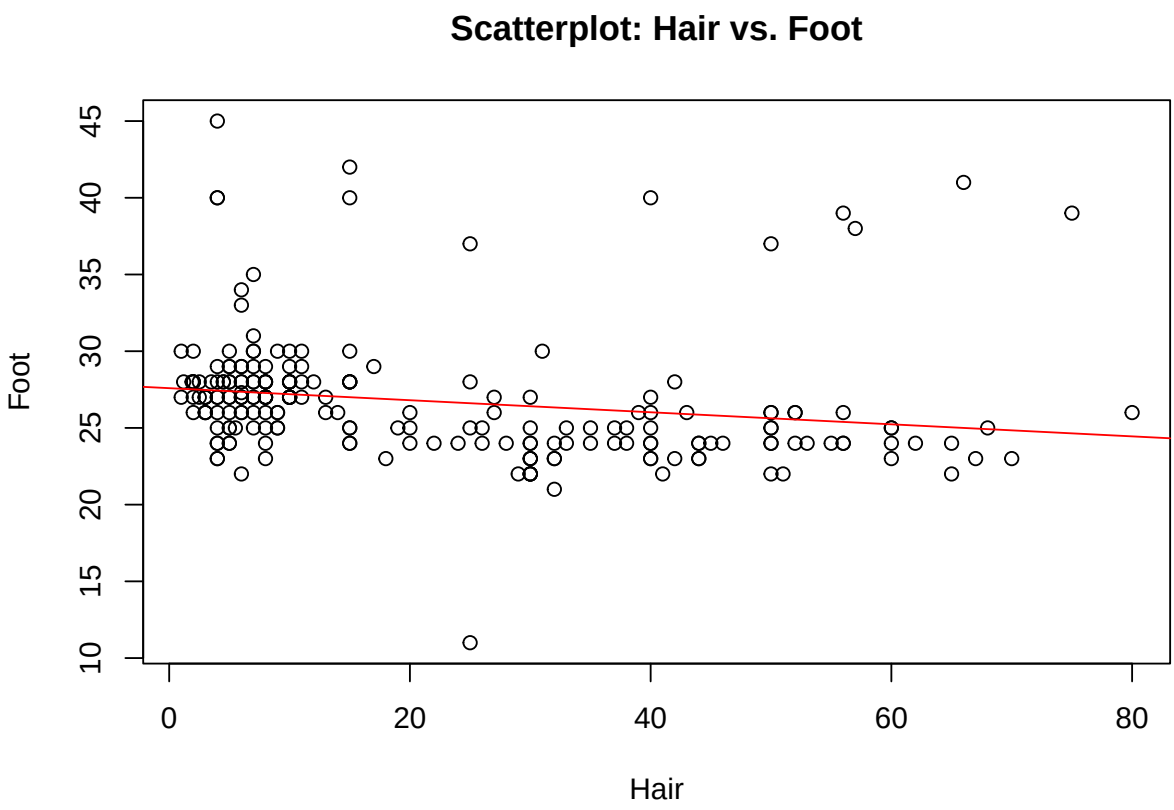



```r
correlation_a <- cor(data$height, data$foot, use = "complete.obs")
cat("Korrelation (height, foot):", correlation_a, "\n")
```

```
## Korrelation (height, foot): 0.2839648
```

2. **Analog für (b) und (c):**

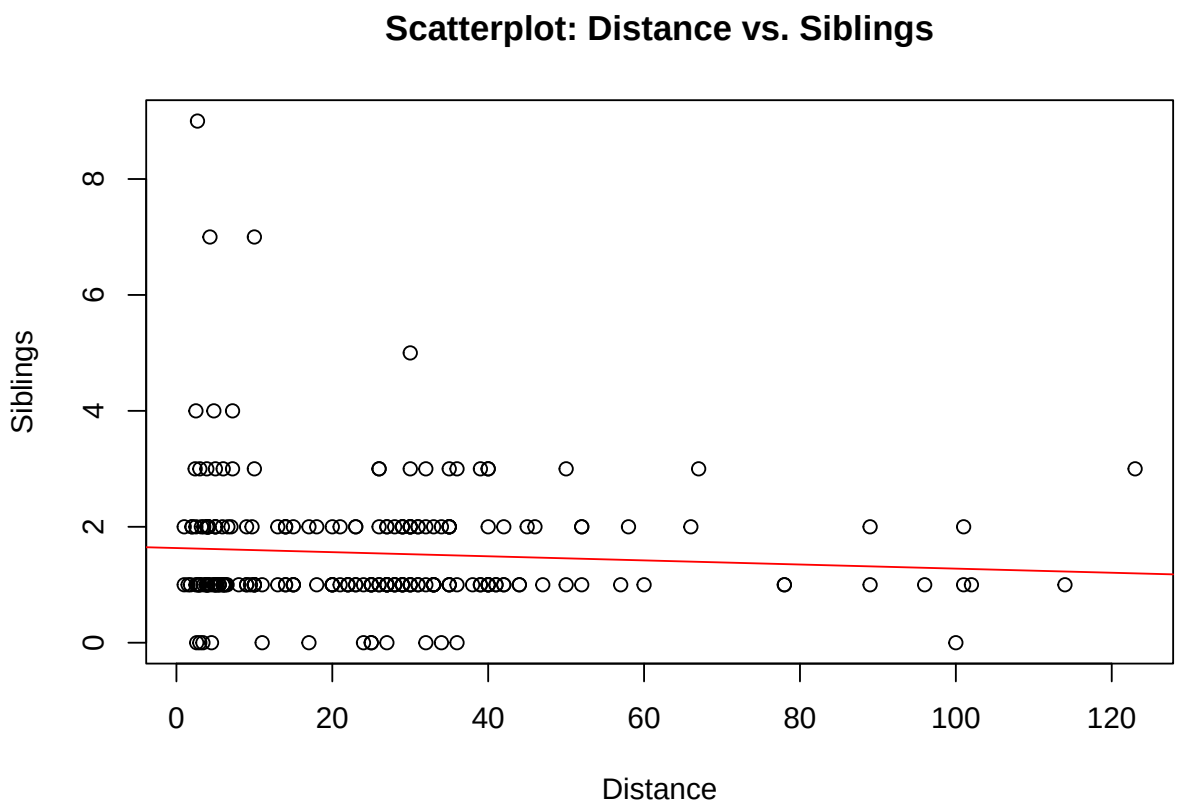
```r
# (b) hair vs. foot
plot(data$hair, data$foot,
     main = "Scatterplot: Hair vs. Foot",
     xlab = "Hair", ylab = "Foot")
lm_model_b <- lm(foot ~ hair, data = data)
abline(lm_model_b, col = "red")
```




```
correlation_b <- cor(data$hair, data$foot, use = "complete.obs")
cat("Korrelation (hair, foot):", correlation_b, "\n")
```

```
## Korrelation (hair, foot): -0.1886152
```

```
# (c) distance vs. siblings
plot(data$distance, data$siblings,
     main = "Scatterplot: Distance vs. Siblings",
     xlab = "Distance", ylab = "Siblings")
lm_model_c <- lm(siblings ~ distance, data = data)
abline(lm_model_c, col = "red")
```




```r
correlation_c <- cor(data$distance, data$siblings, use = "complete.obs")
cat("Korrelation (distance, siblings):", correlation_c, "\n")
```

```
## Korrelation (distance, siblings): -0.07244853
```

3. **Erwartete Ergebnisse:**
   - (a) Es besteht ein positiver Zusammenhang (z.B. r 0.28) mit einer Anpassungsgleichung wie: `foot = 0.12 * height + 4.84`
   - (b) Ein negativer Zusammenhang (z.B. r -0.19) mit: `foot = -0.04 * hair + 27.59`
   - (c) Kein eindeutiger Zusammenhang (z.B. r -0.07)

Aufgabe 17: Zusammenhang zwischen Höhe und Haar – Geschlechtervergleich

Aufgabenstellung:
17. Untersuchen Sie den Zusammenhang zwischen `height` und `hair` in Bezug auf das Geschlecht.
(a) Berechnen Sie die Korrelation für den Gesamtdatensatz.
(b) Berechnen Sie die Korrelation nur für Frauen und vergleichen Sie.
(c) Berechnen Sie die Korrelation nur für Männer und vergleichen Sie.

Schritt-für-Schritt-Erklärung:

1. **Gesamtkorrelation berechnen:**

```
corr_total <- cor(data$height, data$hair, use = "complete.obs")
cat("Korrelation (gesamt):", corr_total, "\n")
```

```
## Korrelation (gesamt): -0.6186634
```

2. **Korrelation für Frauen:**

```
corr_female <- cor(data$height[data$gender == "Female"],
                   data$hair[data$gender == "Female"], use = "complete.obs")
cat("Korrelation (Frauen):", corr_female, "\n")
```

```
## Korrelation (Frauen): 0.006966206
```

3. **Korrelation für Männer:**

```
corr_male <- cor(data$height[data$gender == "Male"],
                 data$hair[data$gender == "Male"], use = "complete.obs")
cat("Korrelation (Männer):", corr_male, "\n")
```

```
## Korrelation (Männer): 0.1163308
```

4. **Erwartete Ergebnisse:**
   - Gesamtkorrelation: ca. -0.62
   - Nur Frauen: ca. 0.01
   - Nur Männer: ca. 0.12

ies illustriert, wie der aggregierte Datensatz andere Korrelationen ergeben kann als getrennte Gruppen (d.h. wir haben hier einen Fall von "Simpson's paradox").

Aufgabe 18: Kovarianz bei identischen Datensätzen

Aufgabenstellung:
18. Was laesst sich über die Formel für die Kovarianz sagen, wenn beide Listen aus denselben Daten bestehen?

Schritt-für-Schritt-Erklärung:

- Wenn Sie die Kovarianz eines Datensatzes mit sich selbst berechnen, entspricht dies der Varianz.
- Formal gilt:

$$s_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  also ist $\mathrm{Cov}(x, x) = \mathrm{Var}(x)$.

Aufgabe 19: Schätzung der Korrelation aus Streudiagrammen

Aufgabenstellung:
19. Schätzen Sie die Korrelationen aus den folgenden Streudiagrammen (Diagramme a bis i).

Schritt-für-Schritt-Erklärung:

- Vergleichen Sie visuell die Streuung und Richtung der Punktwolken.
- Diskutieren Sie, wie eng bzw. locker die Punkte um eine gedachte Regressionslinie liegen.
- Erwartete Schätzungen (basierend auf den Lösungen):
  - (a) r 0.9
  - (b) r -0.6
  - (c) r 0.3
  - (d) r 0
  - (e) r 0.2
  - (f) r 0.5
  - (g) r -0.3
  - (h) r -0.8
  - (i) r 0

Aufgabe 20: Vergleich von Streudiagrammen und Stichprobenumfang

Aufgabenstellung:
20. Ordnen Sie die folgenden vier Streudiagramme in aufsteigender Reihenfolge des Korrelationswertes.

Schritt-für-Schritt-Erklärung:

- Obwohl die Diagramme optisch unterschiedlich erscheinen können, liegt der tatsächliche Korrelationswert in allen Fällen bei r = 0.7.
- Der Unterschied entsteht durch variierenden Stichprobenumfang (von links nach rechts verdoppelt sich dieser).

Aufgabe 21: Schätzung der Korrelation bei nicht-linearem Zusammenhang

Aufgabenstellung:
21. Betrachten Sie die folgende Punktwolke. Wie hoch würden Sie den Korrelationskoeffizienten schätzen?

Schritt-für-Schritt-Erklärung:

- Aufgrund eines offensichtlichen quadratischen (nicht-linearen) Zusammenhangs wird der lineare Korrelationskoeffizient nur einen geringen linearen Effekt erfassen.
- (Lösung: r ≈ -0.17, sagt hier aber nicht viel aus)

Aufgabe 22: Analyse der Aktienmärkte 2008

Aufgabenstellung:

22. Die Tabelle "Stock 2008" enthält prozentuale Rückgänge der Aktien zwischen Januar und Anfang Oktober.

(a) Bestimmen Sie Mittelwert und Median.

(b) Bestimmen Sie das erste und dritte Quartil.

(c) Enthalten die Daten Ausreisser? Konstruieren Sie ein Boxplot.

(d) Welches Perzentil würden Sie für Belgien melden?

Schritt-für-Schritt-Erklärung:

1. **Berechnungen:**

```
# Daten aus WDDA_03 laden
data <- read_excel("../data/WDDA_03.xlsx", sheet = "Stock_2008")

mean_stock <- mean(data$Fall, na.rm = TRUE)
median_stock <- median(data$Fall, na.rm = TRUE)
quantiles_stock <- quantile(data$Fall, probs = c(0.25, 0.75), na.rm = TRUE)

cat("Mittelwert Stock 2008:", mean_stock, "\n")
```

```
## Mittelwert Stock 2008: 38.96786
```

```
cat("Median Stock 2008:", median_stock, "\n")
```

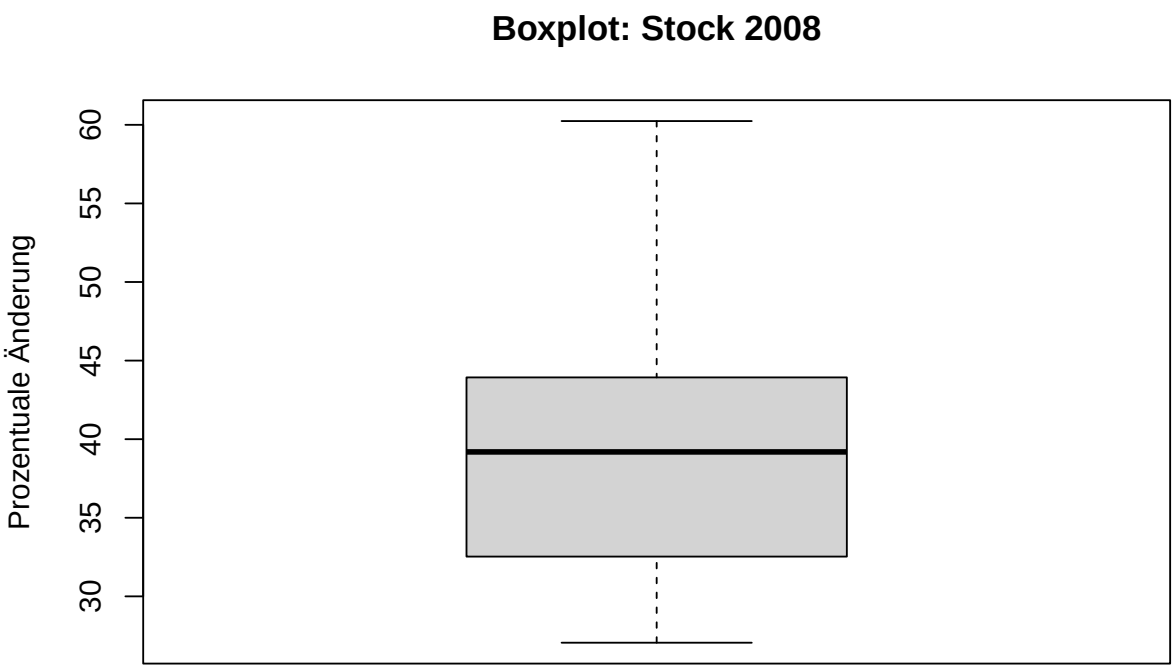
```
## Median Stock 2008: 39.19
```

```
cat("Q1:", quantiles_stock[1], "Q3:", quantiles_stock[2], "\n")
```

```
## Q1: 32.61 Q3: 43.815
```

2. **Boxplot erstellen:**

```
boxplot(data$Fall,
        main = "Boxplot: Stock 2008",
        ylab = "Prozentuale Änderung")
```



Es gibt keine Ausreisser. Der Boxplot zeigt, dass alle Werte innerhalb der Whiskers liegen.

3. **Perzentil für Belgien berechnen:**

```r
# Belgiens Wert identifizieren
belgium_value <- data$Fall[data$Country == "Belgium"]
cat("Belgiens Wert:", belgium_value, "\n")
```

```
## Belgiens Wert: 43.7
```

```r
# Perzentil berechnen
belgium_percentile <- mean(data$Fall < belgium_value, na.rm = TRUE) * 100
cat("Perzentil für Belgien:", round(belgium_percentile), "\n")
```

```
## Perzentil für Belgien: 71
```

Die Berechnung zeigt, dass Belgien im 71. Perzentil liegt. Hier ist eine intuitive Erklärung:

- Ein Perzentil gibt an, welcher Prozentsatz der Daten unter einem bestimmten Wert liegt
- Wir berechnen es, indem wir zählen, wie viele Länder einen niedrigeren Wert als Belgien haben, und teilen diese Anzahl durch die Gesamtzahl der Länder
- Die Formel `mean(data$Fall < belgium_value)` nutzt eine elegante Eigenschaft von R: Der Ausdruck `data$Fall < belgium_value` erzeugt einen Vektor von TRUE/FALSE-Werten

- In R wird TRUE als 1 und FALSE als 0 interpretiert, wenn wir den Mittelwert berechnen
- Daher gibt `mean(data$Fall < belgium_value)` den Anteil der Werte, die kleiner als Belgiens Wert sind
- Multipliziert mit 100 erhalten wir das Perzentil

Das 71. Perzentil bedeutet, dass Belgien einen höheren Aktienrückgang als 71% der anderen Länder erlebte, aber einen niedrigeren Rückgang als die restlichen 29%. Dies positioniert Belgien im oberen Drittel der am stärksten betroffenen Länder während der Finanzkrise 2008.

Aufgabe 23: Korrelation der Aktienindizes

Aufgabenstellung:
23. Die Tabelle "DAX CAC" enthält die Indexstände des deutschen DAX und des französischen CAC 40 für die ersten 10 Wochen 2015.
(a) Berechnen Sie den Korrelationskoeffizienten.
(b) Wie stark ist der lineare Zusammenhang?

Schritt-für-Schritt-Erklärung:

1. **Berechnungen:**

```
# Daten aus WDDA_03 laden
data <- read_excel("../data/WDDA_03.xlsx", sheet = "DAX_CAC")

# Angenommen, die Spalten heissen 'DAX' und 'CAC40'
corr_index <- cor(data$DAX, data$CAC, use = "complete.obs")
cat("Korrelationskoeffizient (DAX, CAC40):", corr_index, "\n")
```

```
## Korrelationskoeffizient (DAX, CAC40): 0.9231529
```

2. **Ergebnis:**
   Erwartete Lösung: r $\approx$ 0.923, was auf einen sehr starken positiven linearen Zusammenhang hindeutet.

Aufgabe 24: Zusammenhang zwischen Factual Reporting und Bias

Aufgabenstellung:
24. Betrachten Sie die Grafik (bei mediabiasfactcheck.com). Welche Art von Zusammenhang, falls vorhanden, könnte zwischen den Variablen Factual Reporting und Bias bestehen?

Schritt-für-Schritt-Erklärung:

- Diskutieren Sie qualitativ:
  - Da die Bias-Skala in beide Richtungen offen ist, ist es sinnvoll, den absoluten Bias zu betrachten.
  - Es wird erwartet, dass ein extremer Bias mit einem niedrigeren Niveau sachlicher Berichterstattung einhergeht.

Aufgabe 25: Analyse der ideologischen Verteilungen

Aufgabenstellung:
25. Die Grafik aus The Economist zeigt die ideologischen Verteilungen von Demokraten und Republikanern über die Zeit.
(a) Erläutern Sie, was mit der Verteilung der Demokraten passiert ist.
(b) Erläutern Sie, was mit der Verteilung der Republikaner passiert ist.
(c) Welche Partei hat im Laufe der Zeit extremere Positionen bezogen?
(d) Erklären Sie, wie das Diagramm die Polarisierung der amerikanischen Politik zeigt.
(e) Gibt es einen Zusammenhang zwischen den Standardabweichungen der beiden Parteien?

Schritt-für-Schritt-Erklärung:

1. **Analyse der Veränderungen:**
   - (a) Bei den Demokraten bleibt der Mittelwert relativ konstant, die Streuung verringert sich.
   - (b) Bei den Republikanern driftet der Mittelwert nach rechts, und die Variation nimmt zu – was darauf hindeutet, dass tendenziell Politiker mit härteren Ansichten aufgenommen werden.
2. **Vergleich der Extreme:**
   - (c) Es werden die Republikaner als Partei mit extremeren Positionen identifiziert.
3. **Polarisierung:**
   - (d) Das Auseinanderdriften der Mittelwerte und das Abnehmen der Überlappung verdeutlichen die zunehmende Polarisierung.
4. **Standardabweichungen:**
   - (e) Es scheint eine negative Assoziation zwischen den Standardabweichungen der beiden Parteien zu bestehen, was nahelegt, dass eine stärkere ideologische Ausbreitung bei einer Partei mit einer Konzentration um den Mittelwert bei der anderen einhergeht.