# Empirische Datenanalyse

*Roman Schulze*

*22 5 2020*

## Abstract

In diesem Dokument wird der *Fish* Datensatz genutzt. Dieser setzt sich aus 159 Beobachtungen und 8 Variablen zusammen. Insgesamt wurden 7 unterschiedliche Fischarten erfasst. Für jeden Fisch liegen spezifische Informationen bezüglich der Länge, des Gewichts, der Breite und der Höhe vor. Weiterhin ist der gehandelte Preis als Information verfügbar.

### Laden der Bibliotheken

```
# Lade notwendige Module
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)
```

### Einlesen der Daten

```
# definiere pfad
setwd("~/Desktop//Daten/Fish")

# lade Daten
data <- read.csv2("Fish.csv")

# Zusammenfassung der Daten
summary(data)
```

```
##       Species       Weight       Length1       Length2       Length3
##  Bream    :35   300    :  6   19     :  6   22     :  7   23.5   :  5
##  Parkki   :11   1000   :  5   20     :  5   35     :  6   22.5   :  3
##  Perch    :56   120    :  5   20.5   :  4   22.5   :  5   25     :  3
##  Pike     :17   500    :  5   22     :  4   40     :  5   28.9   :  3
##  Roach    :20   700    :  5   24     :  3   21     :  4   34     :  3
##  Smelt    :14   145    :  4   25.4   :  3   20     :  3   36.2   :  3
##  Whitefish: 6   (Other):129   (Other):134   (Other):129   (Other):139
```

```
##      Height         Width          Price
##  11.1366:  2   3.525  :  3   3.25   :  2
##  2.2139 :  2   1.1484 :  2   3.31   :  2
##  5.6925 :  2   3.624  :  2   4.06   :  2
##  6.11   :  2   4.144  :  2   4.36   :  2
##  9.6    :  2   4.335  :  2   4.95   :  2
##  1.7284 :  1   6.144  :  2   6.41   :  2
##  (Other):148   (Other):146   (Other):147
```

```r
# Struktur der Daten
str(data)
```

```
## 'data.frame':    159 obs. of  8 variables:
##  $ Species: Factor w/ 7 levels "Bream","Parkki",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Weight : Factor w/ 101 levels "0","10","100",..: 34 41 46 48 51 52 56 49 52 56 ...
##  $ Length1: Factor w/ 116 levels "10","10.1","10.4",..: 46 50 49 57 58 59 59 61 61 64 ...
##  $ Length2: Factor w/ 93 levels "10.5","10.6",..: 44 47 48 54 54 55 55 56 56 57 ...
##  $ Length3: Factor w/ 124 levels "10.8","11.6",..: 60 65 64 67 68 70 69 72 73 74 ...
##  $ Height : Factor w/ 154 levels "1.7284","1.7388",..: 20 30 26 35 29 37 41 34 40 42 ...
##  $ Width  : Factor w/ 152 levels "1.0476","1.1484",..: 66 76 89 81 102 96 106 88 93 97 ...
##  $ Price  : Factor w/ 151 levels "0.13","0.27",..: 75 35 6 57 55 13 19 150 143 10 ...
```

```r
# Konvertiere alle Variablen außer "Species" von Faktoren in numerische Werte
# Quelle: https://stat.ethz.ch/pipermail/r-help/2011-June/280173.html
data[, -1] <- lapply(data[, -1], function(x) as.numeric(levels(x))[x])

# Struktur der Daten nach der Anpassung
str(data) # Abgesehen von Species sind nun alle weiteren SPalten vom Typ "num"
```

```
## 'data.frame':    159 obs. of  8 variables:
##  $ Species: Factor w/ 7 levels "Bream","Parkki",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Weight : num  242 290 340 363 430 450 500 390 450 500 ...
##  $ Length1: num  23.2 24 23.9 26.3 26.5 26.8 26.8 27.6 27.6 28.5 ...
##  $ Length2: num  25.4 26.3 26.5 29 29 29.7 29.7 30 30 30.7 ...
##  $ Length3: num  30 31.2 31.1 33.5 34 34.7 34.5 35 35.1 36.2 ...
##  $ Height : num  11.5 12.5 12.4 12.7 12.4 ...
##  $ Width  : num  4.02 4.31 4.7 4.46 5.13 ...
##  $ Price  : num  3.25 12.3 0.49 2.76 2.48 1.27 1.75 9.34 8.09 1.02 ...
```

# 1.Aufgabe

Stellen Sie grafisch und tabellarisch dar, wie oft jede Fischart im Datensatz vorkommt und interpretieren Sie das Ergebnis.

```r
# Erstellen einer Häufigkeitstabelle
freq_table <- data %>% group_by(Species) %>% count()

# Anpassung der Spaltennamen
colnames(freq_table) <- c("Species", "abs_Count")

# Füge relative Häufigkeit hinzu: Diese ergibt sich, indem die absolute
# Anzahl jeder Spezie durch den gesamten Beobachtungsumfang dividiert wird
freq_table <- freq_table %>% mutate(rel_Count = abs_Count / nrow(data)) %>%
  mutate_if(is.numeric, round, 3)
```

```
## `mutate_if()` ignored the following grouping variables:
## Column `Species`
```

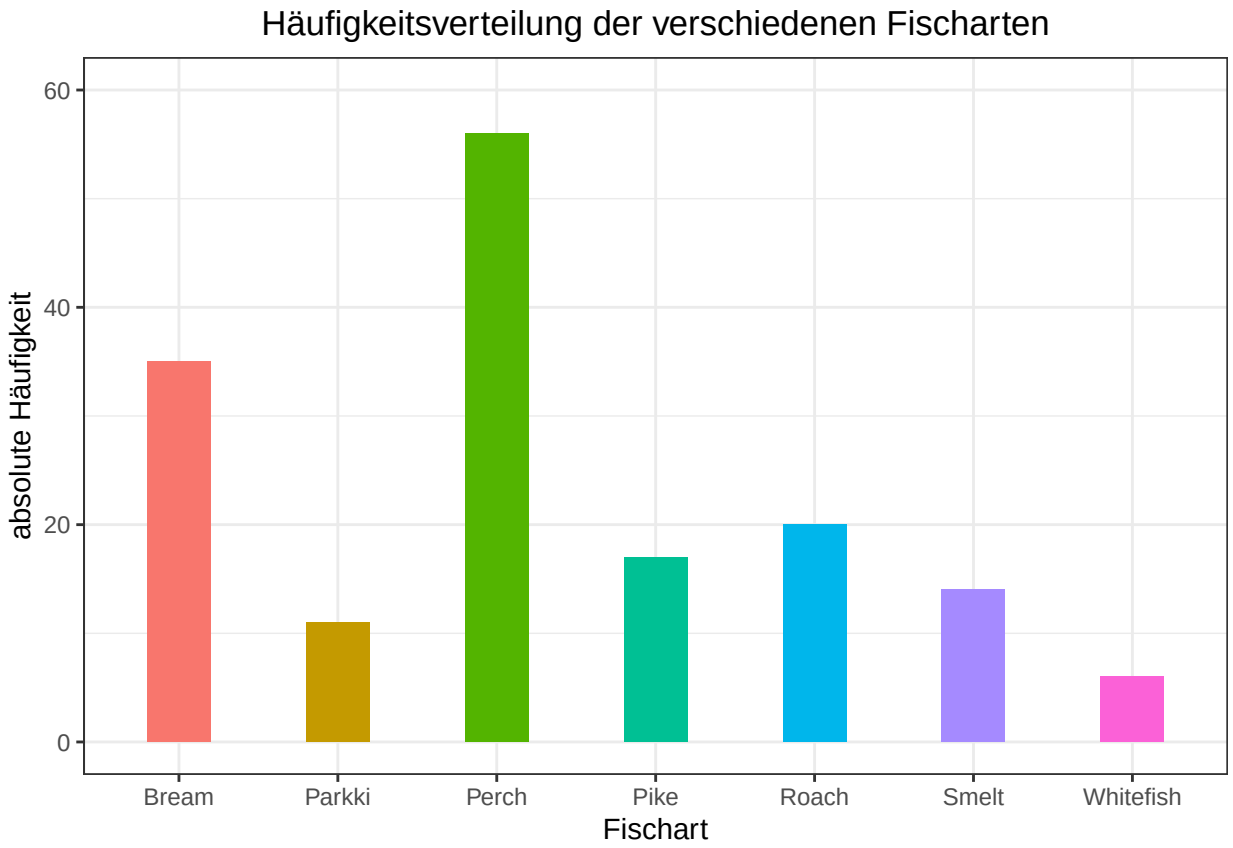
```
# Wiedergabe der Tabelle (tibble table)
print(freq_table)
```

```
## # A tibble: 7 x 3
## # Groups:   Species [7]
##   Species   abs_Count rel_Count
##   <fct>         <dbl>     <dbl>
## 1 Bream            35     0.22
## 2 Parkki           11     0.069
## 3 Perch            56     0.352
## 4 Pike             17     0.107
## 5 Roach            20     0.126
## 6 Smelt            14     0.088
## 7 Whitefish         6     0.038
```

```
# Berechne die durschnittle Anzahl jeder Spezie
round(mean(freq_table$abs_Count), 2) # 22.71
```

```
## [1] 22.71
```

```
# Erstellen einer Häufikeitstabelle

# Wähle den Datensatz und die Variablen
ggplot(data = freq_table) + aes(x = Species, y = abs_Count, fill = Species) +
  # Spezifiziere die Parameter des Barplots
  geom_bar(stat = "identity", width = 0.4) +
  # Anpassung des Stils
  theme_bw() +
  # Füge Titel hinzu
  ggtitle("Häufigkeitsverteilung der verschiedenen Fischarten") +
  # Zentriere Titel und entferne Legende
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none") +
  # x-Achse label
  xlab("Fischart") +
  # y-Achse label
  ylab("absolute Häufigkeit") +
  # y-Achse anpassen
  ylim(0, 60)
```




**Interpretation:**

Für die vorliegende Aufgabe wurde zunächst eine Häufigkeitstabelle angefertigt, die die absolute Häufigkeit jeder Fischart erfasst. Im zweiten Schritt wurde zur Visualisierung ein Balkendiagramm gewählt. In der Grafik ist die Fischart auf der x-Achse abgetragen und die entsprechende absolute Häufigkeit lässt sich der y-Achse entnehmen. Die durchschnittliche absolute Häufigkeit liegt bei näherungsweise 23. Insgesamt gibt es zwei Fischarten, die überdurchschnittlich häufig erfassst wurden. Dabei handelt es sich zum einen um die Spezie "Perch", die mit 56 erfassten etwas mehr als ein Drittel der gesamten Beoabchtungspunkte einnimmt. Als zweithäufigste Spezie mit 35 Beobachtungen folgt die Fischart "Bream". Insgesamt machen die beiden Fischarten mehr als die Hälfte aller Beobachtungspunkte im Datensatz aus. Selten erfasste Fischarten sind zum einen "Parkki" mit 11 Einträgen und der "Whitefish" mit lediglich sechs Beobachtungspunkten.

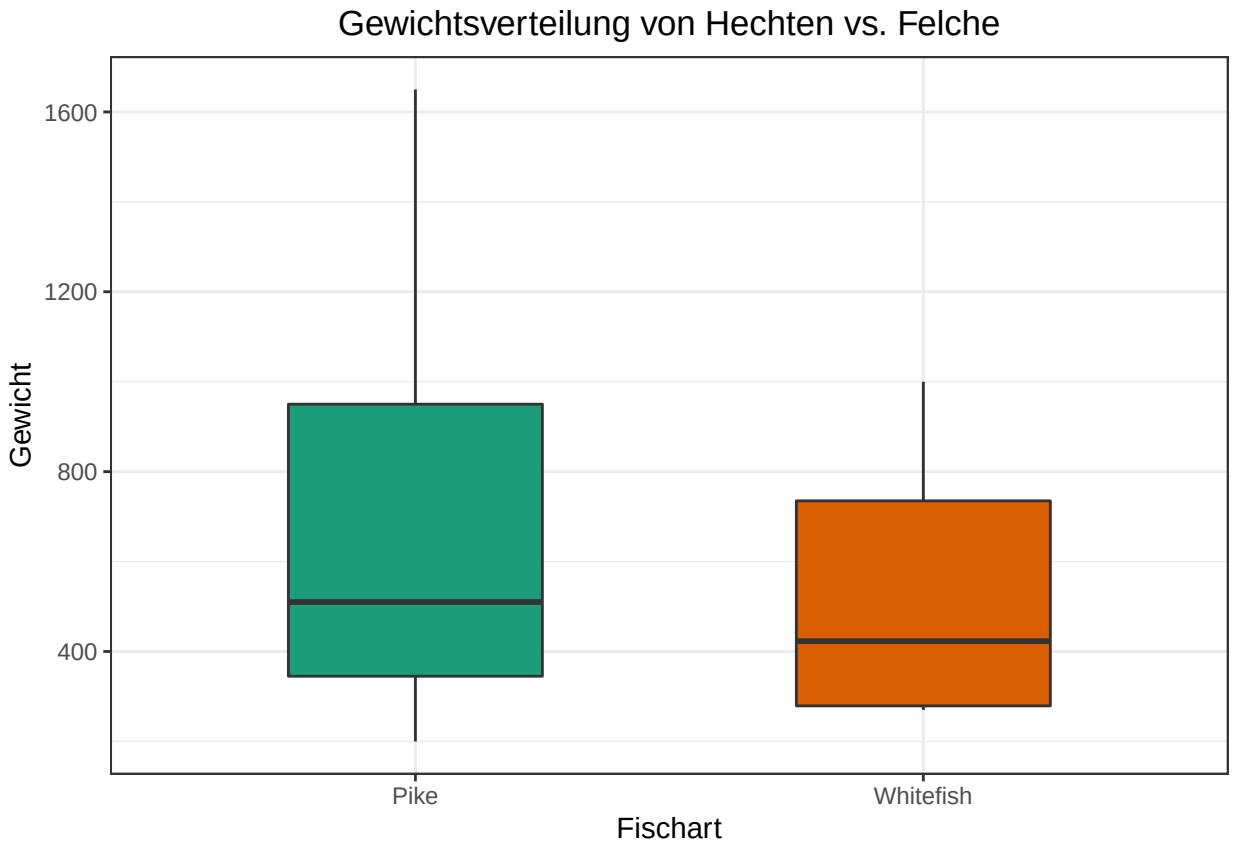
# 2.Aufgabe

### Aufgabe 2a)

Grafische Darstellung der Gewichtsverteilung von Hechten und Felchen: Um die Lage und Streubreite des Gewichts der beiden Fischarten zu vergleichen wird im Folgenden ein Boxplot verwendet.

```
# filtere die beiden Gruppen aus dem Orginaldatensatz
data_weight <- data %>% filter(Species == "Pike" | Species == "Whitefish")

# Veranschaulichung der Gewichtsverteilung anhand eines Boxplots
# fill Parameter dient der Farbgebung
ggplot(data = data_weight) + aes(x = Species, y = Weight, fill = Species) +
```

```r
  # Boxplot, width kontrolliert die Breite der Boxen
  geom_boxplot(width = 0.5) +
  # Anpassung des Stils
  theme_bw() +
  # Füge Titel hinzu
  ggtitle("Gewichtsverteilung von Hechten vs. Felche") +
  # Zentriere Titel und entferne die Legende
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none") +
  # x-Achse
  xlab("Fischart") +
  # y-Achse
  ylab("Gewicht") +
  # Passe die Farben nochmal an
  scale_fill_brewer(palette = "Dark2")
```

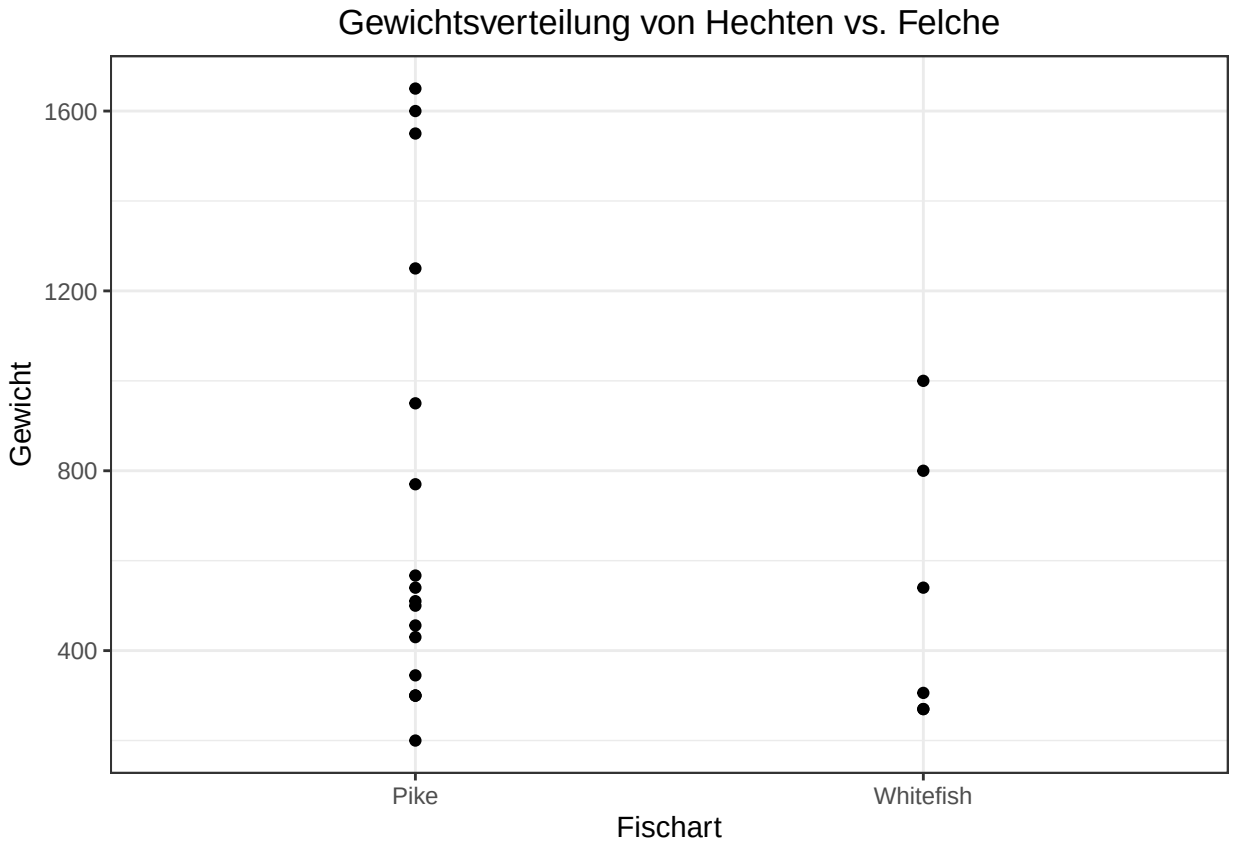



**Interpretation:**

In der vorliegenden Grafik ist das Gewicht auf der y-Achse abgetragen und die beiden Spezien von Interesse auf der x-Achse. Beim Boxplot werden die mittleren 50 Prozent der Verteilug durch die Boxen repräsentiert. Die Whisker geben Auskunft über die Ausreißer. Der Median wird durch horizontale schwarze Linie innierhalb der Box veranschaulicht. Es lässt sich festhalten, dass der Median für die Variable Gewicht bei den Hechten (Pikes) höher ausfällt als bei den Felchen. Bei den Felchen liegt er bei etwas über 400 Gramm und bei den Hechten bei ca. 500 Gramm. Auch der Interquartilsabstand fällt bei den Hechten größer aus, als bei den Felchen. Die Außreißer, repräsentiert durch die Whsiker zeigen, dass auch die Streuung des Gewichts bei den Hechten in der vorliegenden Stichprobe größer ausfällt, als bei den Felchen. Angemerkt sei hier, dass die Stichprobengröße der beiden Subpopulationen relativ klein sind (vgl. die absoluten Häufigkeiten

aus Aufgabe 1) und die dargestellten Verteilungen bei einer umfangreichereicheren Stichprobe von den vorliegenden Ergebnissen abweichen können. Weiterhin besteht die Möglichkeit ein Einzelwertdiagramm anzufertigen, dass die einzelnen Werte der Beobachtungspunkt wiedergibt:

```
# filtere die beiden Gruppen aus dem Orginaldatensatz
data_weight <- data %>% filter(Species == "Pike" | Species == "Whitefish")

# Veranschaulichung der Gewichtsverteilung anhand eines Einzelwertdiagramms
ggplot(data = data_weight) + aes(x = Species, y = Weight) +
  # Boxplot, width kontrolliert die Breite der Boxen
  geom_point() +
  # Anpassung des Stils
  theme_bw() +
  # Füge Titel hinzu
  ggtitle("Gewichtsverteilung von Hechten vs. Felche") +
  # Zentriere Titel und entferne die Legende
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none") +
  # x-Achse
  xlab("Fischart") +
  # y-Achse
  ylab("Gewicht")
```




Um zu überprüfen, ob sich das Gewicht der beiden Gruppen stastistisch signifikant voneinander unterscheidet soll im folgenden ein t-Test auf Mittwelwertgleichheit durchgeführt werden. Die Nullhypothese postuliert, dass die Mittlwerte der beiden Spezien identisch sind, wohingegen die Alternativhypothese das Gegenteil annimmt.

```
# t-test auf Mittwelwertgleichheit
```

```r
# Filtere die Variable Gewicht für die relevanten Fischarten Pike und Whitefish
Pike_weight <- data_weight %>% filter(Species == "Pike") %>% select(Weight)
Whitefish_weight <- data_weight %>% filter(Species == "Whitefish") %>% select(Weight)

# Zweiseitiger t-test
t.test(Pike_weight, Whitefish_weight, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  Pike_weight and Whitefish_weight
## t = 1.0777, df = 14.396, p-value = 0.2989
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -184.9130  560.3247
## sample estimates:
## mean of x mean of y
##  718.7059  531.0000
```

**Interpretation:**

Der geschätzte Mittelwert der Hechte liegt bei etwa 719 Gramm, wohingegen das durchschnittliche Gewicht der Felche 531 Gramm beträgt. Gegeben einem Signifikanzniveau von 0.05, lässt sich festhalten, dass die Nullhypothese nicht verworfen wird, wenn der p-Wert größer als 0.05 ist. Somit folgt: Basierend auf den Stichprobendaten lässt sich die Nullhypothese nicht verwerfen ($0.2989 > 0.05$). Wie bereits erwähnt, soll auch hier nochmal darauf hingewiesen werden, dass der Stichprobenumfang der Teilpopulation sehr gering ist und die vorliegend geschätzten Mittelwerte unter Umständen von den tatsächlichen Mittelwerten der jeweiligen Grundgesamtheit abweichen können. Um sich bei der Interpretation der Ergebnisse sicherer zu sein, bestünde eine Lösungsansatz darin, den Umfang der Zufallsstichproben zu vergrößern.

# 3.Aufgabe

Besteht ein Zusammenhang zwischen Höhe und Gewicht eines Fischs. Gilt das als Ganzes und für die einzelnen Sorten?

Um statistisch zu testen, ob ein Zusammenhang zwischen den Variablen besteht soll im Folgenden ein 1-Stichprobentest durchgeführt werden. Dieser unterstellt in der Nullhypothese, dass die wahre Korrelation in der Grundgesamtheit gleich null ist. Mit anderen Worten: In der Population besteht kein linearer Zusammenhang zwischen zwei Merkmalen. Die Alternativhypothese unterstellt, dass die wahre Korrelation ungleich null ist.

**Test für den gesamten Datensatz:**

```r
# Test auf linearen Zusammenhang zwischen Gewicht und Höhe für den gesamten Datensatz
cor.test(data[, "Weight"], data[, "Height"])
```

```
##
##  Pearson's product-moment correlation
##
## data:  data[, "Weight"] and data[, "Height"]
## t = 13.164, df = 157, p-value < 2.2e-16
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6409616 0.7908323
## sample estimates:
##       cor
## 0.7243453
```

**Interpretation:**

Die Nullhypothese, $r = 0.7243453$ stammt aus einer Grundgesamtheit mit $r$=0 kann zu einem Signifikanzneveau von 0.05 verworfen werden. Wie das Testergebnis zeigt, ist der p-Wert kleiner als 0.05 und das 95 Prozent Konfidenzintervall schließt die Null nicht mit ein. Es lässt sich festhalten, dass die Korrelation statistisch signifikant von null abweicht.

**Test für Subgruppen:**

```
# Test auf linearen Zusammenhang zwischen Gewicht und Höhe für die Spezie Bream

# Filtere Bream Daten
data_Bream <- data %>% filter(Species == "Bream")

# Test auf linearen Zusammenhang zwischen Gewicht und Höhe für Spezie Bream
cor.test(data_Bream[, "Weight"], data_Bream[, "Height"])
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_Bream[, "Weight"] and data_Bream[, "Height"]
## t = 20.989, df = 33, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9303046 0.9821016
## sample estimates:
##       cor
## 0.9645275
```

```
# Test auf linearen Zusammenhang zwischen Gewicht und Höhe für die Spezie Parkki

# Filtere Parkki Daten
data_Parkki <- data %>% filter(Species == "Parkki")

# Test auf linearen Zusammenhang zwischen Gewicht und Höhe für Spezie Parkki
cor.test(data_Parkki[, "Weight"], data_Parkki[, "Height"])
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_Parkki[, "Weight"] and data_Parkki[, "Height"]
## t = 10.88, df = 9, p-value = 1.766e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8635093 0.9908791
## sample estimates:
```

```
##       cor
## 0.9640228
```

```r
# Test auf linearen Zusammenhang zwischen Gewicht und Höhe für die Spezie Perch

# Filtere Perch Daten
data_Perch <- data %>% filter(Species == "Perch")

# Test auf linearen Zusammenhang zwischen Gewicht und Höhe für Spezie Perch
cor.test(data_Perch[, "Weight"], data_Perch[, "Height"])
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_Perch[, "Weight"] and data_Perch[, "Height"]
## t = 28.553, df = 54, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9465301 0.9814584
## sample estimates:
##       cor
## 0.9684407
```

```r
# Test auf linearen Zusammenhang zwischen Gewicht und Höhe für die Spezie Pike

# Filtere Pike Daten
data_Pike <- data %>% filter(Species == "Pike")

# Test auf linearen Zusammenhang zwischen Gewicht und Höhe für Spezie Pike
cor.test(data_Pike[, "Weight"], data_Pike[, "Height"])
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_Pike[, "Weight"] and data_Pike[, "Height"]
## t = 10.341, df = 15, p-value = 3.209e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8289499 0.9772485
## sample estimates:
##       cor
## 0.9364747
```

```r
# Test auf linearen Zusammenhang zwischen Gewicht und Höhe für die Spezie Roach

# Filtere Roach Daten
data_Roach <- data %>% filter(Species == "Roach")

# Test auf linearen Zusammenhang zwischen Gewicht und Höhe für Spezie Roach
cor.test(data_Roach[, "Weight"], data_Roach[, "Height"])
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_Roach[, "Weight"] and data_Roach[, "Height"]
## t = 8.373, df = 18, p-value = 1.273e-07
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7426562 0.9568411
## sample estimates:
##       cor
## 0.8920221
```

```r
# Test auf linearen Zusammenhang zwischen Gewicht und Höhe für die Spezie Whitefish

# Filtere Whitefish Daten
data_Whitefish <- data %>% filter(Species == "Whitefish")

# Test auf linearen Zusammenhang zwischen Gewicht und Höhe für Spezie Whitefish
cor.test(data_Whitefish[, "Weight"], data_Whitefish[, "Height"])
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_Whitefish[, "Weight"] and data_Whitefish[, "Height"]
## t = 9.034, df = 4, p-value = 0.0008317
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7937334 0.9975146
## sample estimates:
##       cor
## 0.9763596
```

```r
# Test auf linearen Zusammenhang zwischen Gewicht und Höhe für die Spezie Smelt

# Filtere Smelt Daten
data_Smelt <- data %>% filter(Species == "Smelt")

# Test auf linearen Zusammenhang zwischen Gewicht und Höhe für Spezie Smelt
cor.test(data_Smelt[, "Weight"], data_Smelt[, "Height"])
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_Smelt[, "Weight"] and data_Smelt[, "Height"]
## t = 11.958, df = 12, p-value = 5.026e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8767474 0.9877211
## sample estimates:
##       cor
## 0.9605119
```

**Interpretation:**

Der Korrelationstest, zeigt für sämtliche Spezien einen p-Wert kleiner als 0.05. Für keine Subgruppe schließt das jeweilig ausgegebene Konfidenzintervall die Null mit ein, sodass in jedem Fall die Nullhypothese verworfen werden kann. Es lässt sich somit festhalten, dass die Korrelation der Grundgesamtheit für jede Spezie zwischen Höhe und Gewicht statistisch signifikant von null abweicht.

# 4.Aufgabe

### 4a) Multiples Regressionsmodell

Bevor ein multiples Regressionsmodell geschätzt wird sollen einige Vorüberlegungen hinsichtlich der Variablenselktion durchgeführt werden. Essentielle Faktoren sind:

1. Prädiktoren sollten hohe Validitäten besitzen.
2. Prädiktoren sollten möglichst unkorreliert sein.
3. Das Modell soll minimal sein, also so wenige Terme wie möglich enthalten.
4. Das Modell soll die Variationen der Variable Preis so gut wie möglich erklären.

Zunächst soll überprüft werden, wie stark die metrischen Variablen mit der Preisvariable und untereinander korrelieren. Dafür wird der Pearson Korrelationskoeffizient herangezogen und eine Korrelationstabelle erstellt. Da sich der Pearson Korrelationskoeffizient nur für metrische Variablen ausgeben lässt, wird die Variable "Spezie" vorliegend nicht berücksichtigt. Der Korrelationskoeffizient ist standardisiert und liegt im Intervall (-1, 1), wobei ein Wert von 1 einen perfekt positiven linearen Zusammenhang ausdeück und -1 einen perfekt negativen Zusammenhang darstellt.

```
# Erstelle eine Korrelationstabelle, um die Abhängigkeiten der Variablen zu prüfen
corr <- round(cor(data[, -1]), 1)

# Veranschaulichung der Tabelle
print(corr)
```

```
##         Weight Length1 Length2 Length3 Height Width Price
## Weight     1.0     0.9     0.9     0.9    0.7   0.9   0.5
## Length1    0.9     1.0     1.0     1.0    0.6   0.9   0.5
## Length2    0.9     1.0     1.0     1.0    0.6   0.9   0.5
## Length3    0.9     1.0     1.0     1.0    0.7   0.9   0.5
## Height     0.7     0.6     0.6     0.7    1.0   0.8   0.3
## Width      0.9     0.9     0.9     0.9    0.8   1.0   0.4
## Price      0.5     0.5     0.5     0.5    0.3   0.4   1.0
```

**Interpretation:**

1. Korrelation Preis und Prädiktor: Der Korrelationstabelle lässt sich entnehmen, dass die Variablen Länge (in allen drei Ausführungen) und die Variable Gewicht mit einem Korrelationskoeffizient von 0.5 von allen metrischen Merkmalen den stärksten Zusammenhang mit der Preisvariable ausfweisen. Anschließend folgt die Variable Weite mit einem Korrelationskoeffizient von 0.4. Die Variable Höhe weist vorliegend mit einem Wert von 0.3 den geringsten linearen Zusammenhang zur Preisvariable auf.
2. Korreltion der Prädiktoren untereinander: Die Prädiktoren weisen untereinander eine recht hohe Korrelation bzw. positive lineare Abhängigkeit auf. Insgesamt findet sich über sämtliche Prädiktorkonstellationen kein Korreltationskoeffizient der unter 0.7 liegt. Die unterschiedlichen Längenmaße sind perfekt positiv miteinander korreliert. Weiterhin weisen die Weite eines Fisches und die Länge mit einem Korrelationskoeffizient von 0.9 eine stark positive Korrelation auf.

Im Folgenden soll überpürft werden inwiefern der Preis mit der Spzie statistisch zusammenhängt. Dafür soll ein einfaches Regressionsmodell geschätzt werden.

0. Zusammenhang Preis und Spezie

```
# Zusammenfassung der einfachen linearen Regression
summary(lm(Price ~ Species, data = data))
```

```
##
```

```
## Call:
## lm(formula = Price ~ Species, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.213  -3.823  -0.742   1.982  15.119
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.3251     0.8849   8.278 5.97e-14 ***
## SpeciesParkki      -1.5133     1.8096  -0.836  0.40431
## SpeciesPerch       -1.4244     1.1280  -1.263  0.20861
## SpeciesPike         4.2984     1.5477   2.777  0.00617 **
## SpeciesRoach       -1.5231     1.4675  -1.038  0.30094
## SpeciesSmelt       -3.3023     1.6555  -1.995  0.04786 *
## SpeciesWhitefish    5.6815     2.3132   2.456  0.01517 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.235 on 152 degrees of freedom
## Multiple R-squared:  0.1642, Adjusted R-squared:  0.1312
## F-statistic: 4.976 on 6 and 152 DF,  p-value: 0.0001112
```

**Interpretation:**

Die Variable Spezie hat einen signifikanten Einfluss auf die zu erklärende Variable Preis, was sich dem p-Wert der F-Statistik entnehmen lässt.Die Koeffizienten sind in Bezug auf die Referenzkategorie "Bream" zu interpretieren. Das Modell weist ein adjustiertes R^2 von 0.1312 auf.

### Schlussfolgerungen für das multiple Regressionsmodell

In Anlehnung auf die vier aufgelisteten Punkte sollen im ersten Schritt die Variablen "Gewicht" und "Spezie" für das multiple Regressionsmodell ausgewählt werden. Weiterhin erscheint es mir intuitiv nachvollziehbar den Preis eines Fisches am Gewicht auszumachen und dabei jedoch zu berücksichtigen um welche Fischssorte es sich handelt. Einige Fischsorten werden seltener gefangen, wodurch ein sich ein geringeres Angebot ergibt, dass unabhängig vom Gewicht zu in einem höheren Preis führen sollte:

```r
# Schätze multiples Regressionsmodell
model <- lm(Price ~ Weight + Species, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = Price ~ Weight + Species, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.6525  -2.6270  -0.2342   2.2302  14.9980
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.539630   1.148537   2.211  0.02853 *
## Weight            0.007746   0.001330   5.823 3.36e-08 ***
```

```
## SpeciesParkki    2.073013   1.752484   1.183  0.23871
## SpeciesPerch     0.400374   1.069688   0.374  0.70871
## SpeciesPike      3.517022   1.409607   2.495  0.01367 *
## SpeciesRoach     2.084636   1.467678   1.420  0.15756
## SpeciesSmelt     1.396641   1.704166   0.820  0.41377
## SpeciesWhitefish 6.354072   2.100476   3.025  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.747 on 151 degrees of freedom
## Multiple R-squared:  0.3174, Adjusted R-squared:  0.2858
## F-statistic: 10.03 on 7 and 151 DF,  p-value: 2.887e-10
```

Anstelle des Gewichts ließe sich auch eine der Längenvariablen nutzen. Dieses Modell weist jedoch ein leicht niedrigeres adjustiertetes R^2 auf, als wenn man die Gewichtsvariable berücksichtigt:

```
# Zusammenfassung der einfachen linearen Regression
model1 <- lm(Price ~ Length3 + Species, data = data)
summary(model1)
```

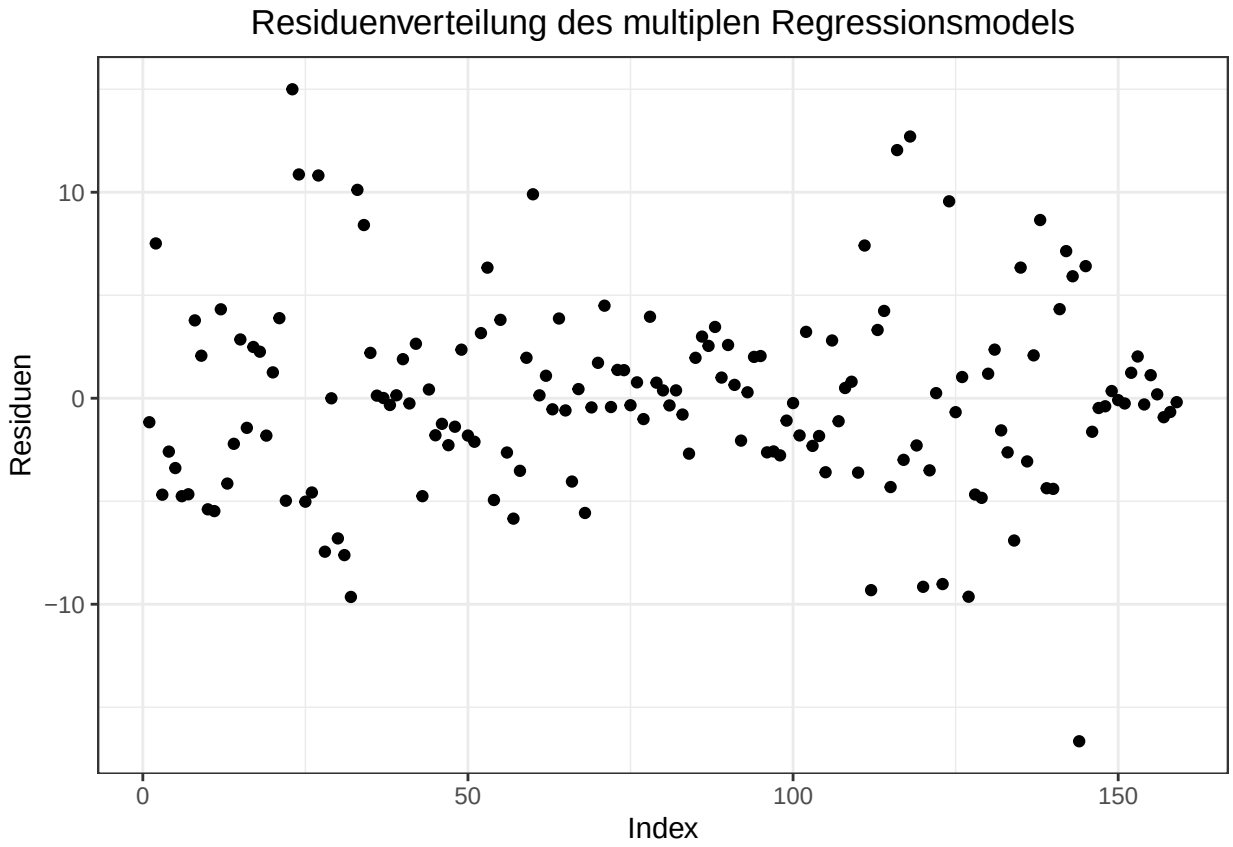
```
##
## Call:
## lm(formula = Price ~ Length3 + Species, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.929  -2.771   0.056   2.322  14.600
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -4.51025    2.20202  -2.048  0.04227 *
## Length3           0.30858    0.05345   5.773 4.29e-08 ***
## SpeciesParkki     3.28923    1.84185   1.786  0.07613 .
## SpeciesPerch      1.28579    1.12682   1.141  0.25564
## SpeciesPike       1.10045    1.51064   0.728  0.46746
## SpeciesRoach      2.60699    1.51248   1.724  0.08682 .
## SpeciesSmelt      4.51054    2.02278   2.230  0.02723 *
## SpeciesWhitefish  6.92746    2.11165   3.281  0.00129 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.754 on 151 degrees of freedom
## Multiple R-squared:  0.3153, Adjusted R-squared:  0.2836
## F-statistic: 9.933 on 7 and 151 DF,  p-value: 3.6e-10
```

Ein umfangreicheres Modell mit mehr Variablen als Gewicht und Spezie führte zu keinem höherem adjustiertem R^2. Dies lässt sich unter Anderem durch die hohe Korrelation der einzelnen metrischen Prädiktoren untereinander begründen. Somit handelt es sich bei dem spezifiziertem Modell um das finale Modell.

### Überprüfung der Annahmen

```
# Überprüfung ob Fehler normalverteilt sind
res_df <- data.frame(Index = 1:nrow(data), Residuals = model$residuals)
```

```r
# Veranschaulichung der Residuen mittels eines Scatterplots
ggplot(res_df) + aes(x = Index, y = Residuals) +
  # Scatterplot
  geom_point() +
  # Titel
  ggtitle("Residuenverteilung des multiplen Regressionsmodels") +
  # Ändere Hintergrundlayer
  theme_bw() +
  # Zentriere Titel und entferne Legende
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none") +
  # Anpassung des x Labels
  xlab("Index") +
  # Anpassung des y Labels
  ylab("Residuen")
```




```r
# Shapiro-Wilk Test auf Normalverteilung der Residuen
shapiro.test(model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.97113, p-value = 0.002058
```

**Überprüfung der Annahmen:**

Die Nullhypothese des Shapiro-Wilk Tests unterstellt die Normalverteilung, die Alternativhypothese das Gegenteil. Für ein angenommenes Signifikanzniveau von 0.05 muss die Nullhypothese auf Normalverteilung der Residuen verworfen werden, sodass die Annahmen des Modells verletzt ist. Schaut man sich die Verteilung der Residuen an, so scheint es keine Anzeichen auf Heteroskedastiziztät in den Residuen zu geben.

**4b) Modelliterpretation:**

Modellspezifikation: Das vorliegend geschätzte multiple Regressionsmodell setzt sich aus der drei Variablen zusammen. Dabei handelt es sich um die zu erklärende Variable "Preis" und die beiden Prädiktoren "Spezie" und "Gewicht". Bei der Variable "Spezie" handelt es sich um eine kategoriale Variable, sodass die Interpretation der jeweiligen Koeffizienten im Hinblick auf die ausgelassene Referenzkategorie "Bream" zu interpretieren sind.

F-Statistik und Signifikanz des Modells: Um die Modellsignifikanz zu überprüfen wird der p-Wert des F-Tests herangezogen. Die Nullhypothese unterstellt, dass alle wahren Steigungskoeffizienten gleich null sind. der p-Wert der F-Statistik zeigt einen p-Wert kleiner als 0.05 bzw. 0.01, sodass die Nullhypothese verworfen werden kann.

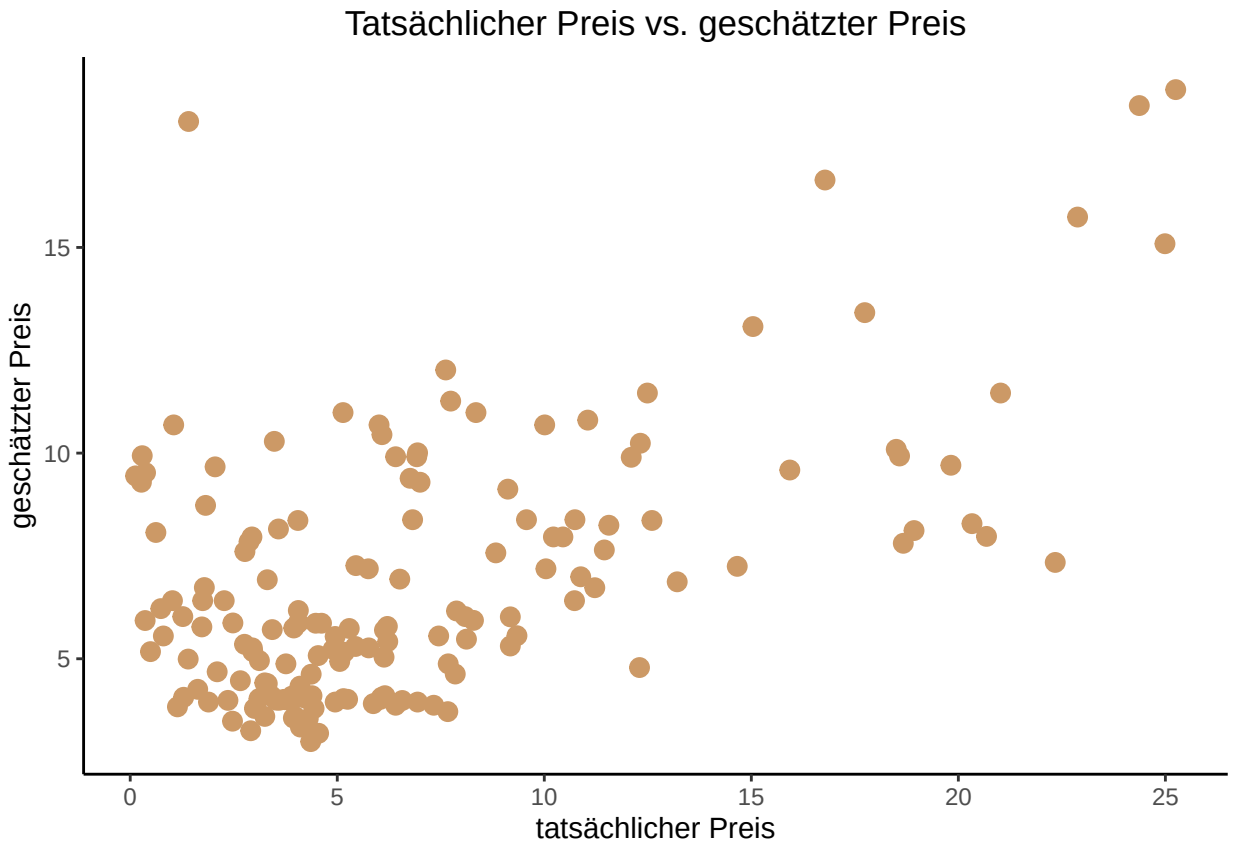
Interpretation der Koeffizienten und Signifikanz: Der Koeffizient der Variable Gewicht sagt aus, dass bei der Erhöhung des Gewichts um eine Einheit, eine Steigerung des Preises um 0.008 Einheiten zu erwarten ist, wenn die Variable "Spezie" unberücksichtigt bleibt. Der t-Test zeigt, dass der Koeffizient statistisch signifikant von null verschieden ist. Bei der Spezie sind die Koeffizienten in Bezug auf die Referenzspezie "Bream" zu interpretieren. Somit lässt sich sagen, dass die Sorte "Whitefish" einen um 6.35 höheren Preis als die Spezie "Bream" hat. Folgt man dem p-Wert der individuellen Teststatistiken so, sind unterscheiden sich ausschließlich die Spezien "Pike" und "Whitefish" signifkant von der Referenzkateogrie. Für die anderen Sorten ist kein statistisch signifikanter Unterschied zu verzeichnen.

Adj R^2 als Gütekriterium: Das adjustierte R^2 gibt wieviel von der Variation in der Variable Preis durch das Modell erklärt wird. Ein Wert von eins würde besagen, dass das Modell die Variation komplett erklärt und ein Wert von null, dass das Modell gar keinen Erklärungsgehalt besitzt. Vorliegend liegt der Wert bei 0.286, sodass etwa 28 Prozent der Variation des Preises durch das Modell erklärt wird.

Die folgende Grafik veranschaulicht den tatsächlichen Preis und den vorhergesagten Preis anhand eines Scatterplots.

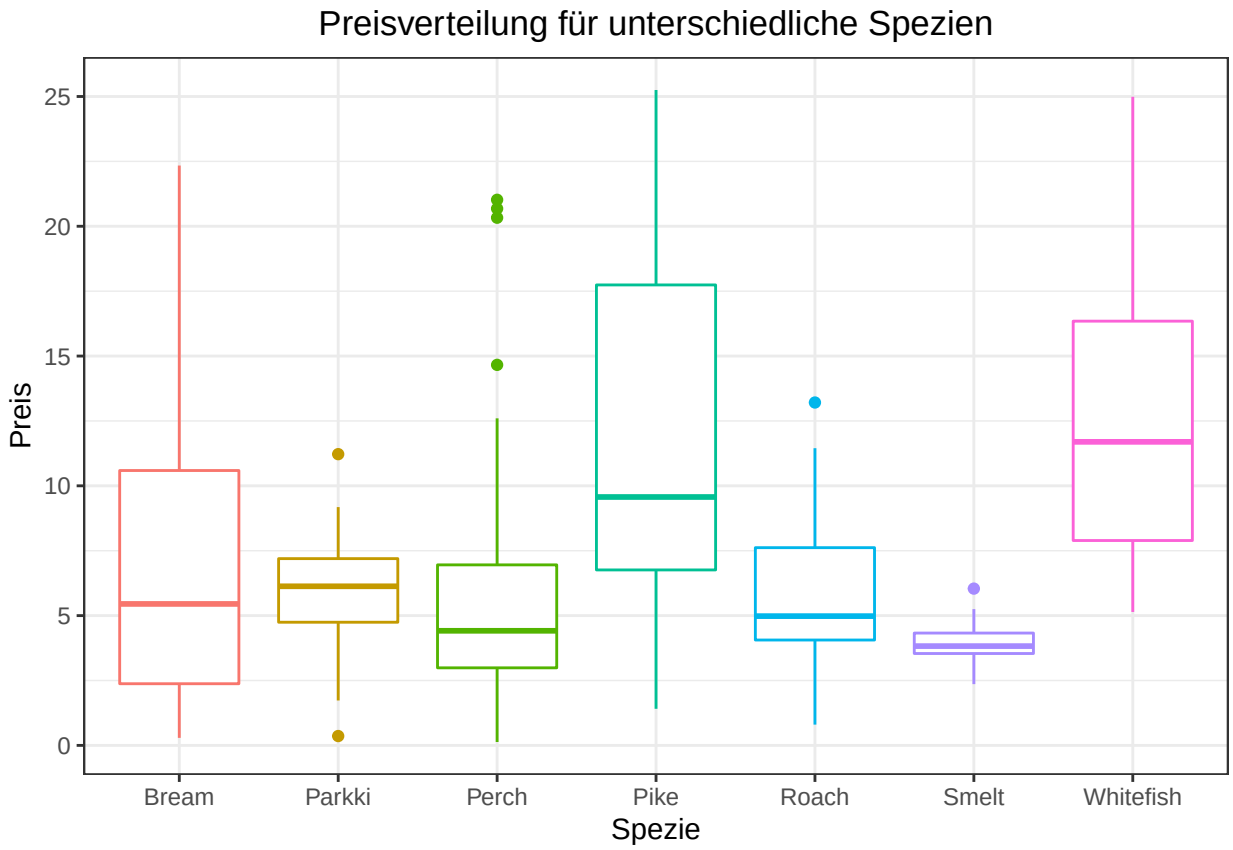
```
# Veranschaulichung der tatsächlichen und vorhergesagten Werte

# Schätzwerte für Preis
price_hat <- predict(model, newdata = data)


# Erstellung eines Dataframes der die tatsächlichen und vorhergesagten Preise beinhaltet
act_pred_df <- data.frame(price = data$Price, price_hat = price_hat)

# Veranschaulichung anhand eines Scatterplots
ggplot(act_pred_df) + aes(x = price, y = price_hat) +
  # Scatterplot
  geom_point(color = "#CC9966", size = 3) +
  # Ändere Hintergrundlayer
  theme_classic() +
  # Füge Titel hinzu
  ggtitle("Tatsächlicher Preis vs. geschätzter Preis") +
  # Zentriere Titel
  theme(plot.title = element_text(hjust = 0.5)) +
```

```
  # Anpassung des x Labels
  xlab("tatsächlicher Preis") +
  # Anpassung des y Labels
  ylab("geschätzter Preis")
```




**Aufgabe 4c)**

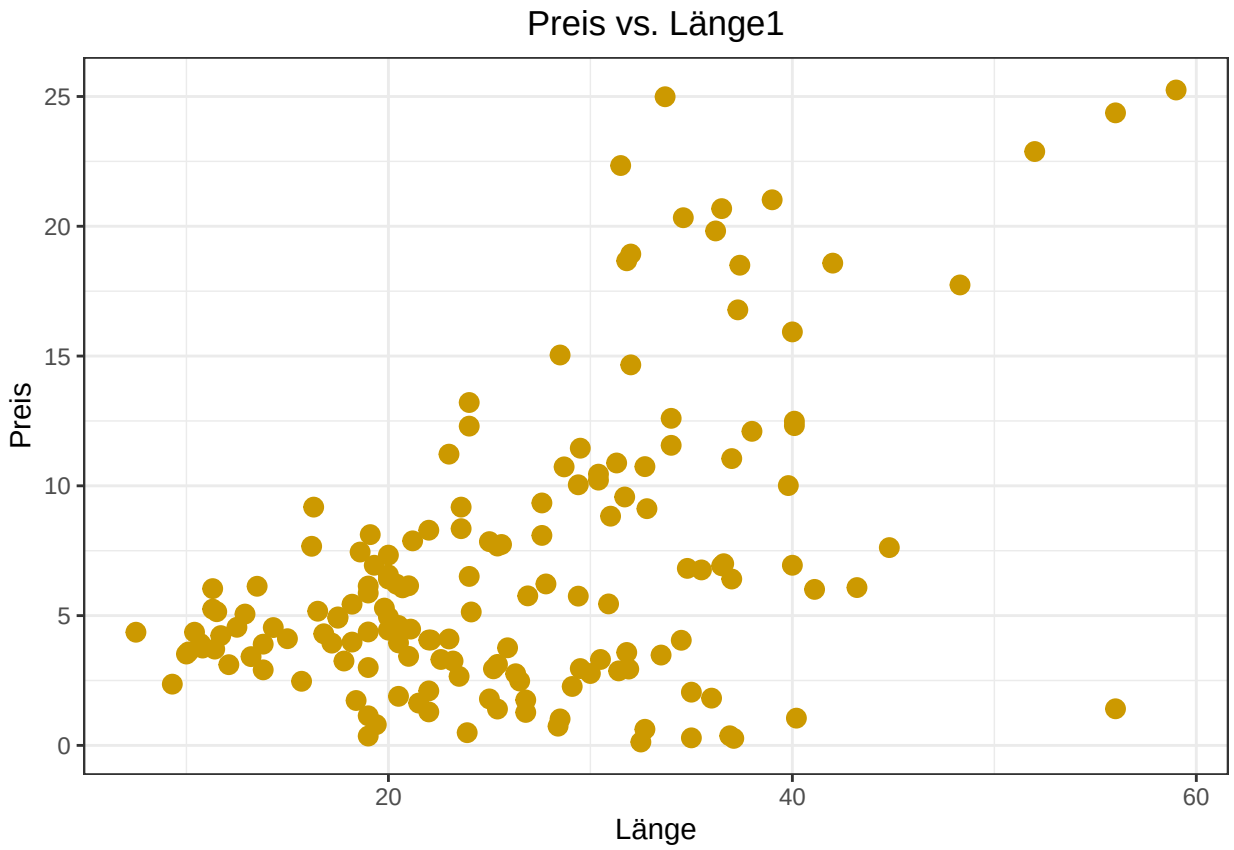
```
ggplot(data) + aes(x = Species, y = Price, colour = Species) +
  # Scatterplot
  geom_boxplot() +
  # Titel
  ggtitle("Preisverteilung für unterschiedliche Spezien") +
  # Ändere Hintergrundlayer
  theme_bw() +
  # Zentriere Titel und entferne Legende
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none") +
  # Anpassung des x Labels
  xlab("Spezie") +
  # Anpassung des y Labels
  ylab("Preis")
```




**Interpretation:**

Dem Boxplot lässt sich die Preisverteilung für die unterschiedlichen Fischarten entnehmen. Der Interquartilsabstand für die Sorten "Pike", "Bream" und "Whitefish" ist basierend auf der Datenlage am größten. Das heißt der Variation der Preise innerhalb des 0.25-Quartils und 0.75-Quartils ist hier am größten. Den geringsten Interquartilsabstand lässt sich für die Spezie "Smelt" beoabchten. Der Median ist für den Whitefish am größten und für den Smelt am gerinsten.

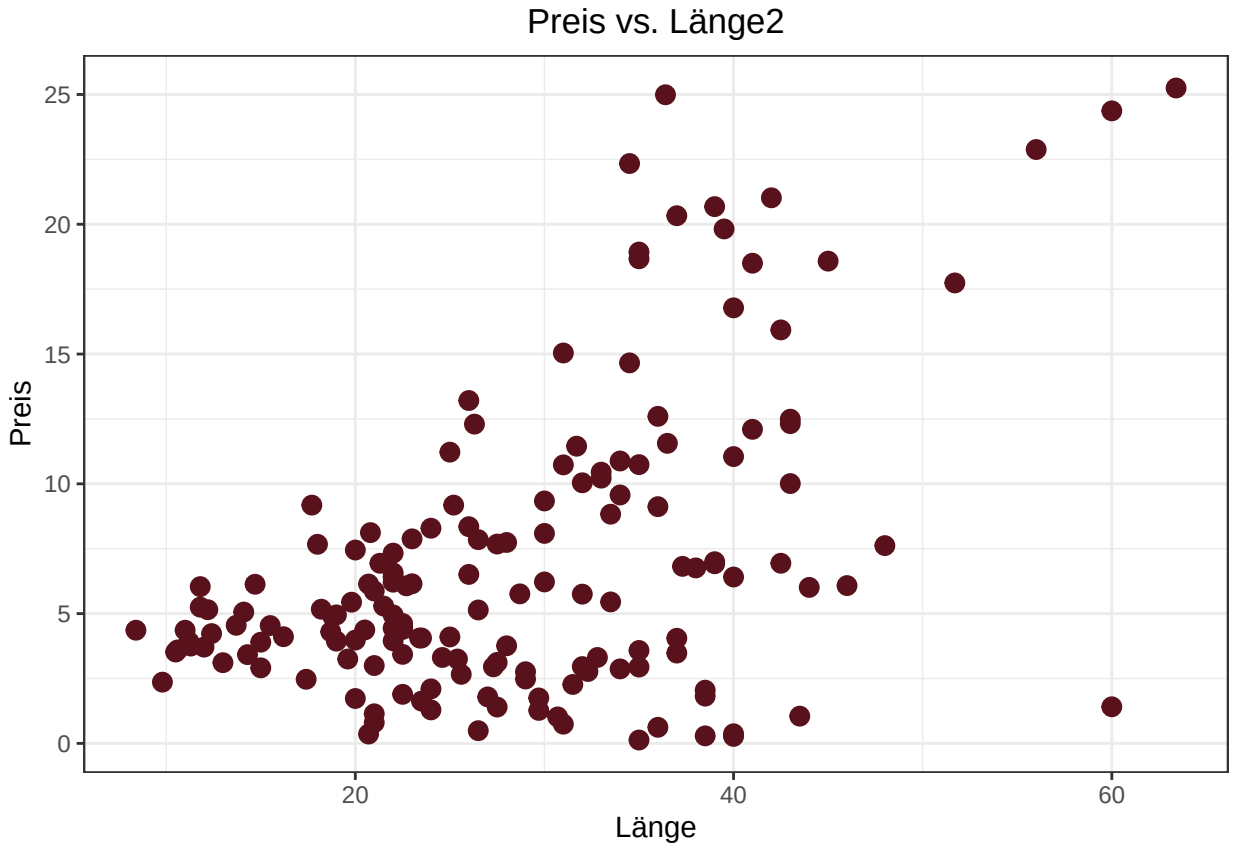
```
ggplot(data) + aes(x = Length1, y = Price) +
  # Scatterplot
  geom_point(color = "#CC9900", size = 3) +
  # Titel
  ggtitle("Preis vs. Länge1") +
   # Ändere Hintergrundlayer
  theme_bw() +
  # Zentriere Titel
  theme(plot.title = element_text(hjust = 0.5)) +
  # Anpassung des x Labels
  xlab("Länge") +
  # Anpassung des y Labels
  ylab("Preis")
```




**Interpretation:**

In der Grafik ist ein Scatterplot abgebildet, mit der Länge der Fische auf der x-Achse und dem Preis auf der y-Achse. Jeder Punkt setzt sich also aus den zwei Komponenten Preis und Länge zusammen. Die Länge erstreckt sich auf ein Intervall von (0, 60) und der Preis auf ein Intervall von (0, 25). Die maximale Länge eines Fisches liegt bei fast 60 Einheiten, wobei der dazugehörige Preis mit ca. 25 Einheiten auch den Maximalpreis darstellt. Der kürzeste Fisch hat einen Preis von fast fünf Einheiten. Weiter folgen zahlreiche längere Fische, die günstiger sind als der kürzeste Fisch, sodass man nicht pauschal sagen: Je länger der Fisch desto höher ist der Preis beziehungsweise umgekehrt je höher der Preis desto länger der Fisch. Dennoch scheint ein tendenziell positiver Zusammenhang zwischen den beiden Variablen zu bestehen. Einen Extrempunkt stellt der Fisch dar, der fast 60 Einheiten lang ist und der Preis knapp eine Einheit beträgt. Außer diesem Fisch gibt es keinen der so lang und gleichzeitig so "erschwinglich" ist.

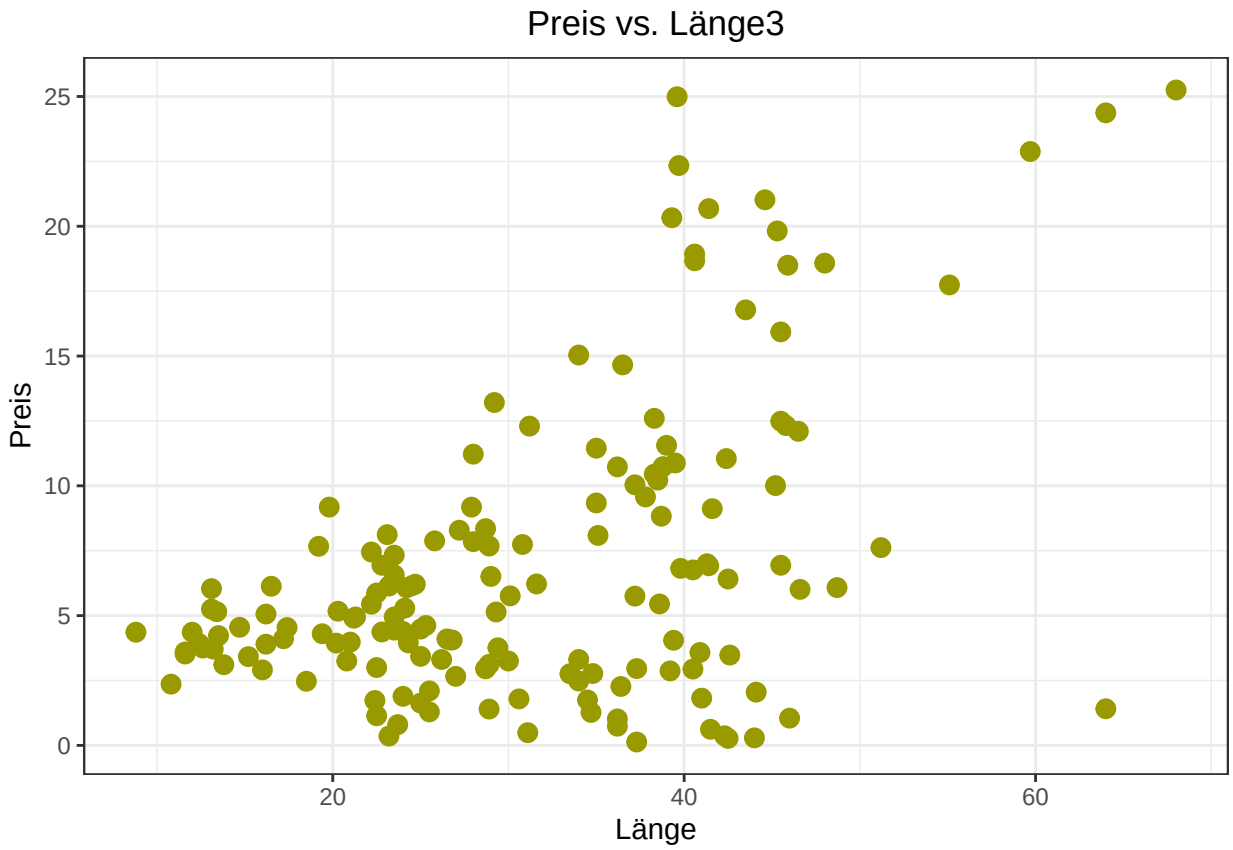
```
ggplot(data) + aes(x = Length2, y = Price) +
  # Scatterplot
  geom_point(colour = "#59121C", size = 3) +
  # Titel
  ggtitle("Preis vs. Länge2") +
   # Ändere Hintergrundlayer
  theme_bw() +
  # Zentriere Titel
  theme(plot.title = element_text(hjust = 0.5)) +
  # Anpassung des x Labels
  xlab("Länge") +
  # Anpassung des y Labels
  ylab("Preis")
```




**Interpretation:**

Ähnlich wie in der vorangegangen Grafik veranschaulicht diese Abbildung auch das Verhältnis von Länge und Preis. Dabei fällt auf, dass die maximale Länge des Längemaßes "Länge 2" jene Länge aus der vorangegangenen Abbildung um ein Paar Einheiten überteigt. Ansonsten scheint die Punktewolke der aus der vorangegangenen Grafik sehr ähnlich und die leicht positive Tendenz bleibt bestehen. Möglichweise wurde für diese Länge die Schwanzflosse mit berücksichtigt.

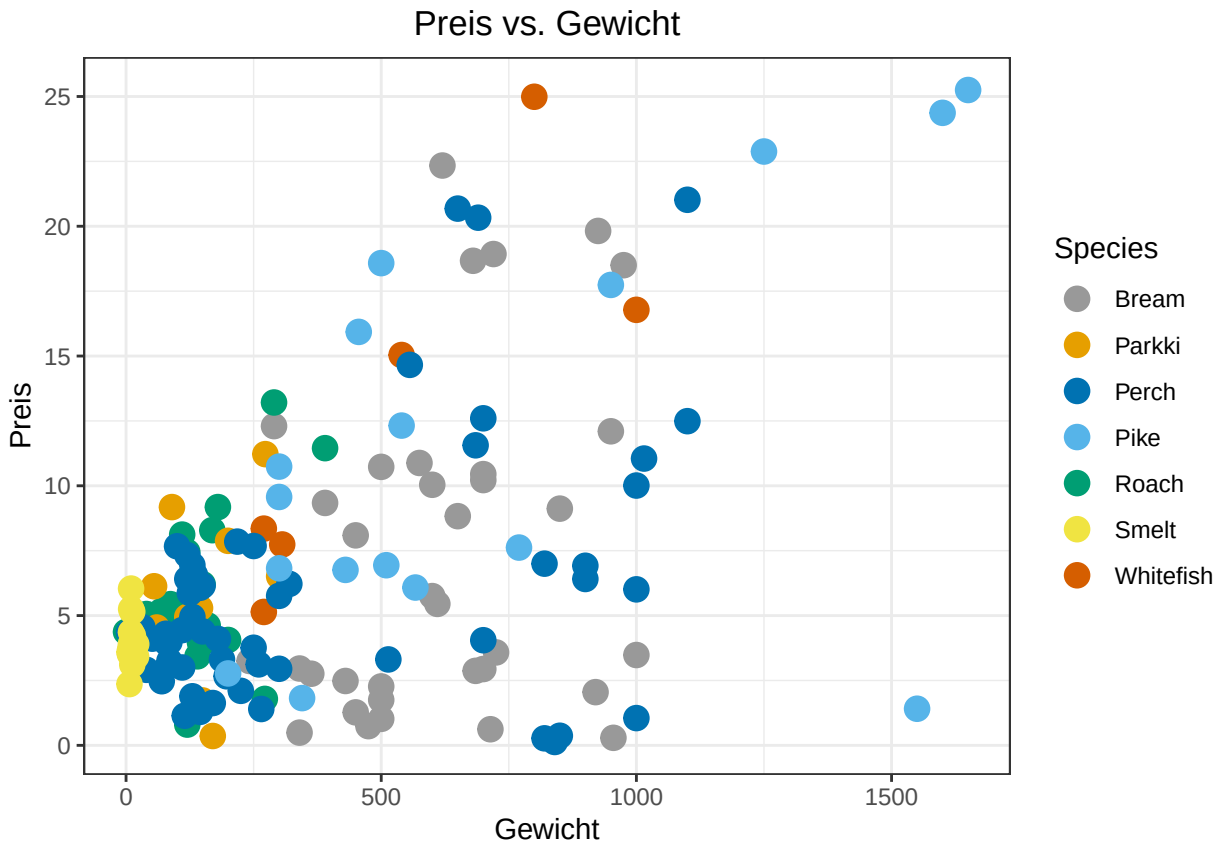
```
ggplot(data) + aes(x = Length3, y = Price) +
  # Scatterplot
  geom_point(colour = "#999900", size = 3) +
  # Titel
  ggtitle("Preis vs. Länge3") +
   # Ändere Hintergrundlayer
  theme_bw() +
  # Zentriere Titel
  theme(plot.title = element_text(hjust = 0.5)) +
  # Anpassung des x Labels
  xlab("Länge") +
  # Anpassung des y Labels
  ylab("Preis")
```




**Interpretation:**

Länge3 ist von allen Längenmaßen jene, die die maximalste Länge aufweist. Dabei ist der Unterschied zur vorherigen Abbildung mit "Länge2" nur sehr geringfügig. Der leicht positive Zusammenhang bleibt auch hier bestehen.

```
# Definiere Farbvektor
color_pal <- c("#999999", "#E69F00",  "#0072B2", "#56B4E9", "#009E73",
          "#F0E442", "#D55E00")


ggplot(data) + aes(x = Weight, y = Price, colour = Species) +
  # Scatterplot
  geom_point(size = 4) +
  # Titel
  ggtitle("Preis vs. Gewicht") +
  # Ändere Hintergrundlayer
  theme_bw() +
  # Zentriere Titel
  theme(plot.title = element_text(hjust = 0.5)) +
  # Anpassung des x Labels
  xlab("Gewicht") +
  # Anpassung des y Labels
  ylab("Preis") +
  # Eigene Farbauswahl
  scale_color_manual(values = color_pal)
```

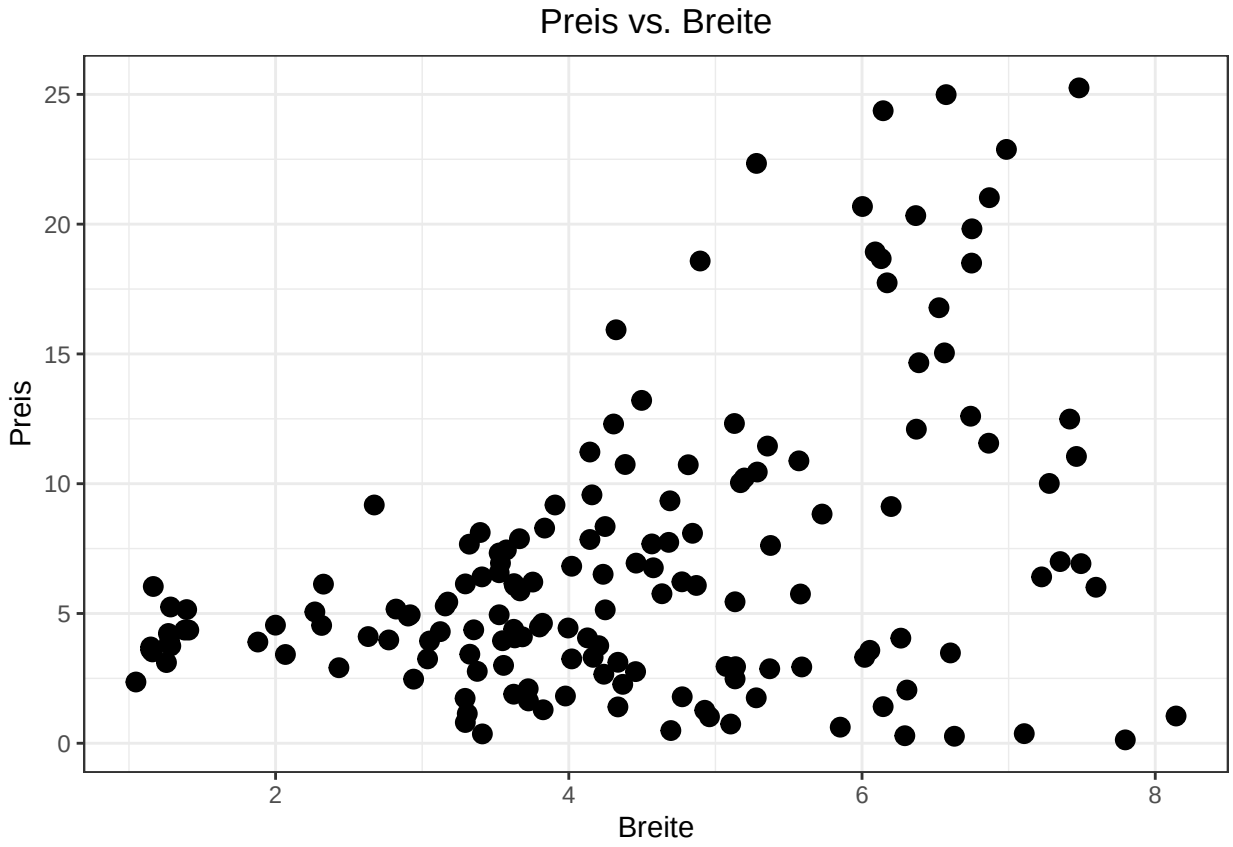



**Interpretation:**

Die vorliegende Grafik veranschaulicht anhand eines Scatterplots das Verhältnis der Variablen Preis und Gewicht eines Fisches. Die Variable Gewicht erstreckt sich von null bis einem Wert von etwa 1600. Insgesamt erweckt die Grafik den Eindruck, dass der Preis und das Gewicht positiv korrelieren. Dabei gibt es sowohl nach oben, als auch nach unten Abweichungen: Einige Fische mit höherem Gewicht sind vergleichsweise günstig, andere relativ leichte Fische sind relativ teuer. Um noch ein bisschen Informationen zu erhalten wurden die Punkte in Abhängigkeit von der Spezie farbig eingebfärbt. Die Zuordnung lässt aus der Legende ableiten. Dabei fällt auf das die Sorten "Smelt" und "Parkki" und "Roach" ein im Vergleich zu den anderen Sorten ziemlich niedriges Gewicht aufweisen. Weiterhin scheinen die Punkte für diese beiden Gruppen ein wenig zentrierter zu. Der Whitefish weist mit sechs Beobachtungspunkten eine vergleichweise größere Streuung auf. Die Sorten "Bream", "Perch" und "Pike" zeigen eine große Variation hinisichtlich beider Variablen auf. Die drei schwersten Fische gehören der Sorte "Pike" an. Interessant ist der Datenpunkt unten rechts in der Ecke. Dieser Fisch der Sorte "Pike" weist für den Datensatz ein ziemlich hohes Gewicht auf, hat jedoch ein ziemlich niedrigen Preis. Das ist insofern überraschend, als das die beiden Datenpunkte oben rechts in der Grafik der gleichen Spezie angehören, etwa gleichviel wiegen, jedoch deutlich teurer sind. Um diesen Punkt besser zu verstehen, wäre es erforderlich weitere Variablen heranzuziehen.

```
ggplot(data) + aes(x = Width, y = Price) +
  # Scatterplot
  geom_point(size = 3) +
  # Titel
  ggtitle("Preis vs. Breite") +
   # Ändere Hintergrundlayer
  theme_bw() +
  # Zentriere Titel
```

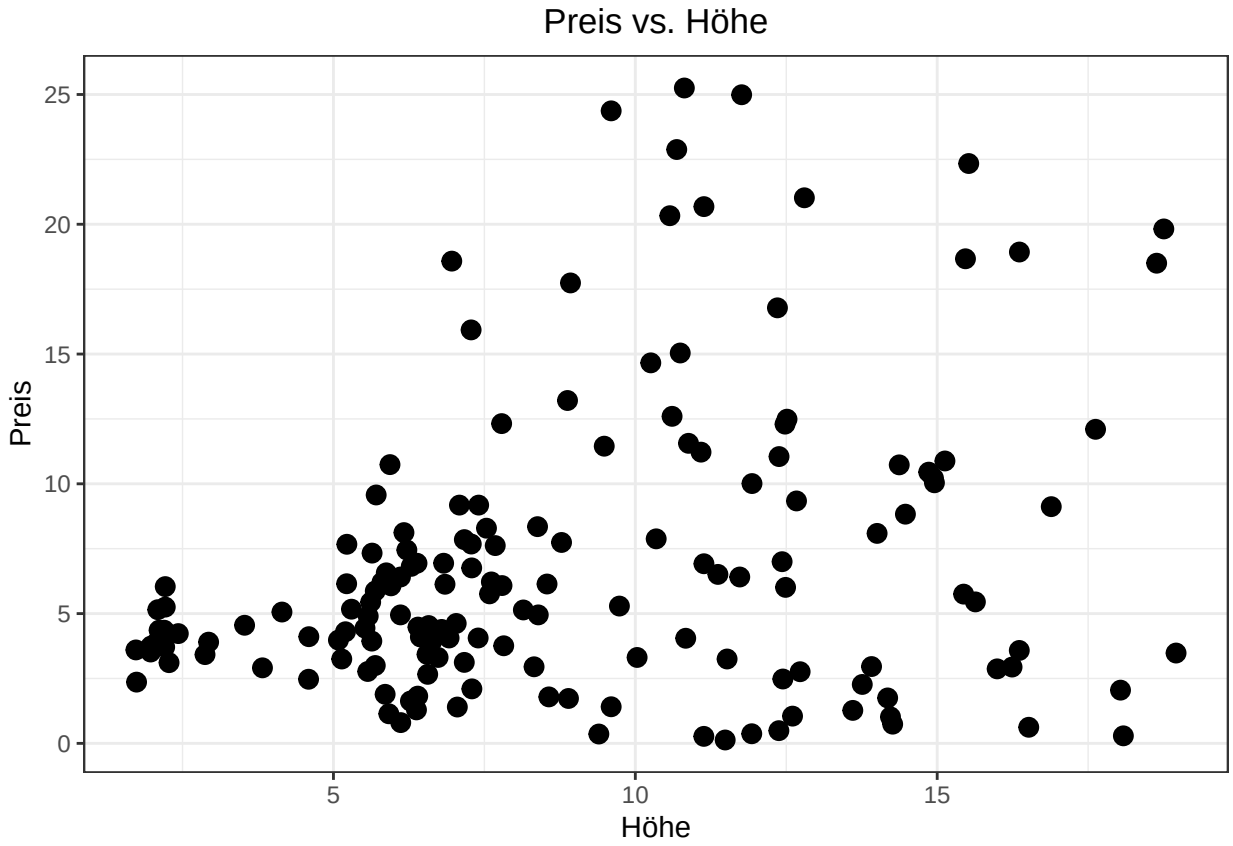
```
  theme(plot.title = element_text(hjust = 0.5)) +
  # Anpassung des x Labels
  xlab("Breite") +
  # Anpassung des y Labels
  ylab("Preis")
```




**Interpretation:**

Zur Gegenüberstellung der Variablen Gewicht und Preis wurde ebenfalls ein Scatterplot gewählt. Die Breite variiert in einem Intervall von null bis etwa acht acht Einheiten. Dabei scheint auch hier ein leicht positiver Zusammenhang der beiden Variablen vorhanden zu sein. Jedoch gibt es auch hier eine Gruppe von Datenpunkten, die eine hohe Breite aufweisen und gleichzeitig einen ziemlich niedrigen Preis haben.

```
ggplot(data) + aes(x = Height, y = Price) +
  # Scatterplot
  geom_point(size = 3) +
  # Titel
  ggtitle("Preis vs. Höhe") +
   # Ändere Hintergrundlayer
  theme_bw() +
  # Zentriere Titel
  theme(plot.title = element_text(hjust = 0.5)) +
  # Anpassung des x Labels
  xlab("Höhe") +
  # Anpassung des y Labels
  ylab("Preis")
```




**Interpretation:**

Wie zuvor wurde auch beim Vergleich der Höhe und des Preises eines Fisches ein Scatterplot zur Veranschaulichung gewählt. die Höhe erstreckt sich von null bis etwa 20. Im Vergleich zu den vorherigen Grafiken scheint nur noch ein sehr leichter positiver Zusammenhang zwischen den beiden Variablen vorhanden zu sein. Dieser begründet sich hauptsächlich durch den niedrigen Preis für Fische mit einer geringen Höhe. Ab einer von etwa 7 Einheiten scheint die Variation des Preises eher zufällig, sodass kein richtiges Muster mehr zu erkennen ist.

**4d)**

```r
# Schätze das vorgegebene Modell
mod_weight <- lm(Weight ~ Length3 + Width, data = data)

# Gebe die Modellzusammenfassung aus
summary(mod_weight)
```

```
##
## Call:
## lm(formula = Weight ~ Length3 + Width, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -258.84  -77.17  -25.78   81.79  441.57
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -520.939     29.332 -17.760  < 2e-16 ***
## Length3       19.487      1.812  10.755  < 2e-16 ***
## Width         70.343     12.479   5.637 7.91e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 126.3 on 156 degrees of freedom
## Multiple R-squared:  0.8771, Adjusted R-squared:  0.8755
## F-statistic: 556.4 on 2 and 156 DF,  p-value: < 2.2e-16
```

```r
# Erstelle einen Dataframe, der die Daten beinhaltet
new_data <- data.frame(Length3 = 30, Width = 4)

# Vorhersage basierend auf den vorgegebenen Angaben
predict(mod_weight, newdata = new_data)
```

```
##        1
## 345.0477
```

**Modellinterpretation:**

Das multiple Regressionsmodell erklärt die Variable Gewicht anhand von zwei Prädiktoren. Dabei handelt es sich um die Länge (genauer gesagt Länge3) und die Breite des Fisches. Die drei zu schätzenden Koeffizienten sind alle hochsignifikant. Basierend auf dem p-Wert der F-Statistik ist auch das Modell als Ganzes signifikant, sodass die Nullhypothese (alle Koeffizienten sind null) verworfen werden kann. Als Bestimmtheitsmaß gibt das adjustierte R^2 einen Wert von 0.8755 aus. Das bedeutet, dass aufgerundet 88 Prozent der Variation der Variable Gewicht durch das geschätzte Modell erklärt werden kann. Der partielle Steigungsparameter der Variable "Length3" gibt die erwartete Veränderung des Gewichts an, wobei die Variable "Width" konstant gehalten wird: Erhöht sich die Länge um eine Einheit, so erhöht sich das Gewicht um 19.487 Einheiten. Umgekehrt gilt dasselbe für die Breite: Erhöht sich die Breite um eine Einheit, so eröht sich das Gewicht um 70.343 Einheiten.

Der vorhergesagte Gewichtsschätzwert für eine Länge von 30 Einheiten und einer Breite von 4 Einheiten liegt bei bei 345.0477 Einheiten.

# 5.Aufgabe

### 5a)

Unterteilung des Datensatzes in drei Gewichtsklassen:

```r
# Kreiere eine neue Spalte "Weight_cat"
data[, "Weight_cat"] <- NA

# Wenn das Gewicht kleiner 120 weise label "Leicht" zu
data[, "Weight_cat"] <- ifelse(data[, "Weight"] < 120, "Leicht", data[, "Weight_cat"])

# Wenn das Gewicht größer gleich 120 und kleiner gleich 650 weise label "Mittel" zu
data[, "Weight_cat"] <- ifelse(data[, "Weight"] >= 120 & data[, "Weight"] <= 650, "Mittel",
                               data[, "Weight_cat"])
```

```r
# Wenn das Gewicht größer als 650 weise label "Schwer" zu
data[, "Weight_cat"] <- ifelse(data[, "Weight"] > 650, "Schwer", data[, "Weight_cat"])

# Konvertiere neue Variable in eine kategoriale Variable
data$Weight_cat <- factor(data$Weight_cat)

# Überprüfung der Datenstruktur
str(data$Weight_cat)
```

```
##  Factor w/ 3 levels "Leicht","Mittel",..: 2 2 2 2 2 2 2 2 2 2 ...
```

**5b)**

Erstellung einer Tabelle, die die Häufigkeiten der Fischssorten und Gewichtsklassen darstellt.

```r
# Kreiere eine Tabelle, die für jede Fischsorte die Anzahl der Gwichtsklassen wiedergibt
freq_table <- table(data$Species, data$Weight_cat)

# Wiedergabe der Tabelle
print(freq_table)
```

```
##
##             Leicht Mittel Schwer
##   Bream          0     20     15
##   Parkki         3      8      0
##   Perch         13     27     16
##   Pike           0     11      6
##   Roach          6     14      0
##   Smelt         14      0      0
##   Whitefish      0      4      2
```

```r
sum(freq_table[, 1]) # Anzahl der leichten Fische im Datensatz
```

```
## [1] 36
```

```r
sum(freq_table[, 2]) # Anzahl der mittleren Fische im Datensatz
```

```
## [1] 84
```

```r
sum(freq_table[, 3]) # Anzahl der schweren Fische im Datensatz
```

```
## [1] 39
```

Die Zeilen in der Tabelle addieren sich zu zur Anzahl der Fischsorte im Fischdatensatz. Die Spalten addieren sich zur Anzahl der Häufikeiten in den jeweiligen Gewichtsklassen. Insgesamt gibt es 36 leichte Fische, 84 Fische die in die Kategorie “Mittel” gehören und 39 Beobachtungen, die als schwer klassifiziert werden.

Für die Durchführung des Chi-Quadrat Unabhängigkeitstests ist es in der Regel erforderlich, dass die einzelnnen Zellhäufigkeiten nicht kleiner als 5 sind. Diese Voraussetzung wird vorliegend vereltzt.

**5c)**

Gibt es einen Zusammenhang zwischen der Fischspzies und dem Gewicht? Mit anderen Worten: Unterscheiden sich die Fischsorten statistisch signifikant hinischtlich des Gewichts?

Nullhypothese: Die beiden Variablen sind voneinander unabhängig.

Alternativhypothese: Beide Variablen sind voneinander abhängig.

```
# Filtere relevante Zeilen
data3x3 <- data %>% filter(Species == "Bream"| Species == "Perch" | Species == "Roach") %>%
  droplevels()

# Kreiere eine Tabelle
table3x3 <- table(data3x3$Species, data3x3$Weight_cat)

# Wiedergabe der Tabelle
print(table3x3)
```

```
##
##         Leicht Mittel Schwer
##   Bream      0     20     15
##   Perch     13     27     16
##   Roach      6     14      0
```

```
# chi-square test
chisq.test(table3x3)
```

```
## Warning in chisq.test(table3x3): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table3x3
## X-squared = 18.851, df = 4, p-value = 0.0008408
```

**Interpretation:**

Der Chi-Square - Test ist signifikant, zu einem Signifikanzniveau von 0.05, als auch zu einem niedrigeren Signifikanzniveau. Die Nullhypothese kann somit verworfen werden. Für die ausgewählten Gruppen lässt sich sagen, dass ein statistisch signifikanter Zusammenhang zwischen der Spezies und dem Gewicht besteht. Das Ergebnis des Tests ist mit Vorsicht zu genießen, da die Annahme von mindestens fünf Beobachtungen pro Zelle nicht erfüllt ist. Aus diesem Grund wird die Meldung "Chi-squared approximation may be incorrect" ausgewiesen.

# 6.Aufgabe

**Kovarianz vs. Korrelation**

Die Kovarianz und die Korrelation sind beides Maße, um den linearen Zusammenhang zweier metrischer Variablen *x,y* zu messen bzw. zu erfassen. Sowohl die Kovarianz als auch der Korrelationskoeffizient sind symmetrisch, d.h. $cov(x, y) = cov(y, x)$ bzw. $r(x, y) = r(y, x)$. Sind $x$ und $y$ voneinader stochastisch unabhängig, so folgt, dass sowohl die Kovarianz, als auch die Korrelation zwischen den Merkmalen null ist. Die Kovarianz ist abhängig von den Maßeinheiten der zu untersuchenden Variablen. Sie kann sich somit, je nachdem welche Variablen untersucht werden, stark voneinander unterscheiden, was eine Vergleichbarkeit verschiedener Kovarianzen schwierig macht. Um dieses Problem zu umgehen, bietet es sich an den Korrelationskoeffizienten zu verwenden, der gegenüber Maßunterschieden beständig ist. Damit ist gemeint, dass der Korrelationskoeffizient, unabhängig von Maßeinheiten der zu untersuchenden Variablen, stets im Intervall (-1, 1) liegt. Dabei spieget der Wert "-1" einen komplett negativen Zusammenhang wieder und der Wert "1" einen perfekt positiven linearen Zusammenhang. Ein Wert von 0 drückt aus, dass zwei metrisch erfasste Variablen keinen linearen Zusammenhang aufweisen. Ein wesentlicher Vorteil des Korrelationskoeffizienten

gegenüber der Kovarianz besteht also darin, dass sich der Korrelationskoeffizient mehrerer unterschiedlicher Variablenkonstellationen vergleichen lässt (vgl. die Korrelationstabelle aus Aufgabe 4). Prinzipiell ist es möglich eine Korrelation auch kausal zu interpretieren: $x$ beinflusst $y$ kausal bzw. umgekehrt. Außerdem lässt sich auch sagen, dass sich beiden Variablen $x$, $y$ wechselseitig beinflussen. Für eine entsprechende Interpretation sind jedoch meistens relevante Zusatzinformationen erforderlich.

# 7.Aufgabe

Bei der Durchführung von Hypothesentests wird eine Nullhypothese aufgestellt und beispielsweise zu einem Signifikanzniveau von fünft Prozent getestet. Die zugrunde liegende Wahrscheinlichkeit die Nullhypothese fälschlicherweise abzulehnen ist somit auf maximal 5 Prozent festegelegt (Fehler 1. Art).

Beim Fehler 2. Art wird hingegen die Nullhypothese fälschlicherweise bestätigt, obwohl die Gegenhypthose korrekt ist. Zum Fehler 2. Art kommt es, wenn die Nullhypothese verworfen werden müsste, der p-Wert eines durchgeführten statistischen Tests jedoch größer als festgelegte Signifikanzniveau ausfällt. Es wird also geschlussfolgert, dass es keinen signifikanten Effekt oder Unterschied gibt, obwohl in der Realität das Gegenteil der Fall ist.
Die Prüfgröße "z_pr" fällt beim Vorliegen des Fehlers 2. Art kleiner aus, als kritische Wert "z_kr". Aus dem festgelegten Signifikanzniveau lässt sich der kritische Wert "z_kr" ableiten. Eine Verkleinerung des Fehlers 1. Art führt zu einer Erhöhung des Fehlers 2.Art und umgekehrt.

# 8.Aufgabe

Ein Konfidenzintervall gibt Auskunft über die Genauigkeit eines zu schätzenden Parameters. Häufig liegt einem zu schätzenden Parameter eine Zufallsstichprobe zugrunde. Im Folgenden soll angenommen werden, dass man an dem Populationsmittwelwert einer normalverteilten Zufallsvariable interessiert ist. Dabei soll es sich um die Variable Alter handeln. Für eine Stichprobe von 1000 zufällig ausgewählten Personen in der Innenstadt Wiens wurde ein durchschnittliches Alter von 35 Jahren erfasst. Die Standardabweichung sei mit 10 Jahren gegeben. Daraus lässt sich nun das Konfidenzintervall ableiten:

```
# Berechne Konfidenzintervall, obere und untere Grenze
KI_low <- 37 - 1.96 * (10 / sqrt(1000))
KI_up <-  37 + 1.96 * (10 / sqrt(1000))

# Gebe obere und untere Grenze aus
print(KI_low)
```

```
## [1] 36.38019
```

```
print(KI_up)
```

```
## [1] 37.61981
```

Dabei ist 1,96 der z-Wert in der Tabelle für die Standardnormalverteilung. Interpretation: Mit einer Vertrauenswahrscheinlichkeit von 95 Prozent liegt das durchschnittliche Alter der Wiener (entspricht hier der Grundgesamtheit) im Bereich zwischen 36.38 und 37.62 Jahren. Je höher die Vertrauenswahrscheinlichkeit sein soll, desto größer wird auch das Intervall. In 95% der Fälle enthält das berechnete 95%-Konfidenzintervall den wahren Wert aus der Grundgesamtheit, der mit dem Punktschätzer, in dieser Anwendung dem Mittelwert, geschätzt wird.