Erste Hausarbeit in Statistik für Wirtschaftsinformatiker

HTW Berlin, Sommersemester 2017

Name, Matrikelnummer: Jenny Rothe
Name, Matrikelnummer: Laura Laugwitz, 544049

Formalitäten

Bitte bearbeiten Sie diese Hausarbeit in Zweiergruppen, in denen beide Studierende bei Herrn Spott oder beide bei Herrn Heimann eingeschrieben sind. Gruppen von drei oder mehr Studierenden sind nicht zugelassen. Setzen Sie bitte Ihre beiden Namen und Matrikelnummern oben ein.

Öffnen Sie das Dokument `titanic.Rmd` in RStudio, lösen Sie alle Aufgaben mit R und **fügen Sie alle Antworten zu diesem R-Markdown-Dokument hinzu, einschließlich des R-Codes**, wie Sie es bereits bei den Übungsblättern getan haben. Zusätzliche handgeschriebene Lösungen oder Erklärungen sind nicht zugelassen ebensowenig wie Lösungen, die mit anderer Software wie z.B. Microsoft Excel erstellt wurden.

Mehr Informationen über R-Markdown-Dokumente finden Sie im Internet unter http://rmarkdown.rstudio.com/lesson-1.html. Sie können die Musterlösungen in RMarkdown zu unseren Übungsblättern als Beispiele heranziehen.

Wenn Sie mit der Bearbeitung fertig sind, erzeugen Sie bitte in RStudio mit dem Knopf **Knit PDF** ein PDF-Dokument oder wählen alternativ über das Dreieck neben **Knit PDF** die Option **Knit HTML**, um ein HTML-Dokument zu erzeugen. Für **Knit PDF** ist die Installation einer LaTeX-Distribution wie MikTeX für Windows (miktex.org) oder MacTeX für Mac OS X (www.tug.org/mactex/) erforderlich. **Knit HTML** funktioniert auch ohne LaTeX. **Drucken Sie das so erzeugte Dokument aus und geben Sie es in Papierform ab**.

Abgabe

- **Elektronisch in Moodle bis spätestens Montag, 22.05.2017 um 16:00**
  - sowohl das RMarkdown-Quelldokument
  - als auch das daraus erzeugte PDF- oder HTML-Dokument

- **UND in Papierform spätestens am 22.05. bzw. 23.05.** bei Herr Heimann bzw. Herrn Spott in den Übungen oder der Vorlesung. Spätere Abgaben werden nicht berücksichtigt und führen automatisch zu einer Bewertung mit 0 Punkten.

Die Ergebnisse aller Hausarbeiten werden zusammen mit 30% gewichtet, das Ergebnis der Klausur mit 70%.

Wichtig: Sobald Sie eine Hausarbeit abgeben, hat damit Ihre Prüfungsleistung für das Sommersemester 2017 begonnen, die mit der Klausur abgeschlossen wird. Wenn Sie eine Hausarbeit abgeben, aber die Klausur nicht im Sommersemester 2017 mitschreiben, sind Sie automatisch durchgefallen und die Punkte der Hausarbeiten verfallen. Diese Regelung ist in der Prüfungsordnung festgelegt.

Viel Erfolg,

Ihre Dozenten Martin Spott und Michael Heimann

Stand 12.05.2017

Aufgaben

Die Aufgaben befassen sich mit Passagierdaten des Kreuzfahrtschiffes Titanic, das 1912 auf ihrer Jungfernfahrt gesunken ist. Der Datensatz `titanic_data.csv` enthält Informationen über knapp 900 der geschätzen 2224 Personen, die zum Zeitpunkt des Untergangs an Bord waren.

Die Merkmale sind

- `PassengerId`: Passenger ID
- `Survived`: 0=no, 1=yes
- `Pclass`: Passenger Class
- `Name`
- `Sex`
- `Age`
- `SibSp`: Number of Siblings/Spouses Aboard
- `Parch`: Number of Parents/Children Aboard
- `Ticket`: Ticket Number
- `Fare`
- `Cabin`
- `Embarked`: Port of Embarkation

Aufgabe 1

a) Lesen Sie den Datensatz `titanic_data.csv` in R ein und weisen Sie ihn der Variablen `titanic_data` zu!

```
titanic_data <- read.csv("titanic_data.csv")
View(titanic_data) # Zusätzlicher Eindruck
str(titanic_data) # Zusätzlicher Eindruck
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 41
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : Factor w/ 681 levels "110152","110413",..: 525 596 662 50 473 276 86 396 345 133
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

b) Bestimmen Sie für jedes Merkmal, ob es qualitativ/quantitativ ist und auf welcher Skala es definiert ist (Nominal-, Ordinal- oder Kardinalskala)!

```
library(knitr)
Merkmale <- c("PassengerId", "Survived", "Passenger Class", "Name",
              "Sex", "Age", "# of Sibling", "# of Parents/Children",
              "Ticket #", "Fare", "Cabin", "Port of Embarkation")
QualiQuanti <- c("quantitativ", "quali?", "quali?", "qualitativ",
                 "qualitativ", "quantitativ", "quantitativ", "quantitativ",
                 "qualitativ", "quantitativ", "quali?", "qualitativ")
Skala <- c("Kardinal", "Nominal", "Ordinal?", "Nominal",
           "Nominal", "Kardinal", "Kardinal", "Kardinal",
```

```
            "Nominal?", "Kardinal", "Ordinal?", "Nominal?")

Titanic_Merkmale <- data.frame(Merkmale, QualiQuanti, Skala)
kable(Titanic_Merkmale)
```

Merkmale	QualiQuanti	Skala
PassengerId	quantitativ	Kardinal
Survived	quali?	Nominal
Passenger Class	quali?	Ordinal?
Name	qualitativ	Nominal
Sex	qualitativ	Nominal
Age	quantitativ	Kardinal
# of Sibling	quantitativ	Kardinal
# of Parents/Children	quantitativ	Kardinal
Ticket #	qualitativ	Nominal?
Fare	quantitativ	Kardinal
Cabin	quali?	Ordinal?
Port of Embarkation	qualitativ	Nominal?

c) Bestimmen Sie die Anzahl der Passagiere im Datensatz mit R-Befehlen!

```
dim(titanic_data) # Gibt die Dimensionen an
```

```
## [1] 891  12
```

```
nrow(titanic_data) # Zählt Reihen
```

```
## [1] 891
```

d) Bestimmen Sie die Anzahl der Merkmale im Datensatz mit R-Befehlen!

```
names(titanic_data) # Bezeichungen der Merkmale
```

```
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"
```

```
ncol(titanic_data) # Zählt Spalten
```

```
## [1] 12
```

e) Fügen Sie eine Spalte namens `Survived2` zum Data-Frame `titanic_data` hinzu, welche die Überlebenden mit dem String `"yes"` und alle anderen mit `"no"` kodiert!

```
titanic_data[, "Survived2"] <- 0
for(i in 1:891){
  if(titanic_data[i, 2]==1){
    titanic_data[i, 13] <- "yes"
  }
  else {
    titanic_data[i, 13] <- "no"
  }
}
View(titanic_data)
```

f) Bestimmen Sie die Anzahl fehlender Werte des Merkmals `Age` mit R-Befehlen!

```r
sum(is.na(titanic_data$Age))
```

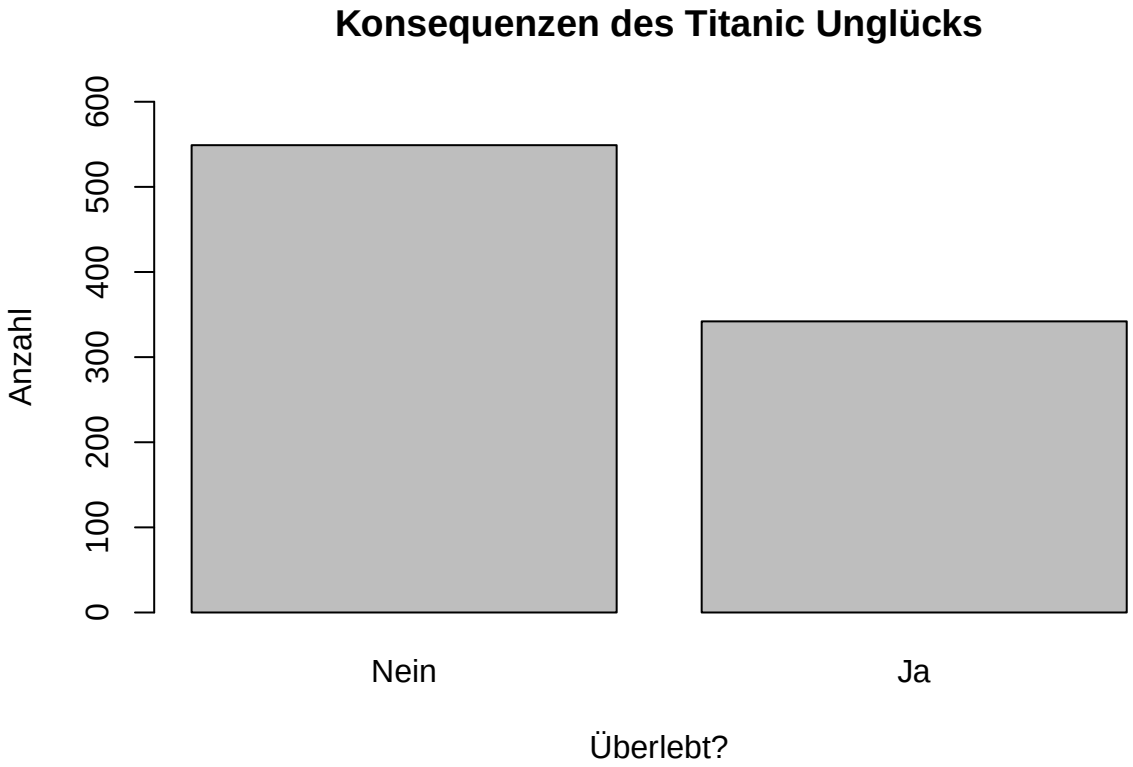
```
## [1] 177
```

Aufgabe 2

Geben Sie den Diagrammen in den folgenden Aufgaben sinnvolle Titel und beschriften Sie alle Axen!

a) Erzeugen Sie ein Balkendiagramm des Merkmals `Survived2`, welches die absolute Anzahl der Überlebenden und Gestorbenen gegenüberstellt!

```r
h <- table(titanic_data$Survived2)
barplot(h, names.arg = c("Nein", "Ja"), main="Konsequenzen des Titanic Unglücks",
        xlab="Überlebt?", ylab ="Anzahl", ylim = c(0, 600))
```

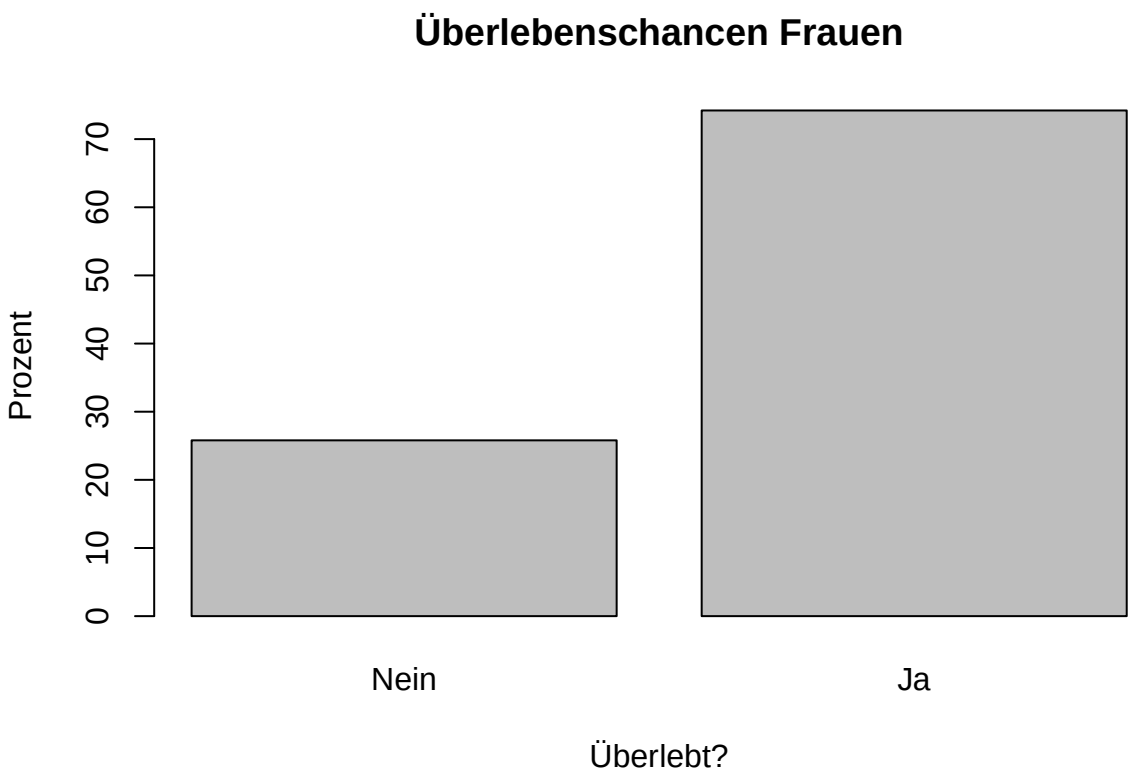


b) Erzeugen Sie weitere Balkendiagramme des Merkmals `Survived2`, diesmal mit den relativen Häufigkeiten in Prozent:

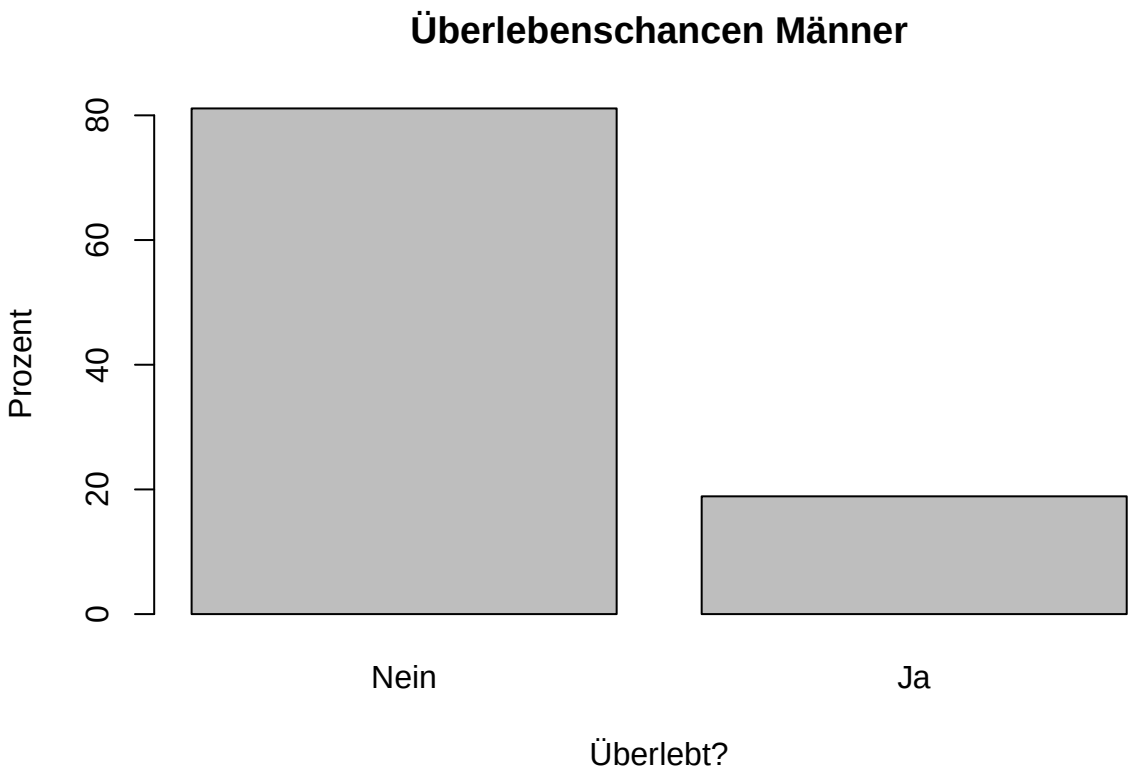
(i) für Frauen
(ii) für Männer
(iii) für jede der drei Klassen des Merkmals `Pclass`

```r
# we need the row-wise proportion, ie, the proportion of each sex that survived, as separate groups
# So we need to tell the command to give us proportions in the 1st dimension which stands for the r
# http://trevorstephens.com/kaggle-titanic-tutorial/r-part-2-the-gender-class-model/

f_gender <- prop.table(table(titanic_data$Sex, titanic_data$Survived), 1)*100
f_women <- f_gender[1,]
f_men <- f_gender[2,]
f_class <- prop.table(table(titanic_data$Pclass, titanic_data$Survived), 1)*100
f_1 <- f_class[1,]
```

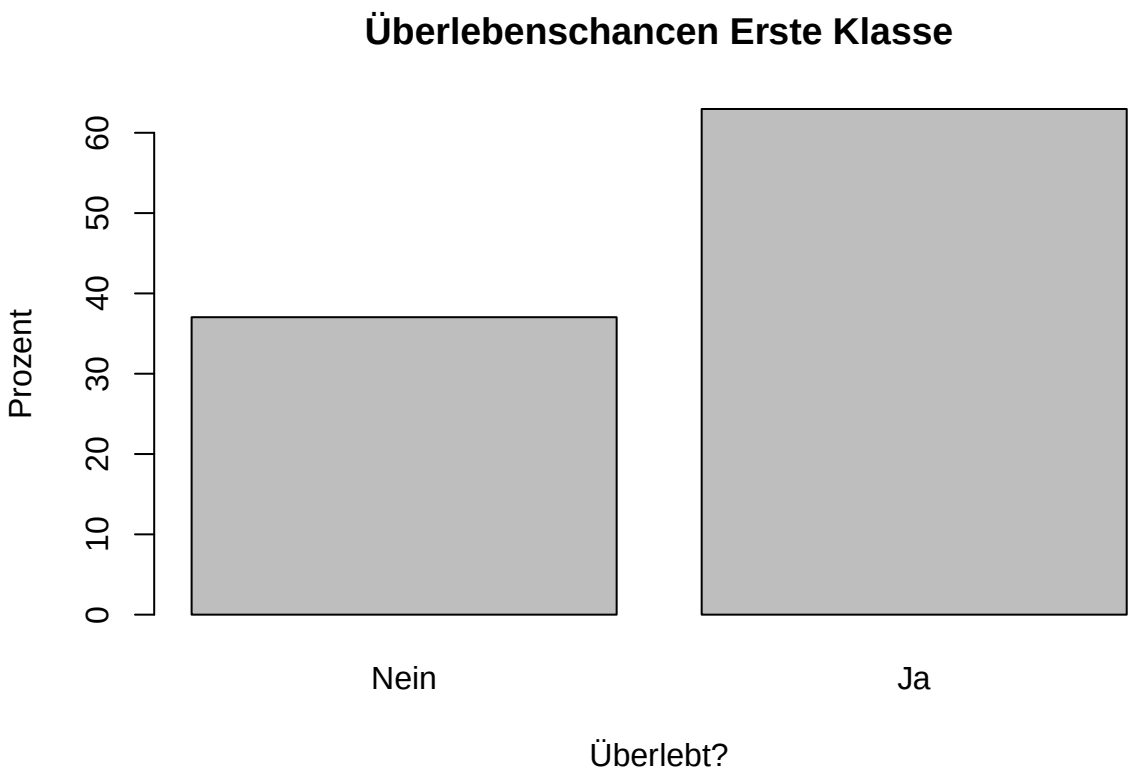
```r
f_2 <- f_class[2,]
f_3 <- f_class[3,]

barplot(f_women, names.arg = c("Nein", "Ja"), main="Überlebenschancen Frauen",
        xlab="Überlebt?", ylab ="Prozent")
```



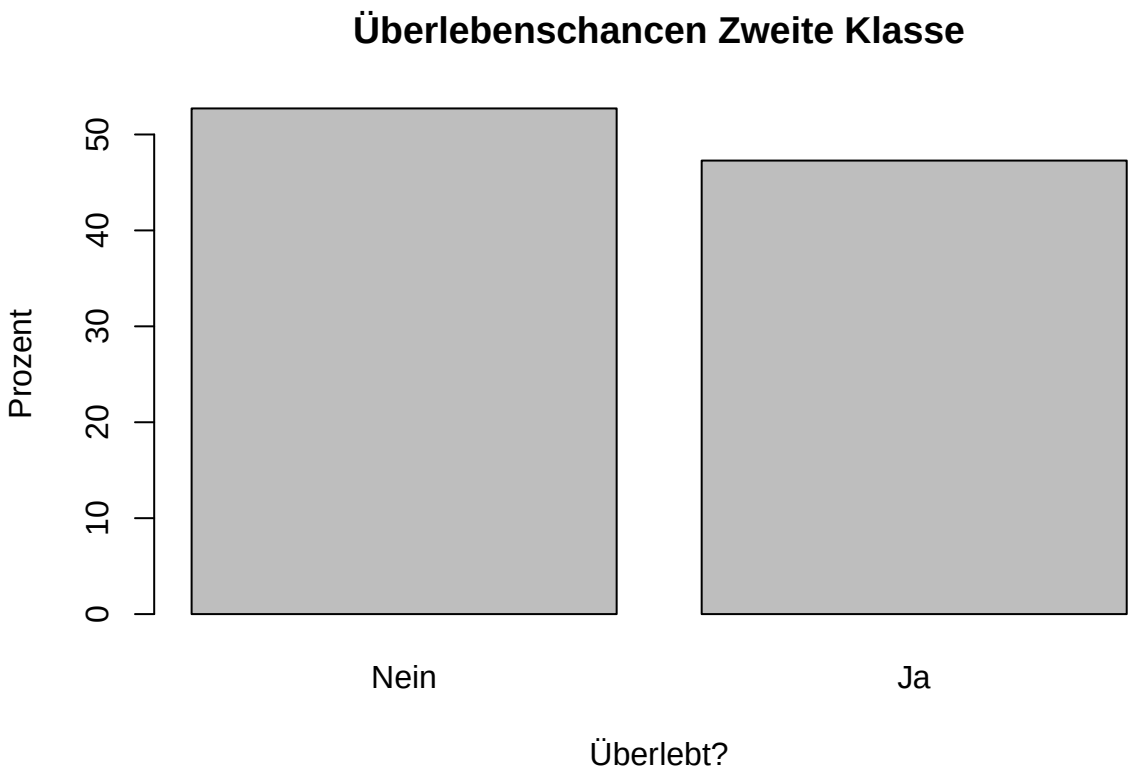

```r
barplot(f_men, names.arg = c("Nein", "Ja"), main="Überlebenschancen Männer",
        xlab="Überlebt?", ylab ="Prozent")
```



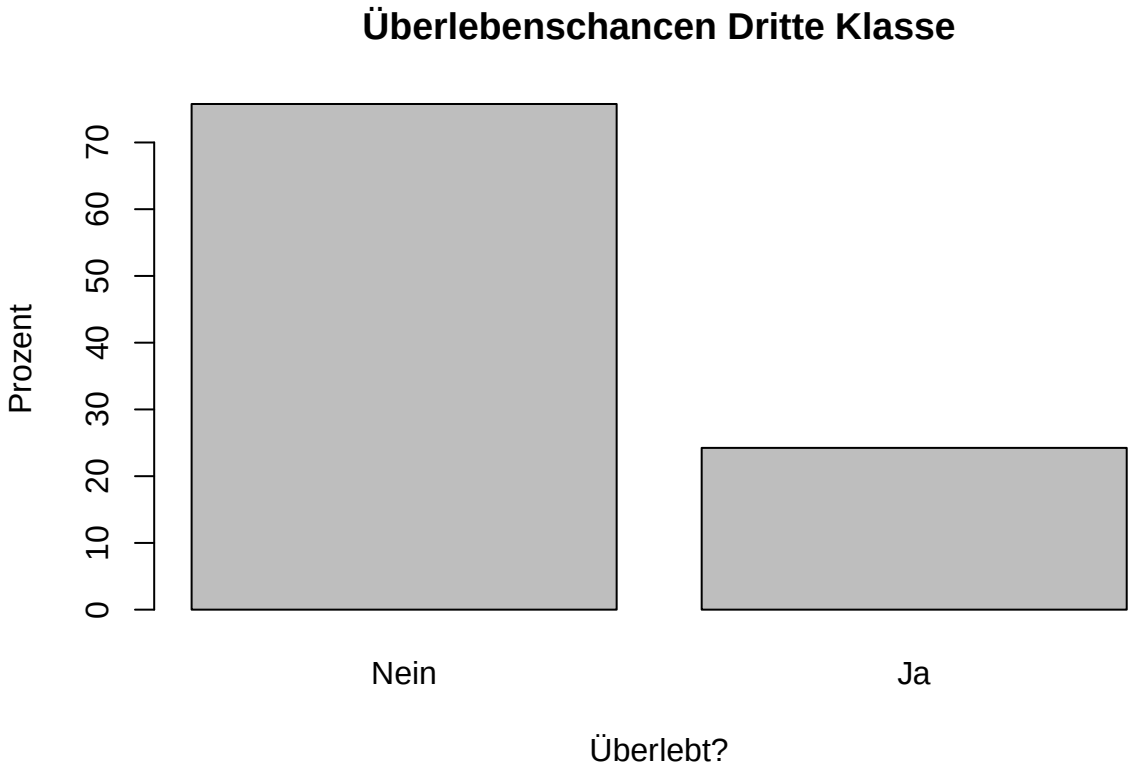

```
barplot(f_1, names.arg = c("Nein", "Ja"), main="Überlebenschancen Erste Klasse",
        xlab="Überlebt?", ylab ="Prozent")
```




```
barplot(f_2, names.arg = c("Nein", "Ja"), main="Überlebenschancen Zweite Klasse",
        xlab="Überlebt?", ylab ="Prozent")
```




```
barplot(f_3, names.arg = c("Nein", "Ja"), main="Überlebenschancen Dritte Klasse",
        xlab="Überlebt?", ylab ="Prozent")
```




```
# weitere Überlegungen bezüglich der Verhältnisse
h_gender <- table(titanic_data$Sex)
```

```
h_gender
```

```
##
## female   male
##    314    577
```

```
f_gender_relative_to_all <- prop.table(table(titanic_data$Sex, titanic_data$Survived))*100
f_gender_relative_to_all
```

```
##
##                   0         1
##   female  9.090909 26.150393
##   male   52.525253 12.233446
```

c) Beschreiben Sie kurz, was wir in b) über die Überlebenschancen verschiedener Passagiergruppen lernen!

Wir sehen, dass insgesamt mehr Passagiere gestorben sind als überlebt haben. Innerhalb der Gruppe der Frauen gab es mehr Überlebende als Verstorbene - Frauen hatten also eine recht gute Überlebenschance. Innerhalb der Gruppe der Männer gab es mehr Verstorbene als Überlebende - Männer hatten also eher schlechte Überlebenschancen. Für die Passagiere der ersten Klasse waren die Überlebenschancen recht gut, für die zweite Klasse fast 50/50, für die dritte Klasse schlecht. (Genauere Angaben?)

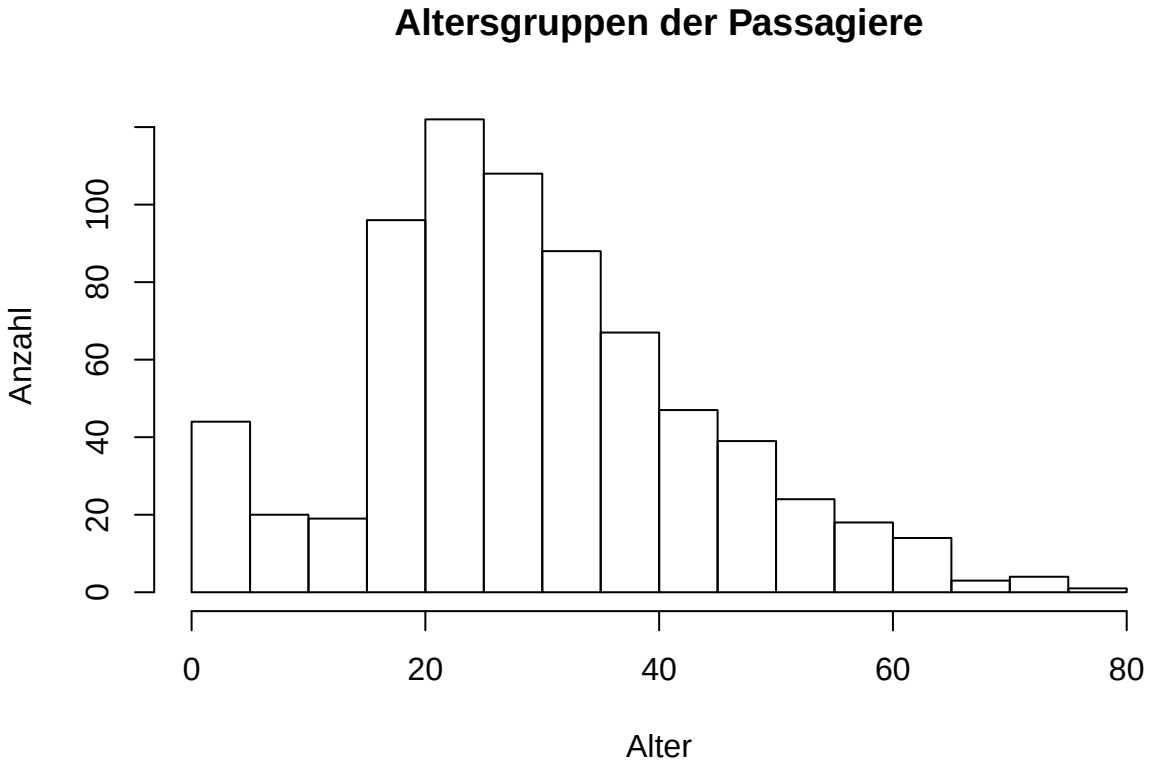
Aufgabe 3

Geben Sie den Diagrammen in den folgenden Aufgaben sinnvolle Titel und beschriften Sie alle Axen!

Erstellen Sie die Histogramme in a) und b) mit der Gruppierung [0,5), [5,10), [10, 15) usw! Beachten Sie, zu welcher Gruppe die Grenzwerte gehören.

a) Erzeugen Sie ein Histogramm absoluter Häufigkeiten des Merkmals `Age`!

```
bins <- c(0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80)
hist(titanic_data$Age, breaks = bins, main = "Altersgruppen der Passagiere", xlab = "Alter", ylab =
```



b) Erzeugen Sie jeweils ein Histogramm relativer Häufigkeiten des Merkmals `Age` für die Überlebenden und die Gestorbenen!

c) Beschreiben Sie kurz, was wir in b) über die Überlebenschancen verschiedener Altersgruppen lernen!