Applied Data Analysis, Übung 1

Henry Haustein

Task 1

```
install.packages("readxl")
library(readxl)
data = read_excel("data.xlsx", na = "NA")
```

Task 2

```
str(data)
summary(data)
head(data, 1)["age"]
tail(data, 1)["age"]
dim(data)[1]
```

Wir stellen fest, dass die numerische Kodierung bei einigen Variablen nicht gut umgesetzt ist: Manchmal starten die Level bei 0, mal bei 1 und nicht immer sind alle Zahlen genutzt. Zudem ist in der Spalte *Alter* nicht das Alter verzeichnet, sondern das Geburtsdatum.

Die erste Person wurde 1967 geboren, die letzte 1996. Insgesamt gibt es 437 Beobachtungen in der Tabelle.

Task 3

```
data$gender = factor(data$gender)
levels(data$gender) = c("male", "female", "diverse")
data$employment = factor(data$employment)
levels(data$employment) = c("student", "employed", "unemployed")
data$education = factor(data$education)
levels(data$education) = c("no degree", "secondary", "intermediate
    ", "high school", "academic")
data$play_frequency = factor(data$play_frequency)
levels(data$play_frequency) = c("never", "every few months", "
    every few weeks", "1-2 days a week", "3-5 days a week", "daily
    ")
data$treatment = factor(data$treatment)
levels(data$treatment) = c("control", "lootbox in task reward", "
    lootbox picture", "badge")
data$age = sapply(data$age, function(year) {2016-year})
data$rt6 = as.numeric(data$rt6)
data$rt7 = as.numeric(data$rt7)
```

```
14 data$rt8 = as.numeric(data$rt8)
15 data$rt9 = as.numeric(data$rt9)
16 data$rt10 = as.numeric(data$rt10)
17 data$rt11 = as.numeric(data$rt11)
18 data$rt12 = as.numeric(data$rt12)
19 data$rt13 = as.numeric(data$rt13)
20 data$rt14 = as.numeric(data$rt14)
```

Statt

```
1 data$rt6 = as.numeric(data$rt6)
```

könnte/sollte man

```
1 data$rt6 = sapply(data$rt6, as.numeric)
```

nutzen, aber das erzeugt benannte Listen in den Spalten, statt den Typ auf `num` zu ändern. Deswegen habe ich das angepasst.

Task 4

```
1 subsetControl = subset(data, treatment == "control")
2 subsetLootPic = subset(data, treatment == "lootbox picture")
3 summary(data$tasks_completed)
4 summary(subsetControl$tasks_completed)
5 summary(subsetLootPic$tasks_completed)
```

Wir sehen, dass der Mittelwert über alle Daten bei 10.06 liegt, in der Kontrollgruppe bei 7.788 und in der Lootbox-Picture-Gruppe bei 8.807.