Applied Data Analysis, Übung 2

Henry Haustein

Task 1

```r
install.packages("dplyr")
library(dplyr)

# Bekannt aus Uebung 1
install.packages("readxl")
library(readxl)
data = read_excel("data.xlsx", na = "NA")
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$rt7)
data$rt8 = as.numeric(data$rt8)
data$rt9 = as.numeric(data$rt9)
data$rt10 = as.numeric(data$rt10)
data$rt11 = as.numeric(data$rt11)
data$rt12 = as.numeric(data$rt12)
data$rt13 = as.numeric(data$rt13)
data$rt14 = as.numeric(data$rt14)

data = data %>% filter(age <= 80) %>% filter(age >= 18)
data = data %>% filter(`total time` <= 2000000)
data = data %>% mutate(old = age > 25)
summary(data)

data %>% group_by(treatment) %>% summarise(completeTasks_mean =
```

```
    mean(tasks_completed), completeTasks_median = median(tasks_
    completed), completeTasks_var = sd(tasks_completed))
35
36 subsetControl = data %>% filter(treatment == "control")
37 subsetLootPic = data %>% filter(treatment == "lootbox picture")
38 subsetLootInTask = data %>% filter(treatment == "lootbox in task
    reward")
39 subsetBadge = data %>% filter(treatment == "badge")
40 subsetYoung = data %>% filter(old == FALSE)
41 subsetOld = data %>% filter(old == TRUE)
```

121 Personen sind alt und 268 sind jung.

Am meiste Aufgaben wurden in der Gruppe mit der Lootbox im task reward erfüllt, und die Gruppe mit der Badge hatte die größte Standardabweichung.

Task 2

```
1 install.packages("ggplot2")
2 library(ggplot2)
3
4 ggplot(data, aes(x = tasks_completed)) + geom_histogram(binwidth =
     1)
5 ggplot(data, aes(x = tasks_completed, fill = treatment)) + geom_
    histogram(binwidth = 1)
6 ggplot(data, aes(x = tasks_completed)) + geom_histogram(binwidth =
     1) + facet_wrap(~ treatment)
7
8 ggplot(data, aes(x = treatment, y = tasks_completed)) + geom_
    boxplot()
9 ggplot(data, aes(y = tasks_completed)) + geom_boxplot() + facet_
    wrap(~ treatment)
10
11 ggplot(data, aes(x = `total time`)) + geom_density()
12 ggplot(data, aes(x = `total time`, color = treatment)) + geom_
    density()
13 ggplot(data, aes(x = `total time`)) + geom_density() + facet_wrap(
    ~ treatment)
```

Task 3

```
1 t.test(data$`total time`, mu = 320000, alternative = "two.sided",
    conf.level = 0.95)
2 t.test(subsetYoung$`total time`, mu = 320000, alternative = "two.
    sided", conf.level = 0.95)
3 t.test(subsetOld$`total time`, mu = 320000, alternative = "two.
    sided", conf.level = 0.95)
4
5 t.test(subsetYoung$tasks_completed, mu = mean(subsetOld$tasks_
    completed), alternative = "two.sided", conf.level = 0.95)
```

```
6
7 pairwise.t.test(data$tasks_completed, data$treatment, p.adjust.
      method = "none", pool.sd = FALSE)
```

Bei allen Personen und den jungen Personen ist der p-value unter 5%, das heißt wir können H_0 nicht ablehnen. Bei den alten Personen ist der p-value bei 50%, wir lehnen H_0 ab.

Bei allen andern Tests sind die p-values unter 5%, auch da lehnen wir H_0 ab. Die Mittelwerte sind also verschieden voneinander.