1.2 Hands-On: Wrangling

Marco Wähner

2022-03-25

Aufgabenstellung - Lösung

Aufgabe 1

Import der GLES Vor- und Nachwahlbefragung 2021

```
library(rio)
getwd()
```

```
## [1] "C:/Users/Marco/Desktop/Lehre/einfuehrung-in-R-hhu"
```

```
df <- import("data/gles_2021_gesamt.sav")
```

Aufgabe 2

Geben Sie die Variablennamen des Datensatzes aus.

```
names(df)
```

Aufgabe 3

Im Datensatz sind sowohl die Vorwahl- als auch die Nachwahl-Daten. In der Variable "sample" ist die jeweilige Befragung gespeichert. Geben Sie die Eigenschaften der Variable "sample" aus.

```
attributes(df$sample)
```

```
## $label
## [1] "Stichprobe"
##
## $format.spss
## [1] "F32.0"
##
## $labels
##  GLES Querschnitt 2021, Vorwahl GLES Querschnitt 2021, Nachwahl
##                               7                               8
```

Aufgabe 4

Filtern Sie nun nach der Variable “sample” und speichern Sie die Befragten aus der Nachwahl-Befragung im neuen Dataframe “df_post” ab.

```
#option 1
df_post <- df[df$sample == 8, ]

#option 2
df_post_2 <- subset(df, sample == 8)
```

Aufgabe 5

Nutzen Sie nur die Daten aus der Nachwahl-Befragung. Die Variable q68 enthält Angaben zur Internetnutzung der Befragten. Erstellen Sie eine neue Variable „online“ und klassifizieren Sie die Variable in folgende Kategorien: 1. Seltener als 1 Tag die Woche 2. 1 bis einschließlich 3 Tage die Woche 3. 4 bis einschließlich 5 Tage die Woche 4. 6 bis einschließlich 7 Tage die Woche Befragte die angeben, dass Sie nie das Internet nutzen (Wert 8) oder einen Internetzugang haben (Wert 9) werden als Missing Values definiert. Geben Sie die absolute Häufigkeit der neuen Variablen an. Wie viele Missing Values hat diese neue Variable?

```
df_post$online <- NA

df_post$online[df_post$q68 == 0] <- 1
df_post$online[df_post$q68 == 1 | df_post$q68 == 2 | df_post$q68 == 3] <- 2
df_post$online[df_post$q68 == 4 | df_post$q68 == 5] <- 3
df_post$online[df_post$q68 == 6 | df_post$q68 == 7] <- 4


#kontrolle
table(df_post$q68, useNA = "ifany")
```

```
##
##  -99  -97  -93  -73    0    1    2    3    4    5    6    7    8    9
##   42  324   43    7   48   47   43   76   84  146  128 2359   51   26
```

```
table(df_post$online, useNA = "ifany")
```

```
##
##    1    2    3    4 <NA>
##   48  166  230 2487  493
```

Aufgabe 6

Die Variable „q73b“ beinhaltet Informationen darüber, ob die Befragten in den vergangenen 12 Monaten an einer Demonstration teilgenommen haben. Erstellen Sie eine Dummy-Variable mit dem Namen “demo_dummy”, die den Wert „0“ annimmt, wenn die Befragten nicht an einer Demonstration teilgenommen haben und den Wert „1“ annimmt, wenn die Befragten an einer Demonstration teilgenommen haben. Wie hoch ist der Anteil der Befragten, die an einer Demo teilgenommen haben? Wie viele fehlende Werte hat die ursprüngliche Variable?

```r
attributes(df_post$q70b)
```

```
## NULL
```

```r
#attributes sind durch subsetting NULL

table(df_post$q73b)
```

```
##
##  -99  -93  -73    0    1
##   78    1    4 2955  386
```

```r
sum(is.na(df$q73b))
```

```
## [1] 0
```

```r
#Missing Values müssen kodiert werden
df_post$q73b[df_post$q73b %in% c(-99,-93,-73)] <- NA

sum(is.na(df_post$q73b))
```

```
## [1] 83
```

```r
df_post$demo_dummy <- ifelse(df_post$q73b == 1, 1, 0)

table(df_post$demo_dummy)
```

```
##
##    0    1
## 2955  386
```

```r
sum(is.na(df_post$demo_dummy))
```

```
## [1] 83
```

```r
table(df_post$demo_dummy, useNA = "ifany")
```

```
##
##    0    1 <NA>
## 2955  386   83
```

Aufgabe 7

Berechnen Sie aus dem Geburtsjahr (Variable d2a) das Alter der Befragten(zur Erinnerung: Die Erhebung wurde 2021 durchgeführt). Speichern Sie das Alter der Befragten als "age" im Datensatz zur Nachwahl-Befragung ab. Was ist das Problem mit der Klasse der ursprünglichen Variable d2a? Nutzen Sie ggf. die Funktion "as.numeric", um die Klasse der Variable zu ändern.

```r
#Variable hat KLasse character
class(df_post$d2a)
```

```
## [1] "character"
```

```r
#Klasse umwandeln
df_post$d2a <- as.numeric(df_post$d2a)
```

```
## Warning: NAs durch Umwandlung erzeugt
```

```r
class(df_post$d2a)
```

```
## [1] "numeric"
```

```r
#Alter aus Geburtsjahr
df_post$age <- 2021 - df_post$d2a
```

##Aufgabe 8 Wählen Sie folgende Variablen aus dem Datensatz zur Nachbefragung aus: online, demo_dummy, age und q1. Diese Variablen sollen in einem neuen Datensatz gespeichert werden. Exportieren Sie den neuen Datensatz im .RDATA-Format mit den vier Variablen.

```r
#subsetting
df_sub <- df_post[ ,c("age", "online", "demo_dummy", "q1")]

save(df_sub, file = "data/df_export.RDATA")
```