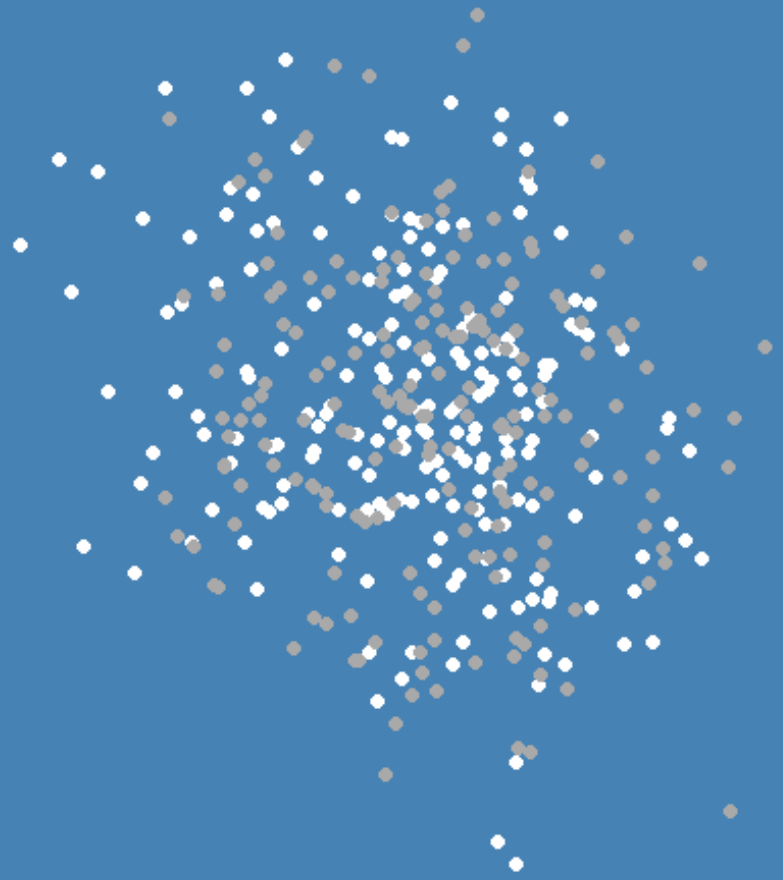
Introduction to R

2.5 Transforming Variables

Class conversions, mutate(), recode()

Lion Behrens, M.Sc.



University of Mannheim
Chair of Social Data Science and Methodology
Chair of Quantitative Methods in the Social Sciences

Data Import

```
library(haven)
ess10 <- haven::read_dta("./dat/ESS10.dta")
dim(ess10) # check dimensionality of data frame
```

```
## [1] 18060   513
```

```
print(ess10[1:10, 1:10])
```

```
## # A tibble: 10 × 10
##    name       essro…¹ edition prodd…²  idno cntry dweight pweight nwspol netus…³
##    <chr>        <dbl> <chr>   <chr>   <dbl> <chr>   <dbl>   <dbl> <dbl+> <dbl+l>
##  1 ESS10e01_2      10 1.2     28.06.… 10002 BG      1.03    0.218  80    1 [Nev…
##  2 ESS10e01_2      10 1.2     28.06.… 10006 BG      0.879   0.218  63    5 [Eve…
##  3 ESS10e01_2      10 1.2     28.06.… 10009 BG      1.01    0.218 390    5 [Eve…
##  4 ESS10e01_2      10 1.2     28.06.… 10024 BG      0.955   0.218  60    5 [Eve…
##  5 ESS10e01_2      10 1.2     28.06.… 10027 BG      0.841   0.218 120    5 [Eve…
##  6 ESS10e01_2      10 1.2     28.06.… 10048 BG      0.946   0.218  60    5 [Eve…
##  7 ESS10e01_2      10 1.2     28.06.… 10053 BG      1.01    0.218  30    5 [Eve…
##  8 ESS10e01_2      10 1.2     28.06.… 10055 BG      1.03    0.218  70    5 [Eve…
##  9 ESS10e01_2      10 1.2     28.06.… 10059 BG      0.991   0.218  60    1 [Nev…
## 10 ESS10e01_2      10 1.2     28.06.… 10061 BG      1.05    0.218  60    1 [Nev…
## # … with abbreviated variable names ¹essround, ²proddate, ³netusoft
```

Changing the data type: Class conversions

Class Conversions: Overview base R

In base R, simply take a function as._() and wrap it around the object that should change its data type.

Conceptual

```
as.numeric()
as.character()
as.factor()
as.logical()
```

Class Conversions: Example base R

Sometimes we have to work with messy data, e.g. numeric information coming as a character variable.

Of course, we can convert a character variable into a numeric format.

First, let's create a character variable:

```
# generate character variable storing numeric information
var_char <- as.character(runif(5, min = 0, max = 10))
```

```
print(var_char)
```

```
## [1] "4.07960083568469" "7.34801860526204" "5.82581334980205" "1.63197547197342"
## [5] "8.23638678062707"
```

Second, let's convert it into numeric:

```
# generate character variable storing numeric information
var_numeric <- as.numeric(var_char)
print(var_numeric)
```

```
## [1] 4.079601 7.348019 5.825813 1.631975 8.236387
```

Class Conversions: Example base R

Since we imported our dataset with the package haven, the variable gndr is stored as a variable of class haven_labelled where 1 = Male and 2 = Female.

```
class(ess10$gndr)
```

```
## [1] "haven_labelled" "vctrs_vctr"     "double"
```

```
table(ess10$gndr)
```

```
##
##    1    2
## 8100 9960
```

Class Conversions: Example base R

Let's change this variable into a character variable.

```
ess10$gndr_char <- as.character(ess10$gndr) # transform to character
class(ess10$gndr_char)
```

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

```
table(ess10$gndr_char)
```

```
##
##    1    2
## 8100 9960
```

Class Conversions: Example base R

Alternatively, we can change it into a factor variable.

```
ess10$gndr_fac <- as.factor(ess10$gndr)
class(ess10$gndr_fac)
```

```
## [1] "factor"
```

```
table(ess10$gndr_fac)
```

```
##
##    1    2
## 8100 9960
```

Class Conversions: Example dplyr

The dplyr package let's us easily change the classes of variables by combining the respective functions (as.numeric(), as.character() etc.) with mutate().

```
table(ess10$gndr)
```

```
##
##    1    2
## 8100 9960
```

```
class(ess10$gndr)
```

```
## [1] "haven_labelled" "vctrs_vctr"     "double"
```

Class Conversions: Example dplyr

The dplyr package let's us easily change the classes of variables by combining the respective functions (as.numeric(), as.character() etc.) with mutate().

```
table(ess10$lrscale)
```

```
##
##    0    1    2    3    4    5    6    7    8    9   10
##  718  371  732 1119 1062 5457 1526 1696 1499  541 1031
```

```
class(ess10$lrscale)
```

```
## [1] "haven_labelled" "vctrs_vctr"     "double"
```

Class Conversions: Example dplyr

The dplyr package let's us easily change the classes of variables by combining the respective functions (as.numeric(), as.character() etc.) with mutate().

```
ess10 <- ess10 %>%
  mutate(gndr_char = as.character(gndr),
         lrscale_num = as.numeric(lrscale))
```

```
table(ess10$gndr_char)
```

```
##
##    1    2
## 8100 9960
```

```
class(ess10$gndr_char)
```

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

Class Conversions: Example dplyr

The dplyr package let's us easily change the classes of variables by combining the respective functions (as.numeric(), as.character() etc.) with mutate().

```
ess10 <- ess10 %>%
  mutate(gndr_char = as.character(gndr),
         lrscale_num = as.numeric(lrscale))
```

```
table(ess10$lrscale_num)
```

```
##
##    0    1    2    3    4    5    6    7    8    9   10
##  718  371  732 1119 1062 5457 1526 1696 1499  541 1031
```

```
class(ess10$lrscale_num)
```

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

Recoding values

Simple Recodings

Simple recodings can be performed in base R.

Let's take our variable gender (coded 1/2)

- let it take on values of either 0 or 1
- convert it into a factor variable
- set 0 to "Male" and 1 to "Female"

Simple Recodings

Simple recodings can be performed in base R.

Let's take our variable gender (coded 1/2)

- let it take on values of either 0 or 1
- convert it into a factor variable
- set 0 to "Male" and 1 to "Female"

```
ess10$female <- ess10$gndr - 1
```

```
table(ess10$female)
```

```
##
##    0    1
## 8100 9960
```

Simple Recodings

Simple recodings can be performed in base R.

Let's take our variable gender (coded 1/2)

- let it take on values of either 0 or 1
- convert it into a factor variable
- set 0 to "Male" and 1 to "Female"

```
ess10$female <- as.factor(ess10$female)
```

```
ess10$female <- ifelse(ess10$female == "1",
                       "Female", "Male")
```

```
table(ess10$female)
```

```
##
## Female   Male
##   9960   8100
```

mutate() for Numeric Variables - Transformation of Scale

Let's look at the variable political interest.

```
table(ess10$polintr)
```

```
##
##    1    2    3    4
## 1319 5469 7093 4137
```

Often, we want a variable's range to start of 0 such that the constant of a regression model becomes substantively interpretable if the variable is included as an independent predictor.

```
ess10 <- ess10 %>%
  mutate(pol_interest = polintr - 1)
```

```
table(ess10$pol_interest)
```

```
##
##    0    1    2    3
## 1319 5469 7093 4137
```

dplyr::mutate() - Reverse Scale of Numeric Variable

What if I told you that...

higher values on the variable pol_interest actually indicate lower amounts of political interest?

> From the European Social Survey Main Questionnaire, Wave 10

SECTION B

Now we want to ask a few questions about politics and government.

B1 How interested would you say you are in politics – are you… **READ OUT**…

very interested,	1
quite interested,	2
hardly interested,	3
or, not at all interested?	4

dplyr::mutate() - Reverse Scale of Numeric Variable

For numeric variables, there is a straightforward way to reverse the scale.

- if k is the maximum value of the scale
- *var_reversed* = (*old_var* *-1) + k

```
ess10 <- ess10 %>%
  mutate(pol_interest_res = (pol_interest * -1) + 3)
```

```
table(ess10$pol_interest)
```

```
##
##    0    1    2    3
## 1319 5469 7093 4137
```

```
table(ess10$pol_interest_res)
```

```
##
##    0    1    2    3
## 4137 7093 5469 1319
```

dplyr::mutate() - Recoding

Let's collapse respondents into either being left-leaning, right-leaning, or in the center.

```
table(ess10$lrscale)
```

```
##
##    0    1    2    3    4    5    6    7    8    9   10
##  718  371  732 1119 1062 5457 1526 1696 1499  541 1031
```

dplyr::mutate() - Recoding

Let's collapse respondents into either being left-leaning, right-leaning, or in the center.

```
ess10 <- ess10 %>%
  mutate(lr_binary = as.numeric(lrscale)) %>%
  mutate(lr_binary = recode(lr_binary,
                            `0` = 0,
                            `1` = 0,
                            `2` = 0,
                            `3` = 0,
                            `4` = 0,
                            `5` = 1,
                            `6` = 2,
                            `7` = 2,
                            `8` = 2,
                            `9` = 2,
                            `10` = 2
                            )
  )
```

dplyr::mutate() - Recoding

Let's collapse respondents into either being left-leaning, right-leaning, or in the center.

```
table(ess10$lr_binary, ess10$lrscale)
```

```
##
##         0    1    2    3    4    5    6    7    8    9   10
##   0   718  371  732 1119 1062    0    0    0    0    0    0
##   1     0    0    0    0    0 5457    0    0    0    0    0
##   2     0    0    0    0    0    0 1526 1696 1499  541 1031
```

dplyr::mutate() - Recoding

If you want to generate (or recode) a factor variable, simply use recode_factor().

The functionality stays the same.

```
ess10 <- ess10 %>%
  mutate(gender = as.factor(gndr)) %>%
  mutate(gender = recode_factor(gender,
                                `1` = "Male",
                                `2` = "Female"
                                )
         )
```

dplyr::mutate() - Recoding

We can chain several variable transformations into one single statement...

and generate a whole data wrangling pipeline consisting of only few commands!

```
ess10 <- ess10 %>%
  mutate(gender = as.factor(gndr)) %>%
  mutate(gender = recode_factor(gender,
                                `1` = "Male",
                                `2` = "Female"
                                )
         )
```

dplyr::mutate() - Recoding

We can chain several variable transformations into one single statement...

and generate a whole data wrangling pipeline consisting of only few commands!

```
ess10 <- ess10 %>%
  mutate(gender = as.factor(gndr),
         voted = as.factor(vote)) %>%
  mutate(gender = recode_factor(gender,
                                `1` = "Male",
                                `2` = "Female"),
         voted = recode_factor(voted,
                               `1` = "Yes",
                               `2` = "No",
                               `3` = "Not eligible"
                               )
         )
```

```r
ess10 <- ess10 %>%
  mutate(gender = as.factor(gndr),
         voted = as.factor(vote)) %>%
  mutate(gender = recode_factor(gender,
                                `1` = "Male",
                                `2` = "Female"),
         voted = recode_factor(voted,
                               `1` = "Yes",
                               `2` = "No",
                               `3` = "Not eligible"
                               )
         )
```

```r
ess10 <- ess10 %>%
  mutate(gender = as.factor(gndr),
         voted = as.factor(vote),
         party_choice = as.factor(prtvtefr)) %>%
  mutate(gender = recode_factor(gender,
                                `1` = "Male",
                                `2` = "Female"),
         voted = recode_factor(voted,
                               `1` = "Yes",
                               `2` = "No",
                               `3` = "Not eligible"),
         party_choice = recode_factor(party_choice,
                                      `1`  = "Lutte Ouvriére",
                                      `2`  = "Nouv. Parti Anti-Capitaliste",
                                      `3`  = "Parti Communiste Français",
                                      `4`  = "La France Insoumise",
                                      `5`  = "Parti Socialiste",
                                      `6`  = "Europe Ecologie Les Verts",
                                      `7`  = "La République en Marche",
                                      `8`  = "Mouvement Démocrate",
                                      `9`  = "Les Républicains",
                                      `10` = "Debout la France",
                                      `11` = "Front National",
                                      `12` = "Other",
                                      `13` = "Blank",
                                      `14` = "Null"
                                      )
         )
```

```r
ess10 <- ess10 %>%
  mutate(gender = as.factor(gndr),
         voted = as.factor(vote),
         party_choice = as.factor(prtvtefr)) %>%
  mutate(gender = recode_factor(gender,
                                `1` = "Male",
                                `2` = "Female"),
         voted = recode_factor(voted,
                               `1` = "Yes",
                               `2` = "No",
                               `3` = "Not eligible"),
         party_choice = recode_factor(party_choice,
                                      `1`  = "Lutte Ouvriére",
                                      `2`  = "Nouv. Parti Anti-Capitaliste",
                                      `3`  = "Parti Communiste Français",
                                      `4`  = "La France Insoumise",
                                      `5`  = "Parti Socialiste",
                                      `6`  = "Europe Ecologie Les Verts",
                                      `7`  = "La République en Marche",
                                      `8`  = "Mouvement Démocrate",
                                      `9`  = "Les Républicains",
                                      `10` = "Debout la France",
                                      `11` = "Front National",
                                      `12` = "Other",
                                      `13` = "Blank",
                                      `14` = "Null"
                                      )
         )
```

```r
ess10 <- ess10 %>%
  mutate(gender = as.factor(gndr),
         voted = as.factor(vote),
         party_choice = as.factor(prtvtefr)) %>%
  mutate(gender = recode_factor(gender,
                                `1` = "Male",
                                `2` = "Female"),
         voted = recode_factor(voted,
                               `1` = "Yes",
                               `2` = "No",
                               `3` = "Not eligible"),
         party_choice = recode_factor(party_choice,
                                      `1`  = "Lutte Ouvriére",
                                      `2`  = "Nouv. Parti Anti-Capitaliste",
                                      `3`  = "Parti Communiste Français",
                                      `4`  = "La France Insoumise",
                                      `5`  = "Parti Socialiste",
                                      `6`  = "Europe Ecologie Les Verts",
                                      `7`  = "La République en Marche",
                                      `8`  = "Mouvement Démocrate",
                                      `9`  = "Les Républicains",
                                      `10` = "Debout la France",
                                      `11` = "Front National",
                                      `12` = "Other",
                                      `13` = "Blank",
                                      `14` = "Null"
                                      )
         )
```

```r
ess10 <- ess10 %>%
  mutate(gender = as.factor(gndr),
         voted = as.factor(vote),
         party_choice = as.factor(prtvtefr)) %>%
  mutate(gender = recode_factor(gender,
                                `1` = "Male",
                                `2` = "Female"),
         voted = recode_factor(voted,
                               `1` = "Yes",
                               `2` = "No",
                               `3` = "Not eligible"),
         party_choice = recode_factor(party_choice,
                                      `1`  = "Lutte Ouvriére",
                                      `2`  = "Nouv. Parti Anti-Capitaliste",
                                      `3`  = "Parti Communiste Français",
                                      `4`  = "La France Insoumise",
                                      `5`  = "Parti Socialiste",
                                      `6`  = "Europe Ecologie Les Verts",
                                      `7`  = "La République en Marche",
                                      `8`  = "Mouvement Démocrate",
                                      `9`  = "Les Républicains",
                                      `10` = "Debout la France",
                                      `11` = "Front National",
                                      `12` = "Other",
                                      `13` = "Blank",
                                      `14` = "Null"
                                      )
         )
```

```r
ess10 <- ess10 %>%
  mutate(gender = as.factor(gndr),
         voted = as.factor(vote),
         party_choice = as.factor(prtvtefr)) %>%
  mutate(gender = recode_factor(gender,
                                `1` = "Male",
                                `2` = "Female"),
         voted = recode_factor(voted,
                               `1` = "Yes",
                               `2` = "No",
                               `3` = "Not eligible"),
         party_choice = recode_factor(party_choice,
                                      `1`  = "Lutte Ouvriére",
                                      `2`  = "Nouv. Parti Anti-Capitaliste",
                                      `3`  = "Parti Communiste Français",
                                      `4`  = "La France Insoumise",
                                      `5`  = "Parti Socialiste",
                                      `6`  = "Europe Ecologie Les Verts",
                                      `7`  = "La République en Marche",
                                      `8`  = "Mouvement Démocrate",
                                      `9`  = "Les Républicains",
                                      `10` = "Debout la France",
                                      `11` = "Front National",
                                      `12` = "Other",
                                      `13` = "Blank",
                                      `14` = "Null"
                                      )
         )
```

```r
ess10 <- ess10 %>%
  mutate(gender = as.factor(gndr),
         voted = as.factor(vote),
         party_choice = as.factor(prtvtefr)) %>%
  mutate(gender = recode_factor(gender,
                                `1` = "Male",
                                `2` = "Female"),
         voted = recode_factor(voted,
                               `1` = "Yes",
                               `2` = "No",
                               `3` = "Not eligible"),
         party_choice = recode_factor(party_choice,
                                      `1`  = "Lutte Ouvriére",
                                      `2`  = "Nouv. Parti Anti-Capitaliste",
                                      `3`  = "Parti Communiste Français",
                                      `4`  = "La France Insoumise",
                                      `5`  = "Parti Socialiste",
                                      `6`  = "Europe Ecologie Les Verts",
                                      `7`  = "La République en Marche",
                                      `8`  = "Mouvement Démocrate",
                                      `9`  = "Les Républicains",
                                      `10` = "Debout la France",
                                      `11` = "Front National",
                                      `12` = "Other",
                                      `13` = "Blank",
                                      `14` = "Null"
                                      )
         )
```

Recoding in base R

How can we achieve such recodings using base R?

It works equally well, the code just becomes a little convoluted...

```r
ess10$gender <- NA
ess10$gender[ess10$gndr == 1] <- "Male"
ess10$gender[ess10$gndr == 2] <- "Female"
ess10$gender <- as.factor(ess10$gender)
```

```r
ess10$gender <- NA
ess10$gender[ess10$gndr == 1] <- "Male"
ess10$gender[ess10$gndr == 2] <- "Female"
ess10$gender <- as.factor(ess10$gender)
```

```r
ess10$gender <- NA
ess10$gender[ess10$gndr == 1] <- "Male"
ess10$gender[ess10$gndr == 2] <- "Female"
ess10$gender <- as.factor(ess10$gender)
```

```r
ess10$gender <- NA
ess10$gender[ess10$gndr == 1] <- "Male"
ess10$gender[ess10$gndr == 2] <- "Female"
ess10$gender <- as.factor(ess10$gender)
```

```r
ess10$gender <- NA
ess10$gender[ess10$gndr == 1] <- "Male"
ess10$gender[ess10$gndr == 2] <- "Female"
ess10$gender <- as.factor(ess10$gender)
```

```r
ess10$voted <- NA
ess10$voted[ess10$vote == 1] <- "Yes"
ess10$voted[ess10$vote == 2] <- "No"
ess10$voted[ess10$vote == 3] <- "Not eligible"
ess10$voted <- as.factor(ess10$voted)
```

```r
ess10$party_choice <- NA
ess10$party_choice[ess10$prtvtefr == 1] <- "Lutte Ouvriére"
ess10$party_choice[ess10$prtvtefr == 2] <- "Nouv. Parti Anti-Capitaliste"
ess10$party_choice[ess10$prtvtefr == 3] <- "Parti Communiste Français"
ess10$party_choice[ess10$prtvtefr == 4] <- "La France Insoumise"
ess10$party_choice[ess10$prtvtefr == 5] <- "Parti Socialiste"
ess10$party_choice[ess10$prtvtefr == 6] <- "Europe Ecologie Les Verts"
ess10$party_choice[ess10$prtvtefr == 7] <- "La République en Marche"
ess10$party_choice[ess10$prtvtefr == 8] <- "Mouvement Démocrate"
ess10$party_choice[ess10$prtvtefr == 9] <- "Les Républicains"
ess10$party_choice[ess10$prtvtefr == 10] <- "Debout la France"
ess10$party_choice[ess10$prtvtefr == 11] <- "Front National"
ess10$party_choice[ess10$prtvtefr == 12] <- "Other"
ess10$party_choice[ess10$prtvtefr == 13] <- "Blank"
ess10$party_choice[ess10$prtvtefr == 14] <- "Null"
ess10$party_choice <- as.factor(ess10$party_choice)
```

The Data Wrangling Pipeline (I/III)

```
library(tidyverse)
ess10 <- haven::read_dta("./dat/ESS10.dta")
ess10 <- ess10 %>% # subset variables
  select(country = cntry, # sociodemographics
         gender = gndr,
         education_years = eduyrs,
         trust_social = ppltrst, # multidimensional trust
         trust_parliament = trstprl,
         trust_legalSys = trstlgl,
         trust_police = trstplc,
         trust_politicians = trstplt,
         trust_parties = trstprt,
         trust_EP = trstep,
         trust_UN = trstun,
         left_right = lrscale, # attitudes
         life_satisfaction = stflife,
         pol_interest = polintr,
         voted = vote, # turnout
         party_choice = prtvtefr # party choice
         ) %>%
  filter(country == "FR") # subset cases (only include France)
```

The Data Wrangling Pipeline (I/III)

```
library(tidyverse)
ess10 <- haven::read_dta("./dat/ESS10.dta")
ess10 <- ess10 %>% # subset variables
  select(country = cntry, # sociodemographics
         gender = gndr,
         education_years = eduyrs,
         trust_social = ppltrst, # multidimensional trust
         trust_parliament = trstprl,
         trust_legalSys = trstlgl,
         trust_police = trstplc,
         trust_politicians = trstplt,
         trust_parties = trstprt,
         trust_EP = trstep,
         trust_UN = trstun,
         left_right = lrscale, # attitudes
         life_satisfaction = stflife,
         pol_interest = polintr,
         voted = vote, # turnout
         party_choice = prtvtefr # party choice
         ) %>%
  mutate_at(c("country", "gender", "voted", "party_choice"), as.character) %>% # change types
  mutate_at("pol_interest", as.numeric) %>% # change types
  filter(country == "FR") # subset cases (only include France)
```

The Data Wrangling Pipeline (I/III)

```
library(tidyverse)
ess10 <- haven::read_dta("./dat/ESS10.dta")
ess10 <- ess10 %>% # subset variables
  select(country = cntry, # sociodemographics
         gender = gndr,
         education_years = eduyrs,
         trust_social = ppltrst, # multidimensional trust
         trust_parliament = trstprl,
         trust_legalSys = trstlgl,
         trust_police = trstplc,
         trust_politicians = trstplt,
         trust_parties = trstprt,
         trust_EP = trstep,
         trust_UN = trstun,
         left_right = lrscale, # attitudes
         life_satisfaction = stflife,
         pol_interest = polintr,
         voted = vote, # turnout
         party_choice = prtvtefr # party choice
         ) %>%
  mutate_at(c("country", "gender", "voted", "party_choice"), as.character) %>%
  mutate_at("pol_interest", as.numeric) %>%
  filter(country == "FR") # subset cases (only include France)
```

The Data Wrangling Pipeline (II/III)

```
ess10 <- ess10 %>%
  mutate(gender = recode_factor(gender,
                                `1` = "Male",
                                `2` = "Female"),
         voted = recode_factor(voted,
                               `1` = "Yes",
                               `2` = "No",
                               `3` = "Not eligible"),
         party_choice = recode_factor(party_choice,
                                      `1`  = "Lutte Ouvriére",
                                      `2`  = "Nouv. Parti Anti-Capitaliste",
                                      `3`  = "Parti Communiste Français",
                                      `4`  = "La France Insoumise",
                                      `5`  = "Parti Socialiste",
                                      `6`  = "Europe Ecologie Les Verts",
                                      `7`  = "La République en Marche",
                                      `8`  = "Mouvement Démocrate",
                                      `9`  = "Les Républicains",
                                      `10` = "Debout la France",
                                      `11` = "Front National",
                                      `12` = "Other",
                                      `13` = "Blank",
                                      `14` = "Null")
         )
```

The Data Wrangling Pipeline (III/III)

```
ess10 <- ess10 %>%
  mutate(pol_interest = (pol_interest * -1) + 5, # invert scale
         life_satisfaction = life_satisfaction + 1 # change scale to [1, 11]
         )
```

References

Parts of this course are inspired by the following resources:

- Wickham, Hadley and Garrett Grolemund, 2017. *R for Data Science - Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly.
- Bahnsen, Oke and Guido Ropers, 2022. *Introduction to R for Quantitative Social Science*. Course held as part of the GESIS Workshop Series.
- Breuer, Johannes and Stefan Jünger, 2021. *Introduction to R for Data Analysis*. Course held as part of the GESIS Summer School in Survey Methodology.
- Teaching material developed by Verena Kunz, David Weyrauch, Oliver Rittmann and Viktoriia Semenova.