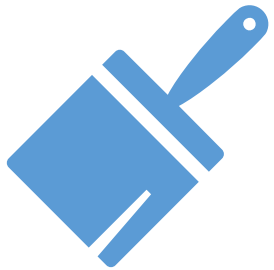
Statistik 2 – Tutorate
Sitzung 2: Datenmanagement

Marco Giesselmann, Rémy Blum, Federica Bruno, Rebecca Hobel, Kristina Trajkovic

Lernziele dieser Sitzung



Datenbereinigung

Löschen irrelevanter Objekte

Variablenwerte rekodieren
(Missings, Klassifizieren, Rechenoperation)

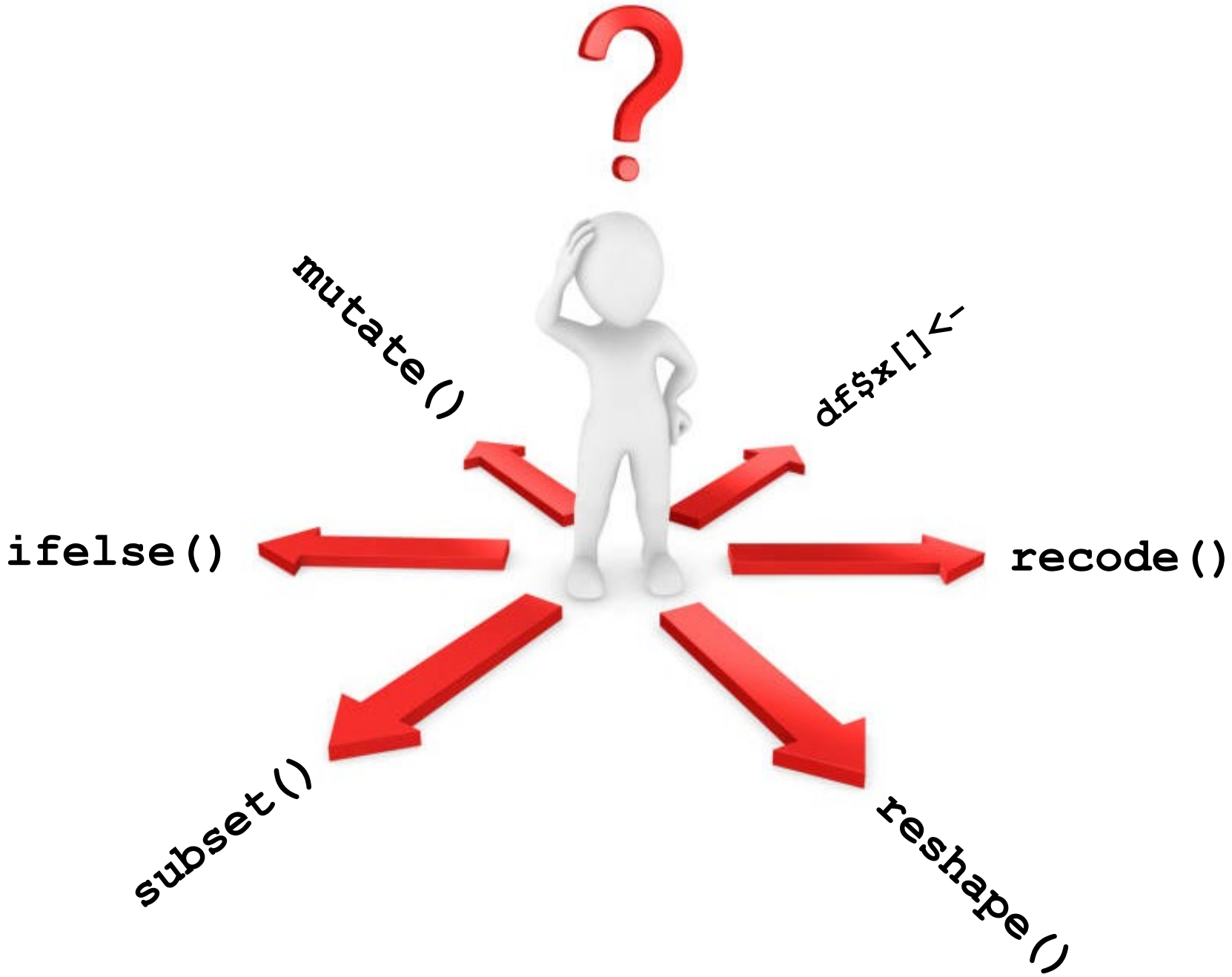
Variablen umbenennen

Datenauswahl

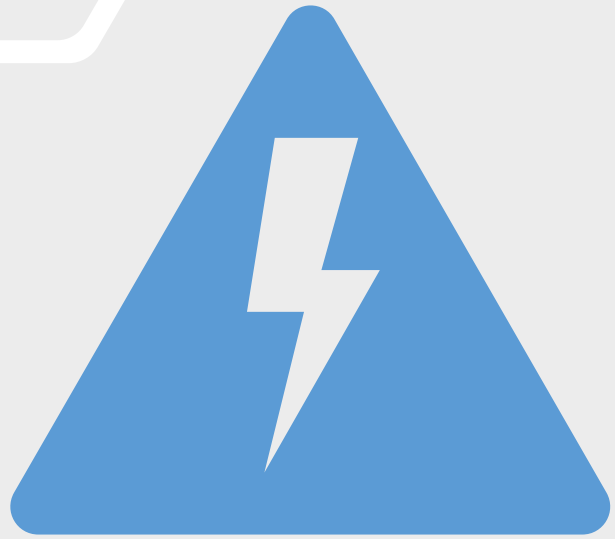
Selektieren: Teildatensätze bilden

Filtern: Teilstichproben bilden

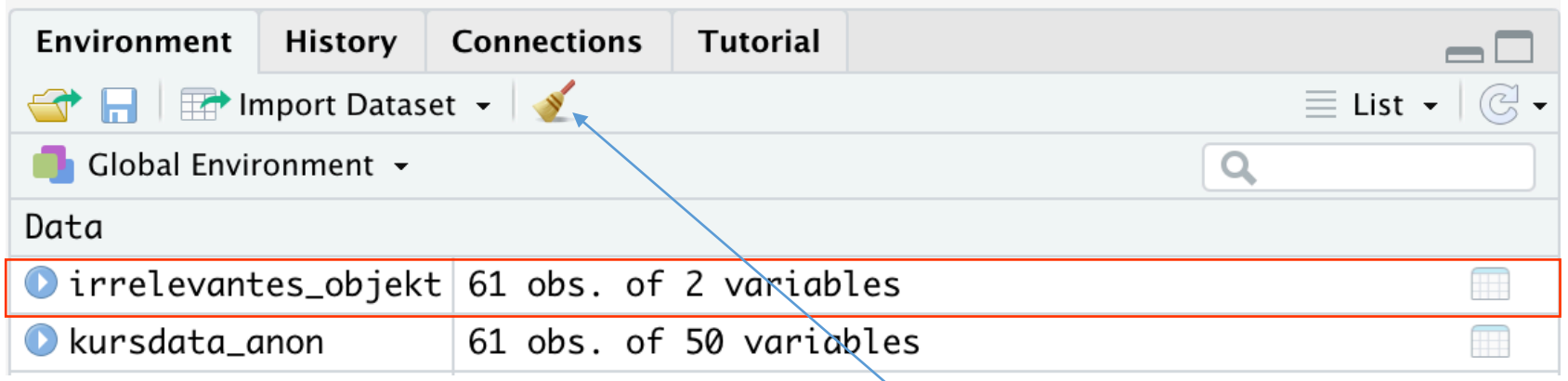
Filtern: Missings eliminieren


mutate()
df$x[]<-
ifelse()
recode()
subset()
reshape()

Bereinigen: Aufräumen



1 Löschen irrelevanter Objekte



remove() löscht ein einzelnes Objekt

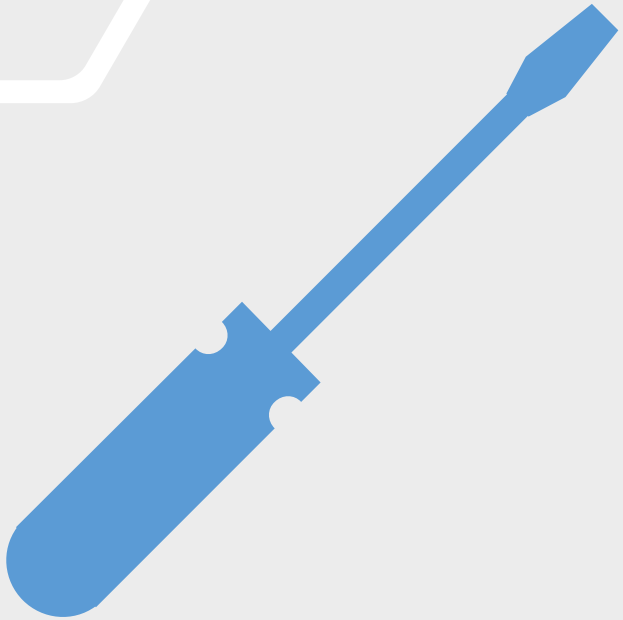
```
remove(irrelevantes_objekt)
```

rm() löscht alle Objekte aus dem Environment:

```
rm(list = ls())
```

Alternativ: Mit dem Besen-Symbol

Bereinigen: Rekodieren



2 Variablenwerte Rekodieren

Wir wollen die Ausprägungen der Variable **rlgblg** rekodieren (**2 zu 0**), dazu erstellen wir die Variable **rlgblg_neu**

`ess8$rlgblg_neu <- ess8$rlgblg` -> *Neue Variable wird gebildet und mit den Werten der alten angereichert*

`ess8$rlgblg_neu <- 0` -> *Die neue Variable wird überschrieben (alle Ausprägungen)*

`ess8$rlgblg_neu <- ess8$rlgblg` -> *Neustart: Zum Schritt 1*

`ess8$rlgblg_neu[ess8$rlgblg == 2] <- 0`

rlgblg Belonging to particular religion or denomination
2
2
1
NA
1
1
2
1
2
1
1
2
1
1
2
1
2
2

rlgblg_neu Belonging to particular religion or denomination
2
2
1
NA
1
1
2
1
2
1
1
2
1
1
2
1
2
2

rlgblg_neu Belonging to particular religion or denomination
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0

rlgblg_neu Belonging to particular religion or denomination
2
2
1
NA
1
1
2
1
2
1
1
2
1
1
2
1
2
2

rlgblg_neu Belonging to particular religion or denomination
0
0
1
NA
1
1
2
1
1
0
1
0
1
1
0
0
1
0
0

So überschreiben wir vergebene Ausprägungen bei Vorliegen der Bedingung in der eckigen Klammer überschrieben („Rekodierung“)

ALLE Ausprägungen wurde mit dem Wert „0“ überschrieben. Sinnvoll?

…durch welchen Befehlszusatz kommen wir hier hin?

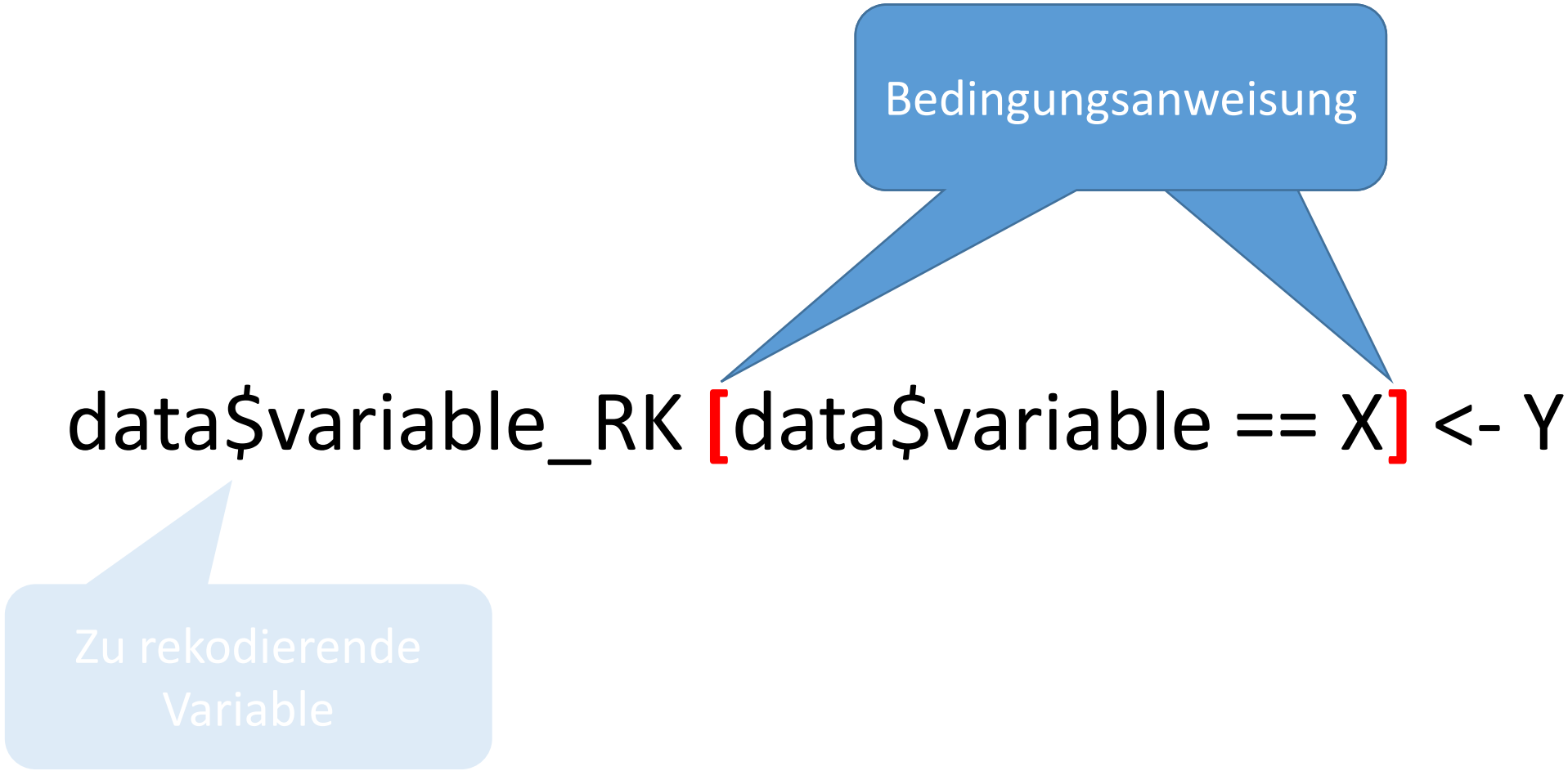
Codestruktur für Rekodierungen:

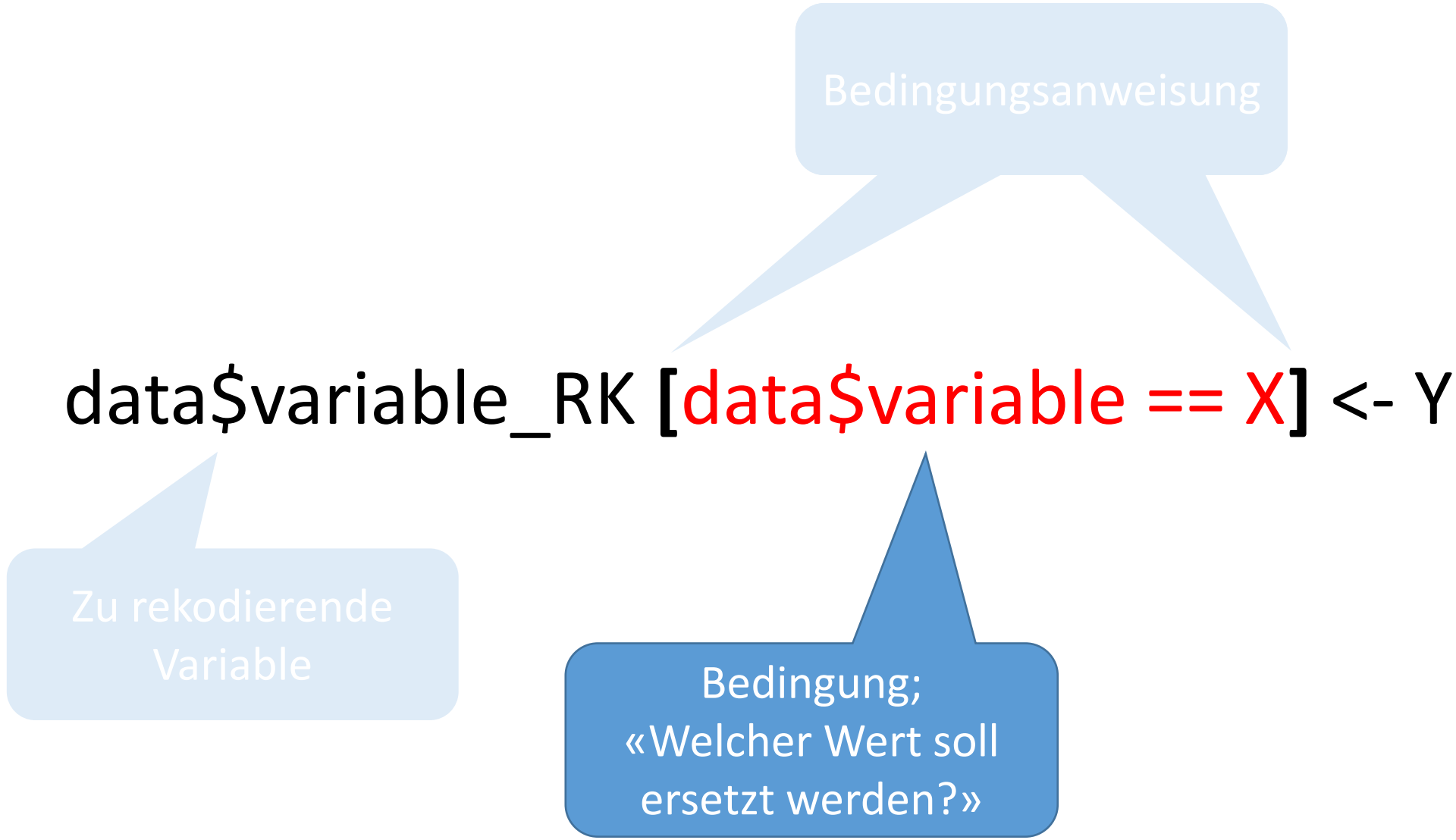
```
data$variable_RK [data$variable == X] <- Y
```

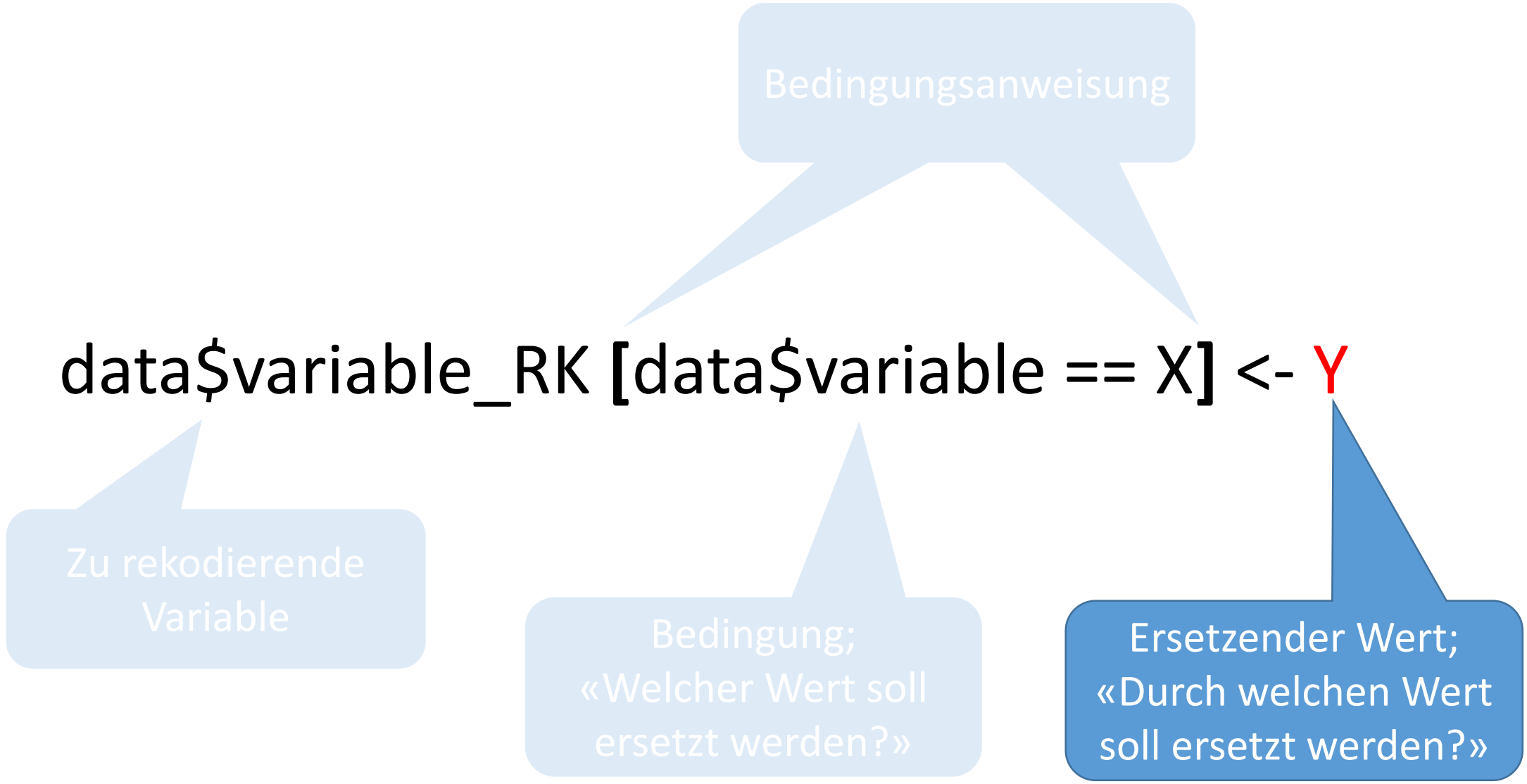
Meistens erstellen wir in der Praxis vorab keine Variablenkopie, sondern legen die neue, recodierte Variable mit der ersten Recodierungsanweisung automatisch an

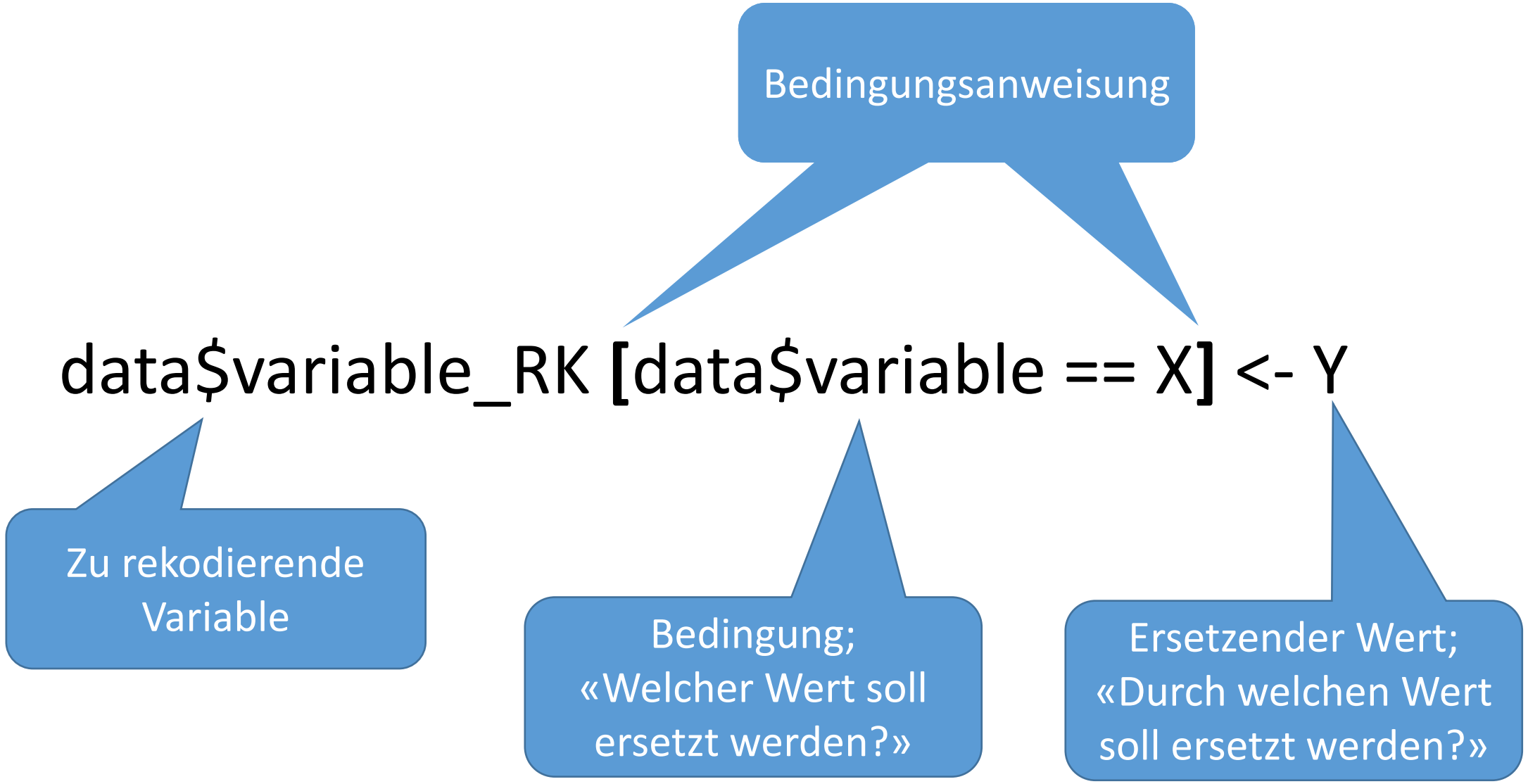
```
data$variable_RK [data$variable == X] <- Y
```

Zu rekodierende, entweder (a) vorab generierte/kopierte oder (b) hier neu angelegte Variable


Bedingungsanweisung
data$variable_RK [data$variable == X] <- Y
Zu rekodierende Variable


Bedingungsanweisung
data$variable_RK [data$variable == X] <- Y
Zu rekodierende Variable
Bedingung; «Welcher Wert soll ersetzt werden?»


Bedingungsanweisung
data$variable_RK [data$variable == X] <- Y
Zu rekodierende Variable
Bedingung; «Welcher Wert soll ersetzt werden?»
Ersetzender Wert; «Durch welchen Wert soll ersetzt werden?»



Frage: In welchen Fällen nehmen wir typischerweise Recodierungen vor?

2.1 Variablenwerte Rekodieren: z.B. Klassifizieren

Wir klassifizieren die metrische Variable **lrscale**.

Frage: Warum bzw. wann könnte dies sinnvoll sein?

Auftrag: Bildet eine Variable mit den Ausprägungen «links», «mitte» und «rechts» auf Basis der Variable «lrscale»

Daten$neue Variable*[*Daten$alte Variable ==] <- *Ausprägung*

!=

<=

>=

<

>

lrscale Placement on left right scale	**lrscale_kat**
0	links
1	links
5	mitte
0	links
5	mitte
5	mitte
4	mitte
5	mitte
5	mitte
5	mitte
5	mitte
8	rechts
8	rechts
5	mitte

Braucht es nicht unbedingt

```
ess8$lrscale_kat<-NA
ess8$lrscale_kat[ess8$lrscale <= 3] <- "links"
ess8$lrscale_kat[ess8$lrscale >= 4 & ess8$lrscale <= 6] <- "mitte"
ess8$lrscale_kat[ess8$lrscale >= 7] <- "rechts"
```

2.1 Variablenwerte Rekodieren: z.B. Klassifizieren

Check ob die Rekodierung erfolgreich war:

(a) Inspektion der (Teil-)Datenmatrix (ggf. per «select»)

(b) Häufigkeitsauszählung/Inspektion alter und neuer Variable

```
> table (ess8$lrscale_kat)

 links  mitte rechts
  8184  20274  10125
> table (ess8$lrscale)
```

Woran erkennst du die erfolgreiche Rekodierung?

```
    0     1     2     3     4     5     6     7     8     9    10
 1463   846  2121  3754  3780 12389  4105  4269  3265   999  1592
```

(c) Kreuztabellierung alter mit neuer Variable

```
table(ess8$lrscale_kat, ess8$lrscale, useNA = "always")

          0     1     2     3     4     5     6     7     8     9    10  <NA>
links  1463   846  2121  3754     0     0     0     0     0     0     0     0
mitte     0     0     0     0  3780 12389  4105     0     0     0     0     0
rechts    0     0     0     0     0     0     0  4269  3265   999  1592     0
<NA>      0     0     0     0     0     0     0     0     0     0     0  5804
```

Woran erkennst du die erfolgreiche Rekodierung?

lrscale Placement on left right scale	**lrscale_kat**
0	links
1	links
5	mitte
0	links
5	mitte
5	mitte
4	mitte
5	mitte
5	mitte
5	mitte
5	mitte
8	rechts
8	rechts
5	mitte

Woran erkennst du die erfolgreiche Rekodierung?

2.2 Variablenwerte Rekodieren: z.B. Umpolung

Aufgabe:

- Was ist **aesfdrk** für ein Variable? (*Konzept, Klasse, Typ, Kategorien, Verteilung*)
- Warum kann hier eine Umkodierung (zwar nicht zwingend, aber) sinnvoll sein?

Tipp: Die Vercodung der Variable entnehmt ihr:

- *attributes()*, oder
- der per *look_for()* generierten Tabelle
- dem Codebuch auf der HP zum ESS
- *table()*, *table(as_factor())*

(Hinweis: as_factor ist eine Variante zu as.factor, bei der die Variablenlabel zu Kategorien werden. Oft sinnvoll)

2.2 Variablenwerte Rekodieren: z.B. Umpolung

Aufgabe:

- Was ist **aesfdrk** für ein Variable? (*Konzept, Klasse, Typ, Kategorien, Verteilung*)
- Warum kann hier eine Umkodierung (zwar nicht zwingend, aber) sinnvoll sein?

```
$label
[1] "Feeling of safety of walking alone in local area after dark"
```

```
$labels
  Very safe        Safe      Unsafe Very unsafe     Refusal  Don't know   No answer
          1           2           3           4          NA          NA          NA
```

- Skizziere kurz die notwendige Rekodierungsprozedur (nicht komplett ausführen, nur durchdenken)
- Finde heraus, wie du der neu kreierten Variable wieder ein Label plus Wertelabel zufügen kannst.

```r
# Umpolung
ess8$aesfdrk_new<-NA
ess8$aesfdrk_new[ess8$aesfdrk == 1] <- 4
ess8$aesfdrk_new[ess8$aesfdrk == 2] <- 3
ess8$aesfdrk_new[ess8$aesfdrk == 3] <- 2
ess8$aesfdrk_new[ess8$aesfdrk == 4] <- 1
```

MA **You**

how can I add a main label and value labels to a numeric variable in R

2.3 Variablenwerte Rekodieren: z.B. Missings

	eisced Highest level of education, ES - ISCED
160	3
161	2
162	3
163	3
164	3
165	55
166	3
167	3
168	5
169	7
170	5
171	2
172	3
173	7
174	3
175	4
176	3
177	3
178	4
179	6
180	5

```
> summary(ess8$eisced)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  1.000   2.000   4.000   4.109   5.000  55.000     129
```

Wo liegt hier das Problem?

2.3 Variablenwerte Rekodieren: z.B. Missings

	eisced Highest level of education, ES - ISCED
160	3
161	2
162	3
163	3
164	3
165	55
166	3
167	3
168	5
169	7
170	5
171	2
172	3
173	7
174	3
175	4
176	3
177	3
178	4
179	6
180	5

eisced Highest level of education, ES - ISCED
3
2
3
3
3
NA
3
3
5
7
5
2
3
7
3
4
3
3
4
6
5

```
> summary(ess8$eisced)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  1.000   2.000   4.000   4.109   5.000  55.000     129
```

Aufgabe: Fehlende Werte als „NA“ kodieren.
Achtung: Bei der Korrektur von Missings erfolgt die Bereinigung typischerweise ohne Anlage einer neuen Variable in der Ausgangsvariable!

```
ess8$eisced[ess8$eisced == 55] <- NA
```

Nun korrekt abgebildet:

```
summary(ess8$eisced)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  1.000   2.000   4.000   4.007   5.000   7.000     217
```

2.4 **Exkurs** Variablenwerte Rekodieren: Variable Werte

Eine Person hat bei der akuten Lebenszufriedenheit keine Angabe gemacht. Wie könnten wird dennoch die akute Lebenszufriedenheit der Person gut approximieren?

Eine Lösung könnte darin bestehen, für diese Person die letztjährige Lebenszufriedenheit zu übernehmen

Wodurch unterscheidet sich die nun geforderte Operation von den bisherigen? Wie setzen wir sie um?

lezufr Lebenszufriedenheit derzeit	**llezufr** Lebenszufriedenheit vor einem Jahr
78	59
81	68
60	75
30	25
16	61
73	53
NA	17
80	75
69	55
70	85

2.4 **Exkurs** Variablenwerte Rekodieren: Variable Werte

```
kursdata_anon$lezufr <- case_when (is.na(kursdata_anon$lezufr) ~ kursdata_anon$llezufr,
                                   !is.na(kursdata_anon$lezufr) ~ kursdata_anon$lezufr)
```

Aufgabe: Übersetzung des Befehls und seiner konkreten Spezifikation in Alltagssprache?

Allen Personen **ohne NA** in **lezufr** wird in der Variable lezufr der Wert aus der Variable **lezufr** zugewiesen.
Allen Personen **mit NA** in lezufr wird in der Variable lezufr der Wert aus der Variable **llezufr** zugewiesen.

lezufr Lebenszufriedenheit derzeit	**llezufr** Lebenszufriedenheit vor einem Jahr
78	59
81	68
60	75
30	25
16	61
73	53
NA	17
80	75
69	55
70	85

lezufr	**llezufr** Lebenszufriedenheit vor einem Jahr
78	59
81	68
60	75
30	25
16	61
73	53
17	17
80	75
69	55
70	85

Bereinigen: Umbenennen

3 Variablen umbenennen

Daten <- **rename**(*Daten*, "*neuer Name*" = "*alter Name*")

Frage: Wann ist es sinnvoll, Variablen umzubenennen?

Die Umbenennung von Variablen ergibt vor allem dann Sinn, wenn die von den Datenprovidern verwendeten Variablenkürzel entweder zu lang oder inhaltlich nicht interpretierbar sind

```
library(dplyr)
ess8 <- rename(ess8, "citizen" = "ctzcntr")
```

Aufgabe: Gebt den Variablen **eiscedf** und **eiscedm** inhaltlich passende Namen.

Tipp: Nutzt wiederum *attributes()*, *Hmisc::contents ()* oder die per *labelled::look_for()* generierte Variablenübersicht (oder orientiert euch am Codebuch zum Datensatz), um herauszufinden, was mit diesen Variablen gemessen wird

Datenauswahl: Selektieren

4 Variablenauswahl mit «select (*daten, var 1, var 2, …*)»

Frage: Wann ist es sinnvoll, den Datensatz auf bestimmte Variablen zu reduzieren?

Aufgabe: erstellet mit **select()** aus dem dplyr Package einen neuen Datensatz, der nur die Variablen **citizen, eduyrs** sowie die ***Identifier*** enthält

```
ess8_ss <- select(ess8, idno, cntry, citizen, eduyrs)
```

Filter

	idno Respondent's identification number	cntry Country	citizen Citizen of country	eduyrs Years of full-time education completed
1	1	AT	2	21
2	2	AT	2	16
3	4	AT	1	13
4	6	AT	1	12
5	10	AT	2	13
6	11	AT	1	13
7	12	AT	1	8
8	13	AT	1	17
9	14	AT	1	14
10	15	AT	1	16
11	16	AT	1	9

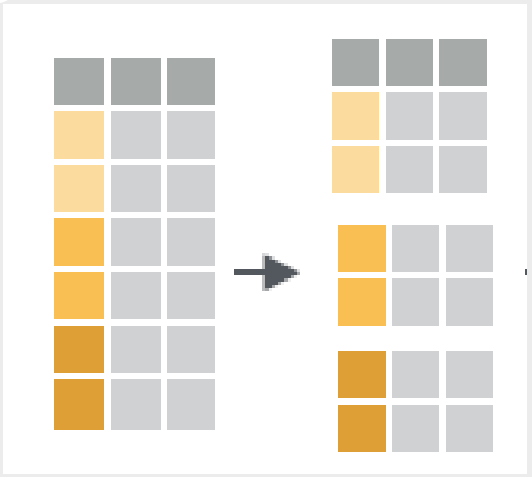
Environment | History | Connections

Import Dataset | Grid

Global Environment

Name	Type	Length	Size	Value
ess8	tbl_df	0	0 B	44387 obs. of 536 variables
ess8_ss	tbl_df	4	1.4 MB	44387 obs. of 4 variables

Datenauswahl: Filtern



5 Fallauswahl mit “filter(*daten*, *bedingung*)”

Aufgabe: Bildet mit **filter()** auf Basis der «citizen»-Variable eine Teilstichprobe von Personen mit Bürgerrecht im Aufenthaltsland

```
citizen <- filter(ess8_ss, citizen == 1)
```

	citizen Citizen of country	eduyrs Years of full-time education completed
row names		21
2	2	16
3	1	13
4	1	12
5	2	13
6	1	13
7	1	8

	citizen Citizen of country	eduyrs Years of full-time education completed
1	1	13
2	1	12
3	1	13

Aufgabe: Bildet nun zusätzlich die “Gegenteilstichprobe” von Personen ohne Bürgerrecht

```
non_citizen <- filter(ess8_ss, citizen == 2)
```

	citizen Citizen of country	eduyrs Years of full-time education completed
row names		21
2	2	16
3	1	13
4	1	12
5	2	13
6	1	13
7	1	8

	citizen Citizen of country	eduyrs Years of full-time education completed
1	2	21
2	2	16
3	2	13

Frage: Wozu könnte das Bilden von Teilstich- und Gegenteilstichproben nützlich sein?

5 *Filtern*: Teilgruppenvergleich

Aufgabe: Vergleicht nun die Bildung (gemessen in Jahren) von Personen mit und ohne Bürgerrecht

```
mean(citizen$eduyrs, na.rm = TRUE)
mean(non_citizen$eduyrs, na.rm = TRUE)
```

```
> mean(citizen$eduyrs, na.rm = TRUE)
[1] 13.02351
> mean(non_citizen$eduyrs, na.rm = TRUE)
[1] 13.27332
```

Fazit?

Hausaufgabe mit Selbstüberprüfung:

https://www.suz.uzh.ch/dataforstat/statistik2/dataueb.html