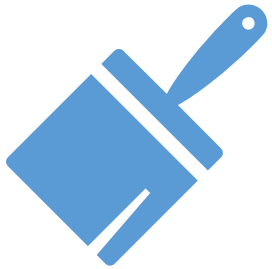
Statistik 1 – Tutorate
Sitzung 4: Datenmanagement in R

Marco Giesselmann, Lea Elina Hofer, Norma De Min, Mara Moos, Rémy Blum

Lernziele dieser Sitzung



Datenbereinigung

Löschen irrelevanter Objekte

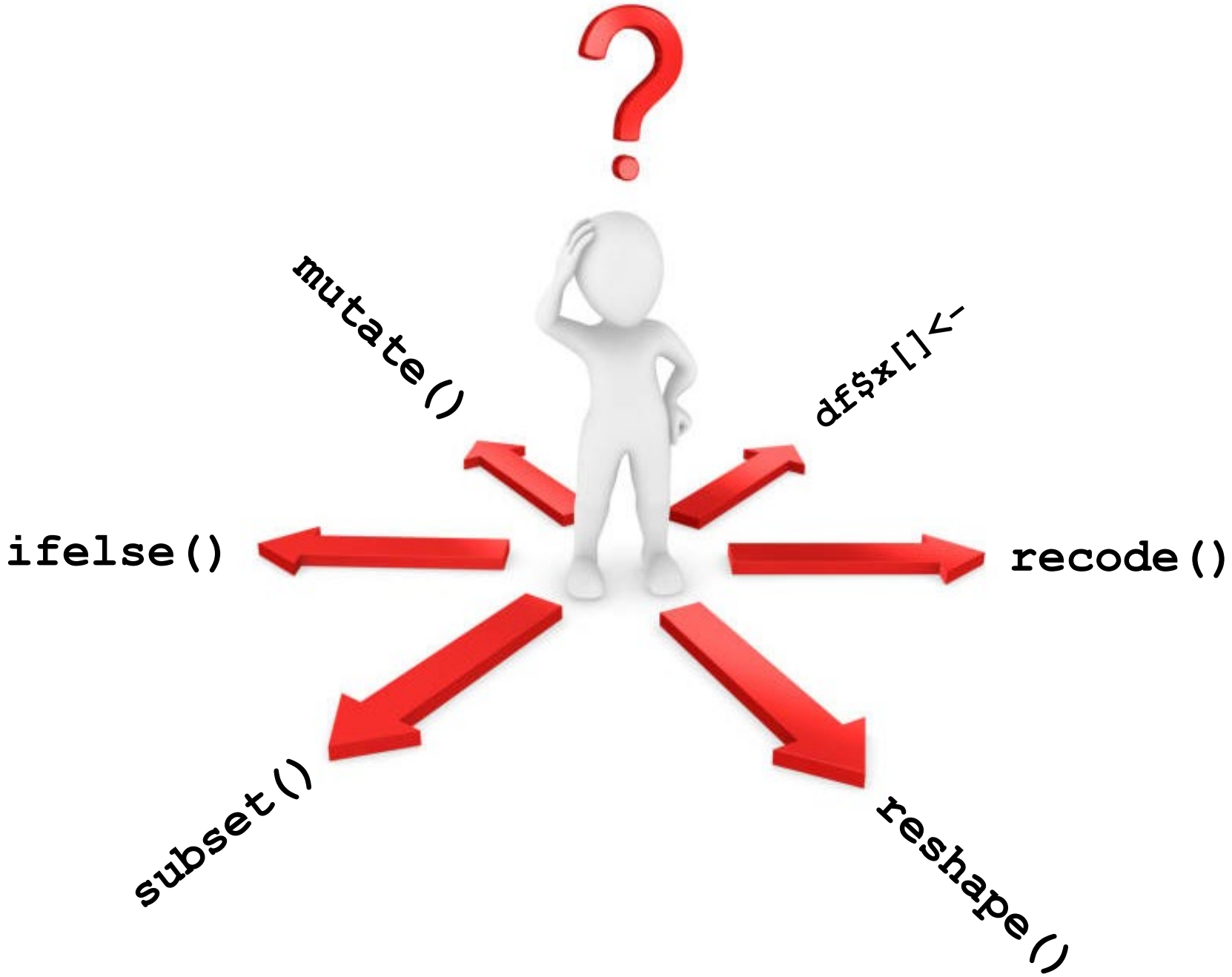
Variablen rekodieren

Variablen umbenennen

Datenauswahl

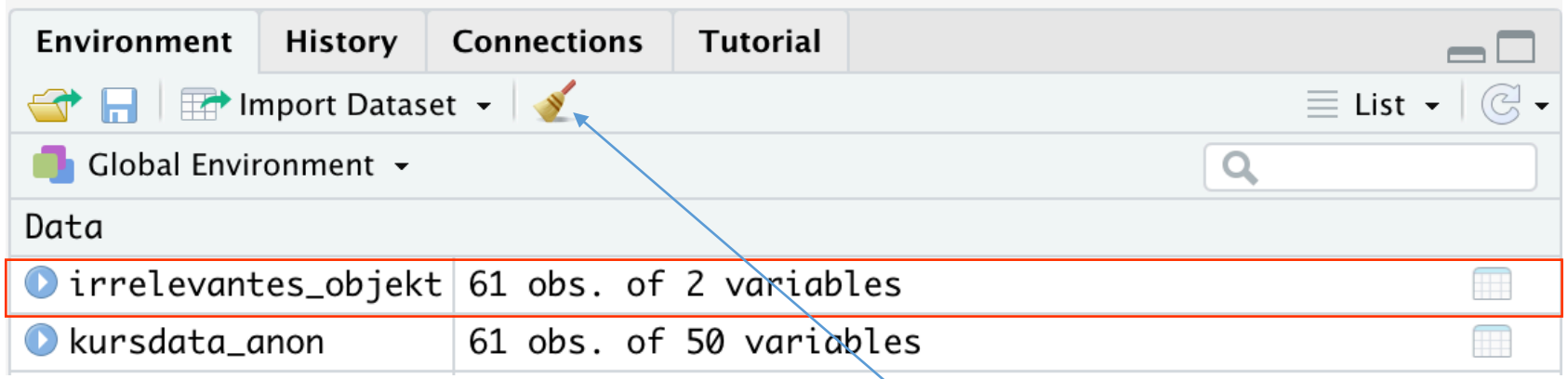
Selektieren: Teildatensätze bilden

Filtern: Teilstichproben bilden


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

Bereinigen: Aufräumen

1 Löschen irrelevanter Objekte



remove() löscht ein einzelnes Objekt

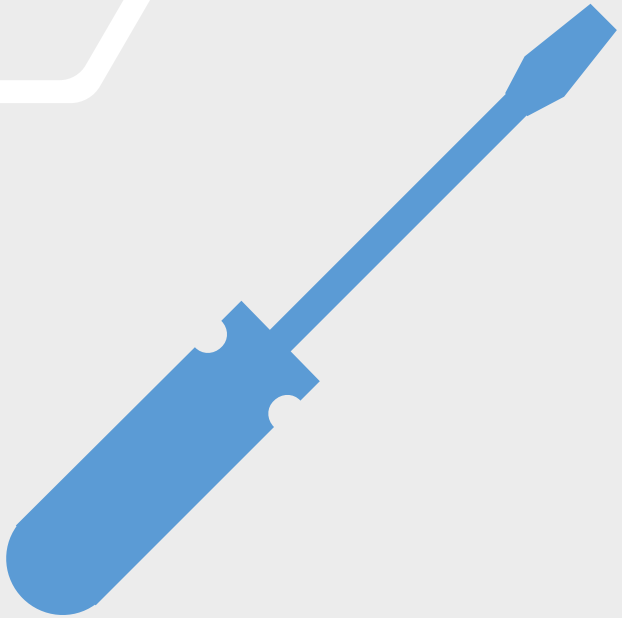
```
remove(irrelevantes_objekt)
```

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 **fach** rekodieren (2 zu 0) dazu erstellen wir die Variable **fach_neu**

fach studiertes hauptfach
1
2
2
2
1
2
1
1
1
1

```
kursdata_anon$fach_neu <- kursdata_anon$fach
```
-> Neue Variable wird gebildet und mit den Werten der alten angereichert

fach_neu studiertes hauptfach
1
2
2
2
1
2
1
1
1
1

```
kursdata_anon$fach_neu <- 0
```
-> Die neue Variable wird überschrieben (alle Ausprägungen)

fach_neu
0
0
0
0
0
0
0
0
0
0

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

```
kursdata_anon$fach_neu <- kursdata_anon$fach
```
-> Neustart: Zum Schritt 1

fach_neu studiertes hauptfach
1
2
2
2
1
2
1
1
1
1

…durch welchen Befehlszusatz kommen wir hier hin?

```
kursdata_anon$fach_neu[kursdata_anon$fach == 2] <- 0
```

fach_neu studiertes hauptfach
1
0
0
0
1
0
1
1
1
1

So soll die recodierte Variable aussehen…

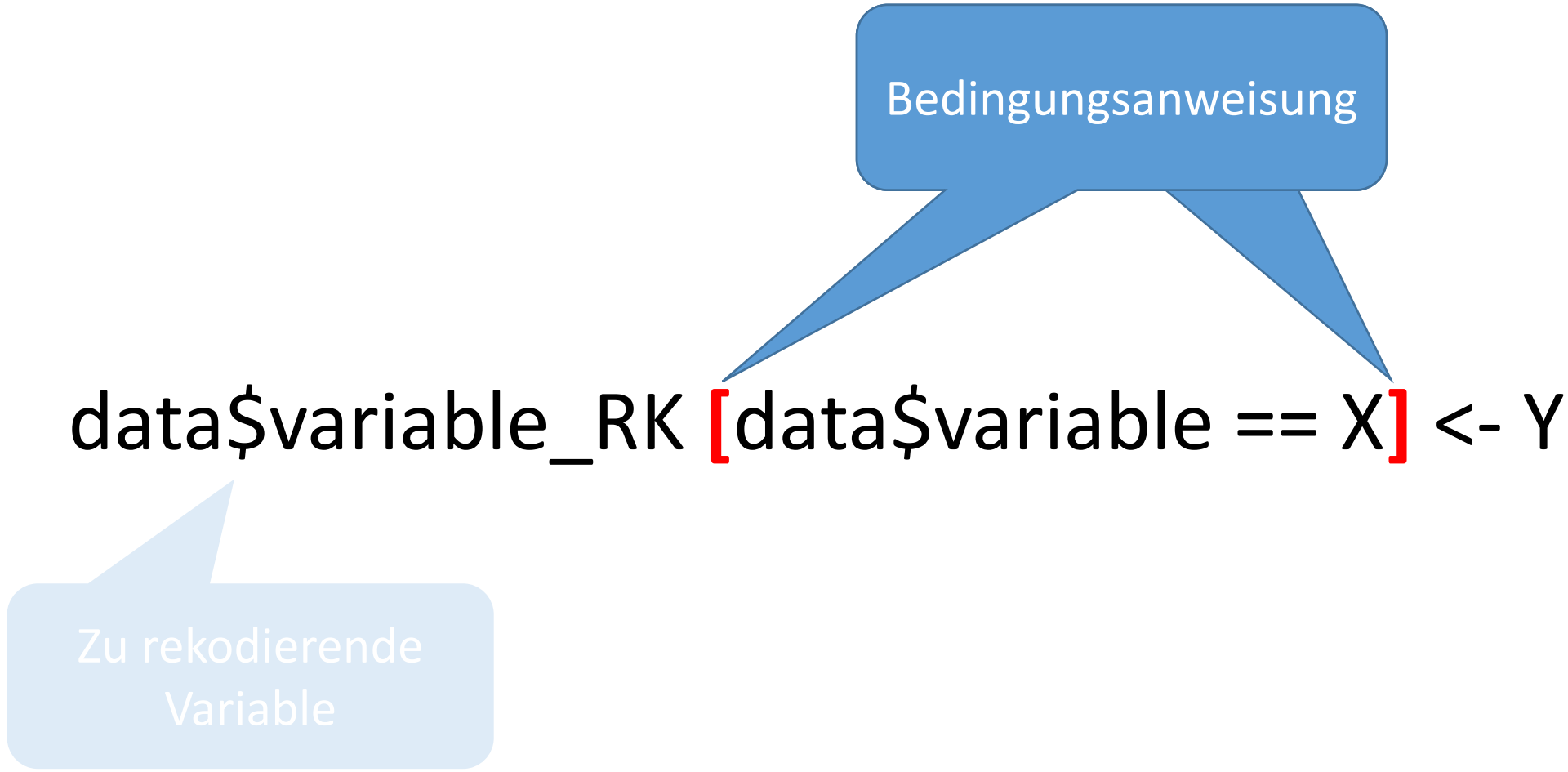
Codestruktur für Rekodierungen:

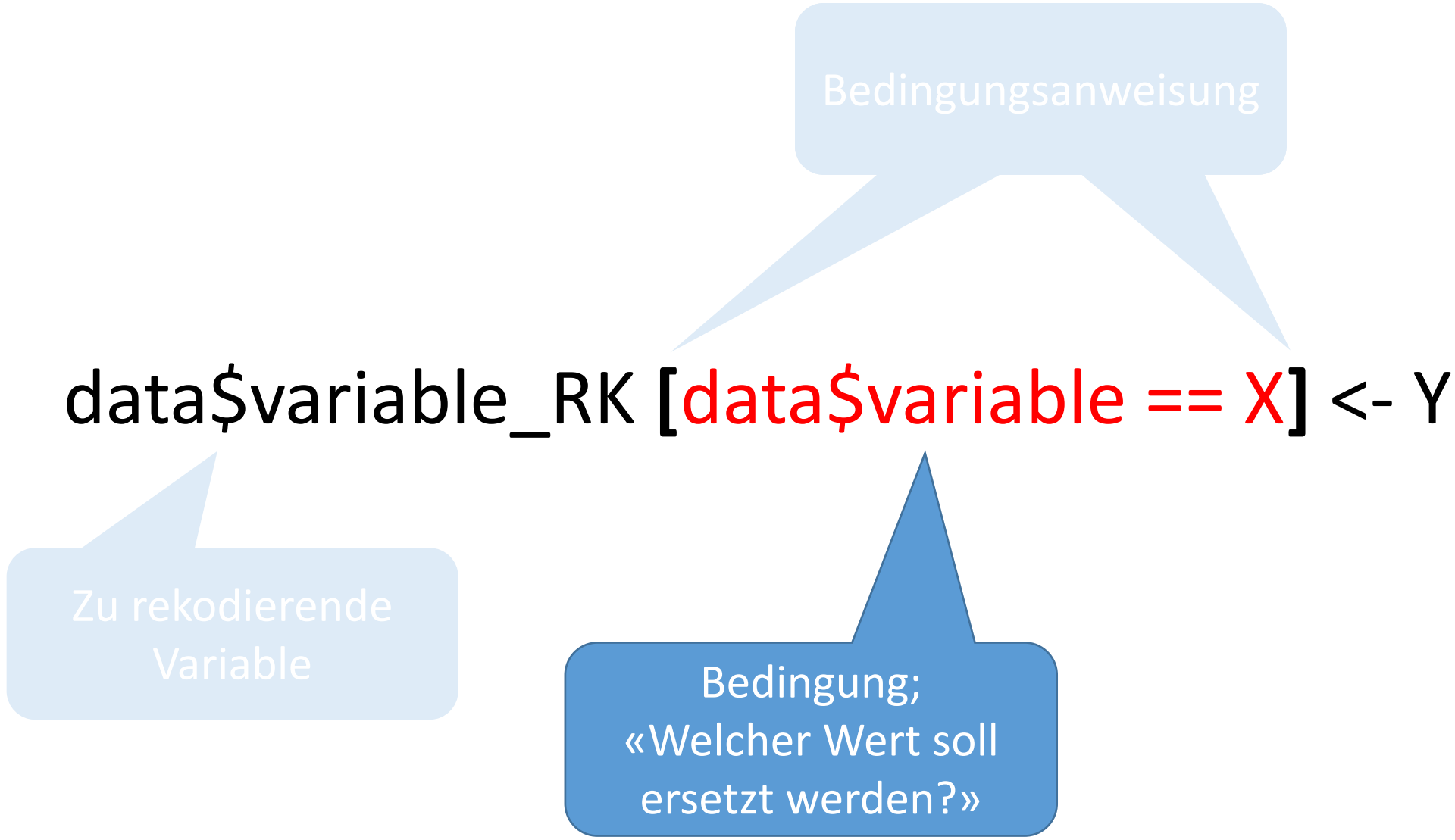
```
data$variable_RK [data$variable == X] <- Y
```

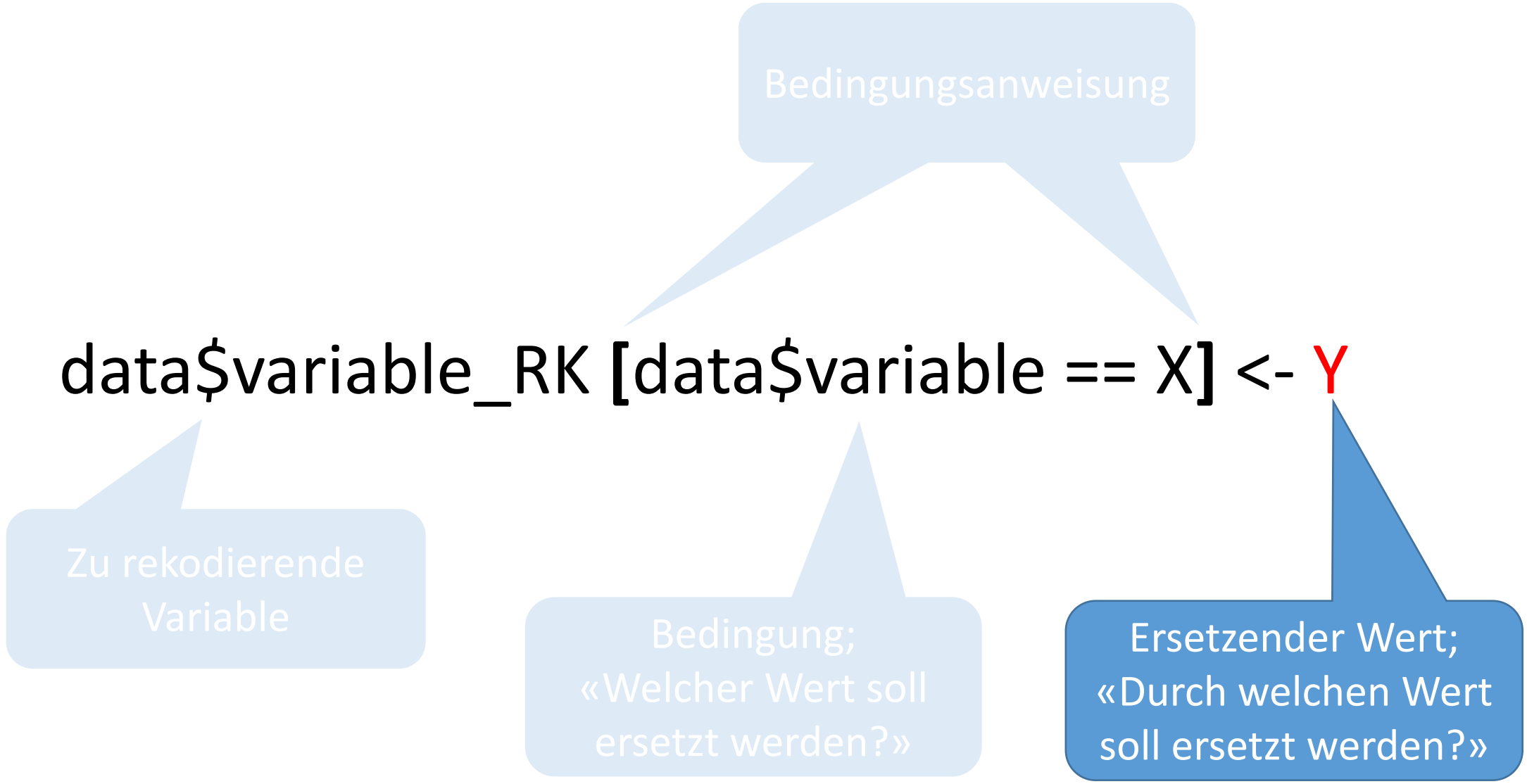
Meistens erstellen wir in der Praxis vorab keine Variablenkopie, sondern legen die neue, recodierte Variable mit der ersten Recodierungsanweisung automatisch an

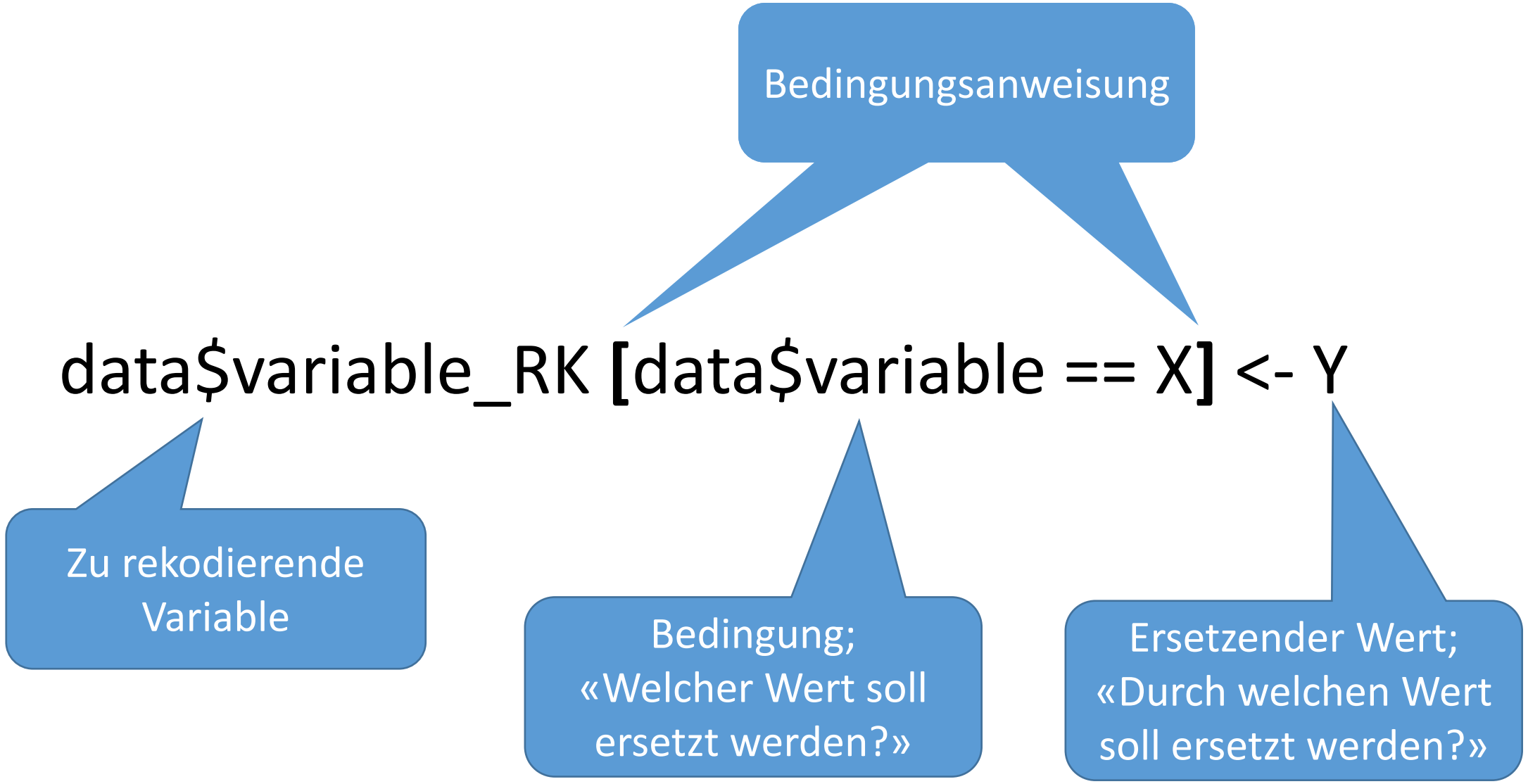
```
data$variable_RK [data$variable == X] <- Y
```

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 - Typische Anwendung: **Umpolung**

Aufgabe: Finde heraus, wie die Variable **intmig** codiert ist und, ob die Reihenfolge der Werte sinnvoll ist.

```
> attributes(kursdata_anon$intmig)
$label
[1] "interesse: Migration und Integration"

$format.stata
[1] "%9.0g"

$class
[1] "haven_labelled" "vctrs_vctr"     "double"

$labels
     sehr      etwas gar nicht
        1          2          3
```

Momentan drückt der tiefste Wert das grösste Interesse am Thema „Migration“ aus.
Dies ist nicht sehr intuitiv, nimm daher eine Rekodierung bzw. konkret: *Umpolung* vor.

intmig interesse: Migration und Integration	intmig_neu
1	3
2	2
2	2
2	2
2	2
3	1
1	3
2	2
1	3
1	3

Achtung: Die Label von Variable & Werten gehen bei solchen kopiebasierten Rekodierungen verloren, manchmal werden sie sogar falsch übertragen. Ihr solltet die neue Bedeutung der Variablenwerte mindestens im Skript dokumentieren, besser aber noch die rekodierte Variable neu labeln (siehe HP).

```
kursdata_anon$intmig_neu[kursdata_anon$intmig == 1] <- 3
kursdata_anon$intmig_neu[kursdata_anon$intmig == 2] <- 2
kursdata_anon$intmig_neu[kursdata_anon$intmig == 3] <- 1
```

2.2 Variablenwerte Rekodieren - Typische Anwendung: **Klassifizieren**

Wir wollen nun die metrische Variable **leftright** klassifizieren**.**

Warum könnte dies sinnvoll sein? Was könnte ein Nachteil sein?

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

Vorab: Alternative Bedingungsoperatoren

- != ungleich
- < kleiner
- > grösser
- <= kleiner/gleich
- >= grösser/gleich
- & beide Bedingungen korrekt
- | eine Bedingung korrekt

Überlege dir eine sinnvolle Klassifikationsstrategie:
Welche Klassen sind sinnvoll?
Wie kann ich die Klassifikation mit den Operatoren vornehmen?

leftright einordnung auf links_rechts skala	**leftright_kat**
15	links
34	mitte
10	links
45	mitte
90	rechts
18	links
50	mitte
40	mitte
50	mitte
5	links
0	links
50	mitte
36	mitte

```
kursdata_anon$leftright_kat[kursdata_anon$leftright <= 33] <- "links"
kursdata_anon$leftright_kat[kursdata_anon$leftright >= 34 & kursdata_anon$leftright <= 66] <- "mitte"
kursdata_anon$leftright_kat[kursdata_anon$leftright >= 67] <- "rechts"
```

2.3 Variablenwerte Rekodieren - Typische Anwendung: **Zusammenfassen**

Aufgabe: Fasst die Variable **rauchen** so zusammen, dass lediglich unterschieden wird, ob jemand jemals geraucht hat (Smoker) oder noch nie geraucht hat (Nonsmoker).

```
> attributes(kursdata_anon$rauchen)
$label
[1] "letzte Woche geraucht?"

$format.stata
[1] "%18.0g"

$class
[1] "haven_labelled" "vctrs_vctr"     "double"

$labels
                ja nein, aber früher    nein, noch nie
                 1                  2                 3
```

rauchen letzte Woche geraucht?	**rauchen_binary**
3	NonSmoker
2	Smoker
1	Smoker
3	NonSmoker
1	Smoker
3	NonSmoker
1	Smoker
1	Smoker
2	Smoker

```
kursdata_anon$rauchen_binary[kursdata_anon$rauchen <= 2] <- "Smoker"
kursdata_anon$rauchen_binary[kursdata_anon$rauchen == 3] <- "NonSmoker"
```

2.4 Variablenwerte Rekodieren - Typische Anwendung: **Missings korrigieren**

lezufr Lebenszufriedenheit derzeit
78
81
60
30
16
73
-99
80
69
70

lezufr Lebenszufriedenheit derzeit
78
81
60
30
16
73
→ *NA*
80
69
70

Untersucht die Variable **lezufr** mit dem Befehl **summary().**
Was fällt euch auf?

```
> summary(kursdata_anon$lezufr)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-99.00   60.00   72.00   66.11   80.00  100.00
```

Fehlende Werte sind hier mit Zahlenwert „-99“ abgebildet. Problem!
Diese sollten in ein „NA“ umgewandelt werden.

```
kursdata_anon$lezufr[kursdata_anon$lezufr == -99] <- NA
```

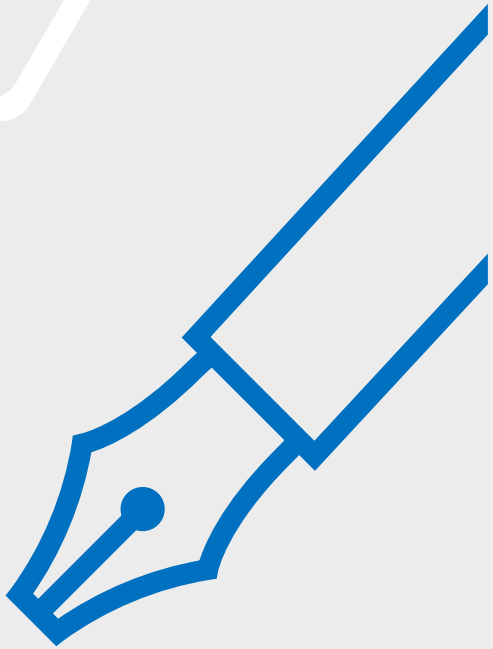
```
> summary(kursdata_anon$lezufr)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   60.00   72.50   68.34   80.00  100.00       1
```

Wieso ist der Mean stärker angestiegen als der Median?

Datenrekodierung im ESS

1. Lade den ESS 8 Datensatz (siehe letzte Sitzung oder HP)
2. Recherchiere den Begriff ES-ISCED. Worum handelt es sich?
3. Bei der Variable **eisced** gibt es eine Ausprägung, die als Missing verstanden werden könnte. Finde zuerst mit **attributes()** und **summary()** heraus, um welche Ausprägung es sich handelt und codiere diese anschliessend in das für R verständliche Missing-Format.
4. Finde mit **attributes()**, **look_for()** oder dem **Codebook** heraus, um was für eine Variable es sich bei **clmchng** handelt und wie diese codiert ist.
5. Wann wäre es sinnvoll, die Variable **clmchng** (Einstellung zum Klimawandel) zu rekodieren?
6. Generiere nun, basierend auf Aufgabe 5, eine dichotome Variable names **clmchng_d**, die definitive Leugner vom Rest abgrenzt - die also lediglich unterscheidet, ob es sich um einen kategorischen Leugner handelt oder nicht.

Bereinigen: Umbenennen



3 Variablen umbenennen

Frage: Wann ist es sinnvoll, Variablen umzubenennen?

```
library(dplyr)
kursdata_anon <- rename(kursdata_anon, "soz_hf" = "fach_neu")
```

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

fach_neu studiertes hauptfach
1
0
0
0
1
0
1
1
1
1

soz_hf studiertes hauptfach
1
0
0
0
1
0
1
1
1
1

Hier macht die Umbenennung der Variable Sinn, da nun auf einen Blick erkenntlich ist, dass die Variable **soz_hf** abfragt, ob das Hauptfach Soziologie ist oder ein anderes Fach.

Insbesondere beim ESS-Datensatz kann der **rename()**-Befehl zudem nützlich sein, um Variablennamen abzukürzen oder deren Verständlichkeit zu erhöhen.

Datenauswahl: Variablen Selektieren

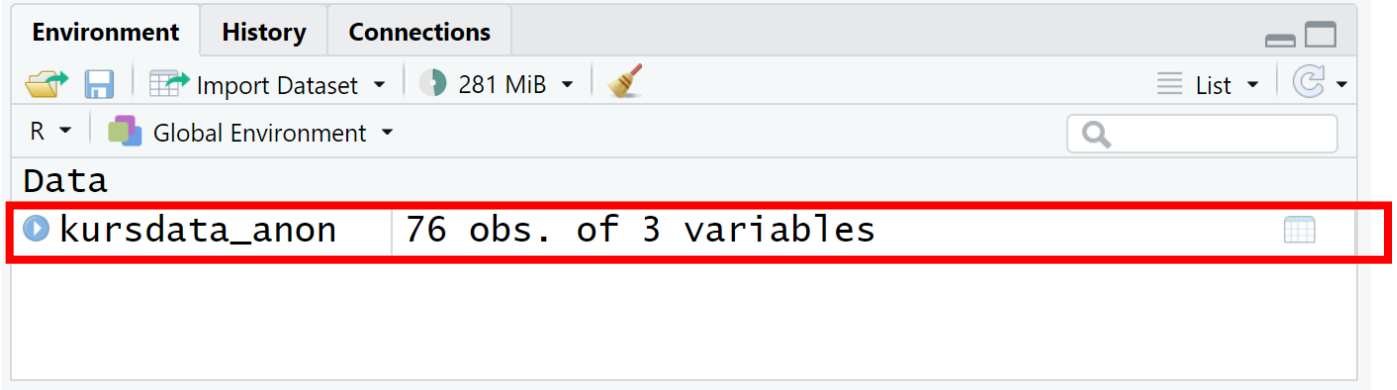
4 Datenauswahl: ***Variablen*** *selektieren*

Fragen: - Warum ist es meist sinnvoll, den Datensatz auf bestimmte Variablen zu reduzieren?
- Recherhiert den (wichtigsten) Befehl zur Variablenauswahl in R.
- Zu welchem Package gehört er?

Erstellt mithilfe des **select()** Befehls aus dem dplyr Package einen reduzierten Datensatz, der nur die Variablen **rauchen_binary** und **lezufr** enthält.

```
kursdata_anon <- select(kursdata_anon, id, rauchen_binary, lezufr)
```

	id	rauchen_binary	lezufr Lebenszufriedenheit derzeit
1	28	NonSmoker	68
2	41	Smoker	74
3	31	Smoker	58
4	34	NonSmoker	65
5	48	Smoker	90
6	50	NonSmoker	46
7	60	Smoker	*NA*
8	54	Smoker	65
9	69	Smoker	76
10	63	Smoker	75



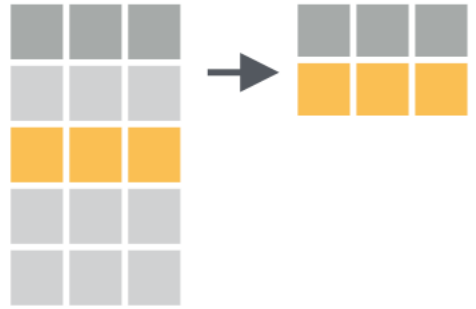
Die "Select" - Prozedur wird zumeist angewendet, um den Ausgangsdatensatz auf die analyserelevanten Variablen zu begrenzen. Da die übrigen, ausgeschlossen Variablen dann nicht mehr gebraucht werden, wird bei Anwendung von "select" meist kein neuer Datensatz angelegt, sondern einfach der Ausgangsdatensatz reduziert

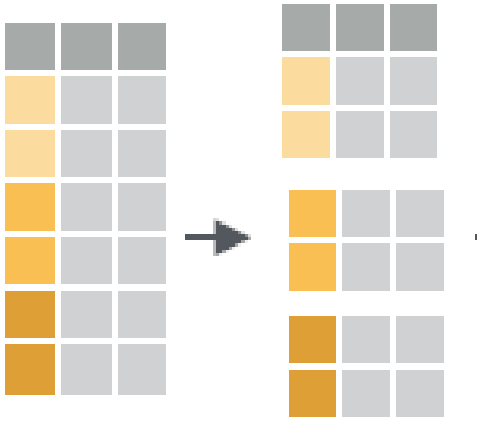
Datenauswahl: Fälle Filtern

5.1 Datenauswahl: ***Fälle*** *filtern*

Fragen:

- Warum ist es manchmal ist es sinnvoll, Analysen auf bestimmte Merkmalsträger/ Fälle zu beschränken?
- Recherhiert den (wichtigsten) Befehl zur Fallauswahl in R.
- Zu welchem Package gehört er?





5.1 Datenauswahl: ***Fälle** filtern*

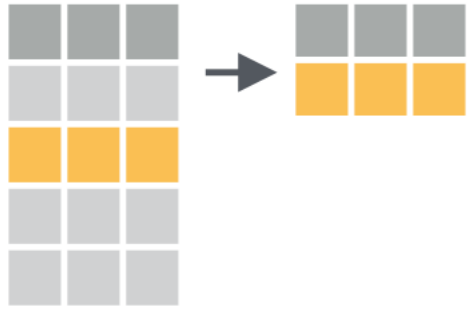
Bilde in einem neuen Datensatz eine Teilstichprobe von Personen, die rauchen.

```
smokers <- filter(kursdata_anon, rauchen_binary == "Smoker"
```

id	rauchen_binary	lezufr Lebenszufriedenheit derzeit
28	NonSmoker	68
41	Smoker	74
31	Smoker	58
34	NonSmoker	65
48	Smoker	90

→

id	rauchen_binary	lezufr Lebenszufriedenheit derzeit
41	Smoker	74
31	Smoker	58
48	Smoker	90



Bilde nun zusätzlich die “Gegenteilstichprobe” von Personen, die nicht rauchen.

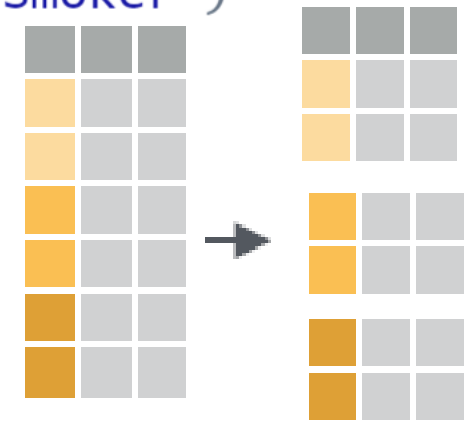
```
no_smokers <- filter(kursdata_anon, rauchen_binary == "NonSmoker")
```

«smokers»

id	rauchen_binary	lezufr Lebenszufriedenheit derzeit
41	Smoker	74
31	Smoker	58
48	Smoker	90

«no_smokers»

id	rauchen_binary	lezufr Lebenszufriedenheit derzeit
28	NonSmoker	68
34	NonSmoker	65



5.1 Datenauswahl: ***Fälle** filtern*

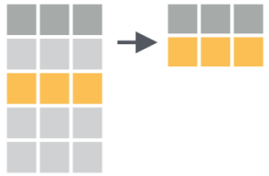
Bilde in einem neuen Datensatz eine Teilstichprobe von Personen, die rauchen.

```
smokers <- filter(kursdata_anon, rauchen_binary == "Smoker"
```

id	rauchen_binary	lezufr Lebenszufriedenheit derzeit
28	NonSmoker	68
41	Smoker	74
31	Smoker	58
34	NonSmoker	65
48	Smoker	90

→

id	rauchen_binary	lezufr Lebenszufriedenheit derzeit
41	Smoker	74
31	Smoker	58
48	Smoker	90



Bilde nun zusätzlich die "Gegenteilstichprobe" von Personen, die nicht rauchen.

```
no_smokers <- filter(kursdata_anon, rauchen_binary == "NonSmoker")
```

«smokers»

id	rauchen_binary	lezufr Lebenszufriedenheit derzeit
41	Smoker	74
31	Smoker	58
48	Smoker	90

«no_smokers»

id	rauchen_binary	lezufr Lebenszufriedenheit derzeit
28	NonSmoker	68
34	NonSmoker	65

25

Frage: Warum ist es bei der «Filter»-Prozedur, anders als bei «select», meistens sinnvoll, neue Datensätze anzulegen, statt den Ausgangsdatensatz zu überschreiben?

Antwort: Der Ausgangsdatensatz wird für die Operationen nach der «Filter»-Prozedur meist noch gebraucht (wie auch hier im konkreten Beispiel: Die zweite Filterung oben bedingt den Ausgangsdatensatz)

5.1 Datenauswahl: ***Fälle** filtern*

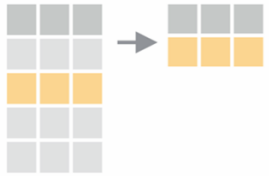
Bilde in einem neuen Datensatz eine Teilstichprobe von Personen, die rauchen.

```
smokers <- filter(kursdata_anon, rauchen_binary == "Smoker"
```

id	rauchen_binary	lezufr Lebenszufriedenheit derzeit
28	NonSmoker	68
41	Smoker	74
31	Smoker	58
34	NonSmoker	65
48	Smoker	90

→

id	rauchen_binary	lezufr Lebenszufriedenheit derzeit
41	Smoker	74
31	Smoker	58
48	Smoker	90



Bilde nun zusätzlich die "Gegenteilstichprobe" von Personen, die nicht rauchen.

```
no_smokers <- filter(kursdata_anon, rauchen_binary == "NonSmoker")
```

Piktogramme kommen von CHEATSHEET…

Data transformation with dplyr : : CHEATSHEET

dplyr functions work with pipes and expect **tidy data**. In tidy data:

Each **variable** is in its own **column**

&

Each **observation**, or **case**, is in its own **row**

pipes

x |> f(y) becomes **f(x, y)**

Summarize Cases

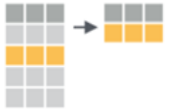
Apply **summary functions** to columns to create a new table of summary statistics. Summary functions take vectors as input and return one value (see back).

summary function

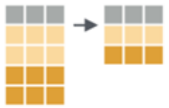
Manipulate Cases

EXTRACT CASES

Row functions return a subset of rows as a new table.



filter(.data, …, .preserve = FALSE) Extract rows that meet logical criteria.
mtcars |> filter(mpg > 20)



distinct(.data, …, .keep_all = FALSE) Remove rows with duplicate values.
mtcars |> distinct(gear)

slice(.data, …, .preserve = FALSE) Select rows by position.
mtcars |> slice(10:15)

Manipulate Variables

EXTRACT VARIABLES

Column functions return a set of columns as a new vector or table.

pull(.data, var = -1, name = NULL, …) Extract column values as a vector, by name or index.
mtcars |> pull(wt)

select(.data, …) Extract columns as a table.
mtcars |> select(mpg, wt)

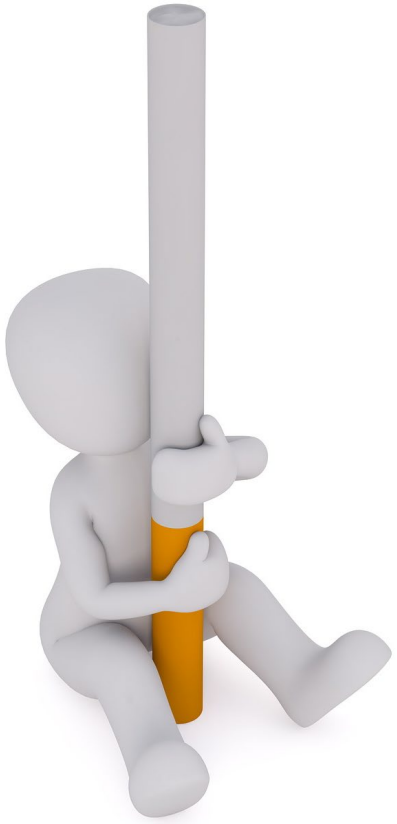
relocate(.data, …, .before = NULL, .after = NULL) Move columns to new position.
mtcars |> relocate(mpg, cyl, .after = last_col())

5.2 *Variablen Filtern*: z.B. für Teilgruppenvergleich

RaucherInnen und NichtraucherInnen im Kurs – wer ist lebenszufriedener?

```
mean(smokers$lezufr, na.rm = TRUE)
mean(no_smokers$lezufr, na.rm = TRUE)
```

```
> mean(smokers$lezufr, na.rm = TRUE)
[1] 69.57895
> mean(no_smokers$lezufr, na.rm = TRUE)
[1] 65.81081
```



Fazit?

Filten und Selektieren mit dem ESS

Aufgabe 1 (Alle):

- Suche im ESS die Variable heraus, die
  (a) sich auf das Land (bzw. das Ländersample) bezieht
  (b) die emotionale Bindung der Befragten zu diesem Land misst
  (c) mit dem Namen ***brncntr*** bezeichnet ist
- Beschränke den Datensatz auf diese drei Variablen. Was messen sie?
- Generiere zwei separate Teildatensätze (‘ess_ch’ und ‘ess_de’), welche jeweils ausschliesslich Befragte aus der Schweiz bzw. aus Deutschland enthalten.

Aufgabe 2 (Anfangsbuchstaben A-L):

- In welchem der beiden Länder ist die emotionale Bindung an das Land grösser?
- Binde als dritte Vergleichgruppe Personen auf den skandinavischen Staaten (Finnland, Norwegen, Schweden) mit in den Vergleich ein.

Aufgabe 3 (Anfangsbuchstaben M-Z):

- Wie unterscheide sich die emotionale Bindung an die Schweiz zwischen in der Schweiz geborenen und zugezogenen Personen?
- Ist der emotionale Bindungsunterschied an das Wohnland zwischen im dort Geborenen und Zugezogenen in Deutschland grosser als in der Schweiz?

Weitere Aufgaben…

- **Übung 2** auf http://www.suz.uzh.ch/dataforstat/ zu lösen.