

R Lesson 7: Data Wrangling (practice)

vanderbi.lt/r

Steve Baskauf



Download ICPSR data

- Instructions link on the lessons homepage (vanderbi.lt/r)
- Create ICPSR account
- Download 2 files for National Longitudinal Study of Adolescent to Adult Health, 1994-2008 (**21600-0001-Data.tsv** and **21600-0022-Data.tsv**)

Issues with the NLSAAH dataset

- **National Longitudinal Study of Adolescent to Adult Health, 1994-2008** = big longitudinal study
- Datasets are huge and difficult to work with, so need to **extract** subset of data
- Data are coded using numbers – need to **transform** to other forms.
- Data are in separate CSV files that need to be **joined**
- We might want to **create new data fields by calculation** from others.

Assignment #1: Extract data

- We want data on age, sex, and parental relationships (from DS0001) and on smoking and body characteristics (from DS0022). See top of script.
- The AID unique identifier is the key to join the two tables.
- Want to save extracted data in file so that we don't need to load the big datasets into memory.

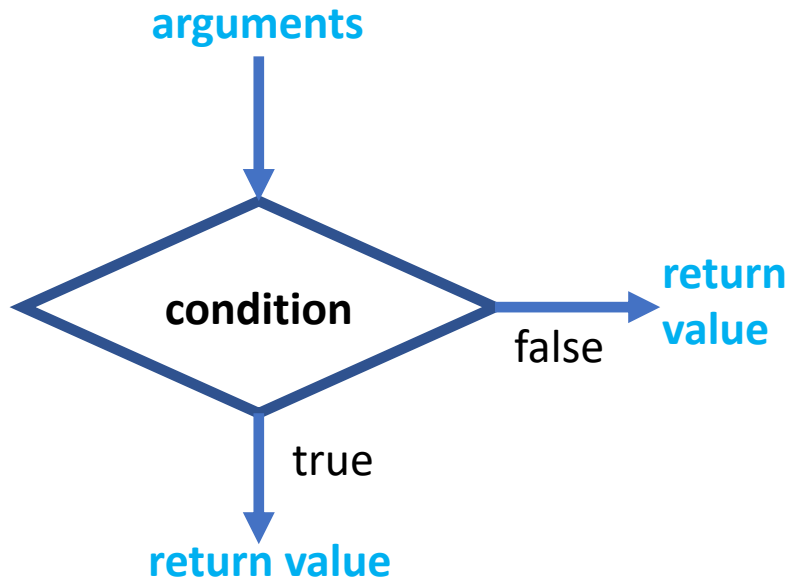
Assignment #2 Calculate BMI

- We want to know BMI to see if it's related to smoking and other factors.
- Need to replace codes for missing values to NA (documentation gives codes for each question)
- Also recode "1" and "2" for sex to "male" and "female" to make interpretation clearer.
- Calculate age
- Convert height and weight to SI (metric)
- Calculate BMI using $\text{mass_kg}/\text{height_m}^2$
- Need to use piping because of so many operations

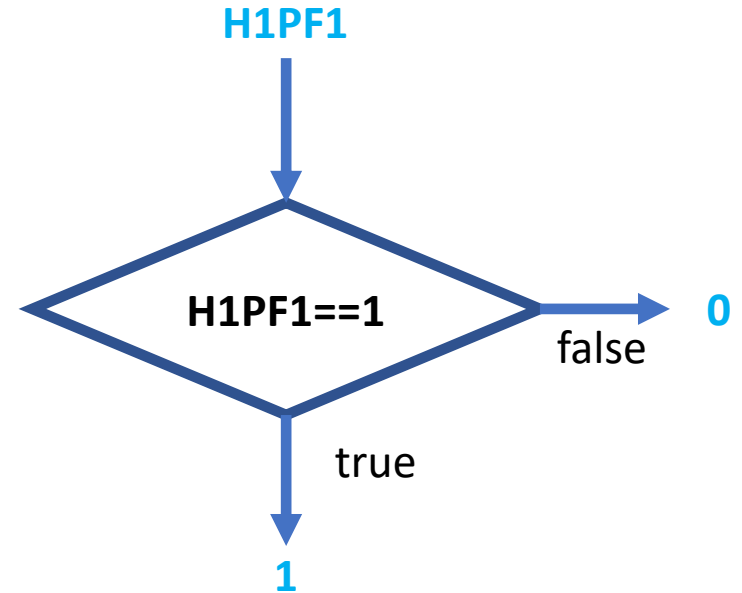
Assignment #3 Calculate "maternal closeness"

- For each maternal relationship question recode 1 to 1 (good) and other values to 0 (1 is most common value).
- Complication: NA values must be preserved during calculation.
- Maternal closeness is 1 for people who have 1 values for every maternal relationship question.
- Need to have an "if" function for this (`ifelse`).

ifelse () function



general pattern



specific example

```
ifelse (H1PF1==1 , 1 , 0)
```

boolean operators for conditions

! is **NOT**

& is **AND**

| is **OR**

- **Examples:**

`!H1PF5==1` H1PF5 isn't equal to 1

`!is.na(H1PF5)` H1PF5 doesn't have an NA value

`H1PF3==1 & H1PF4==1` both H1PF3 and H1PF4 equal 1

`is.na(H1PF2) | is.na(H1PF3)` either H1PF2 or H1PF3 is NA