Introduction to R- Day 2

Alessandra Meddis

Exercise: cleaning data

We consider extract data from Sundby95 survey carried out in Copenhagen 1995 to assess general health of people. You can find the data in: "https://raw.githubusercontent.com/AMeddis/IntrotoR-for-Basic-Statistics/refs/heads/main/data_exercise/sundby_clean.csv"

The data include the variables:

- kon: sex (1/2)
- v75: weight (kg)
- v76: height (cm)
- v17: physical activity (categories 1-4 with 1: most activity)
- v24af: alcohol intake during the last week.

Exercise 0:

Read the data into R and get a summary of its contents.

Exercise 1:

1. Change the name of the columns (in English) so to be easier to understand what they refer to.
2. Use the *str()* function on the data. Check the types of the variables, is there something unexpected?

Exercise 2:

Weight should be a numerical variable, but it has been imported as character. Can you guess why?

1. Add one new variable *weight_num* to the data set, that is the numerical version of weight. You can use *as.numeric()*. What is the warning message suggesting?

   ```
   NOTE that COERCION refers to the process of converting an object from one data type to another.
   The message mentions that NAs (missing values) are introduced
   when converting from character to numeric.
   ```

2. We search for the *NAs introcued by Coercion*

   2a. Run the command: table(data_raw$weight==" "), knowing that missing values for weight (character variable) are indicated with an empty space " ". What is this indicating?

   2b. Run the command: table(is.na(data_raw$weight_num)), knowing that is.na() gives TRUE if the element of the vector is an NA, FALSE otherwise. What is this indicating?

3. So far, we noticed that when converting*weight* into a numeric, 2 missing values have been introduced (by coercion). We would like to identify those values. We can use the function *which*, that provides the index (position) of the elements where the specified condition is TRUE.

   ```
   Run the command: which(data_raw$weight!="" & is.na(data_raw$weight_num)).
   ```

What are these numbers? **Note**: & is the logical operator AND (TRUE where both conditions are fulfilled)

4. We know now which elements we have to modify, you can access to the *weight* and correct manually the weights values. Remember that you can use [ ] to access objects in `R`.
5. Re-create then the numerical version of weight.

Exercise 3

Sex (kon) is a numerical variable, with values 1 and 2. We believe that 1 represents males and 2 females (default). However, we would like to check whether this assumption is correct.

1. We have information on height, calculate the median height by sex, which group is taller? Is the assumption correct?
2. Transform sex (kon) into a factor and assign more appropriate labels.
3. Is there another variable you would like to have as a factor? If yes, transform it.

Exercise 4

We want to calculate the BMI for each patient. The formula is : weight(kg) / height(m) ^2, but we have the height in centimeters.

1. Create a new variable *height_m*, which is the height in meters.
2. Create a new variable for *BMI*.