Exercises for Recap Session 1

Possible solutions

2024-04-18

Exercise 1: Basic object types

Consider the following vector:

```
ex_1_vec <- c(1.9, "2", FALSE)
```

1. What is the type of this vector? Why?

```
typeof(ex_1_vec)
```

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

Atomic vectors only contain objects of the same type, and there is a hierarchy. Elements that themselves are of a type lower in the hierarchy are coerced to the same type as the object highest in the hierarchy. The hierarchy is as as follows:

1. `character`
2. `double`
3. `integer`
4. `logical`

Therefore, the type of `ex_1_vec` is `character`. The underlying reason is that you can, for instance, always transform a `double` value into a `character` but not vice versa.

2. What happens if you coerce this vector into type integer? Why?

```
as.integer(ex_1_vec)
```

```
Warning: NAs introduced by coercion
```

```
[1]  1  2 NA
```

Because `integer` is lower in the hierarchy than `character`, the transformation is not straightforward. By coincidence, the first two elements can actually be coerced into integers (albeit maybe not with the expected result), but there is no way you can transform the logical value `FALSE` into an integer, which is why a missing value is produced.

3. What does `sum(is.na(x))` tell you about a vector `x`? What is happening here?

```
x <- c(1,2,3,NA,NA,8)
```

First, `is.na(x)` creates a vector with logical values indicating whether a value of the original vector is missing (i.e. `NA`):

```
is.na(x)
```

```
[1] FALSE FALSE FALSE  TRUE  TRUE FALSE
```

Then, `sum()` computes the sum over this vecor of boolean values:

```
sum(is.na(x))
```

```
[1] 2
```

Here, `TRUE` counts as one and `FALSE` as zero, so `sum()` gives the number of cases in which `is.na(x)` has evaluated to `TRUE`:

4. Is it a good idea to use `as.integer()` on double characters to round them to the next integer? Why (not)? What other ways are there to do the rounding?

No, because `as.integer()` is not acutally rounding numbers (as, for example, `as.integer(2.1)` would make you think), but only removing the decimal part of the number:

```
as.integer(2.9) # you might expect 2...
```

```
[1] 2
```

Better use `round()`:

```r
round(2.9)
```

```
[1] 3
```

Exercise 2: Define a function

Create functions that take a vector as input and returns:

1. The last value.

```r
get_last_val <- function(x){
  last_val <- x[length(x)]
  return(last_val)
}
```

2. Every element except the last value and any missing values.

```r
get_beginning <- function(x){
  beginning <- x[-length(x)]
  beginning_nonas <- beginning[!is.na(beginning)]
  return(beginning_nonas)
}
```

3. Only even numbers.

   Hint: Use the operation `x %% y` to get the remainder from diving `x` by `y`, the so called 'modulo y'. For even numbers, the modulo 2 is zero.

```r
get_even <- function(x){
  even_nbs <- x[x%%2==0]
  even_nbs_nonas <- even_nbs[!is.na(even_nbs)]
  return(even_nbs)
}
```

Apply your function to the following example vector:

```r
ex_2_vec <- c(1, -8, 99, 3, NA, 4, -0.5, 50)
```

```r
get_last_val(ex_2_vec)
```

```
[1] 50
```

```
get_beginning(ex_2_vec)
```

```
[1]  1.0 -8.0 99.0  3.0  4.0 -0.5
```

```
get_even(ex_2_vec)
```

```
[1] -8 NA  4 50
```

Exercise 3: Lists

1. Create a list that contains three elements called `'a'`, `'b'` and `'c'`. The first element should correspond to a double vector with the elements `1.5`, `-2.9` and `99`. The second element should correspond to a character vector with the elments `'Hello'`, `'3'`, and `'EUF'`. The third element should contain three times the entry `FALSE`.

```
ex_3_list <- list(
  'a' = c(1.5, -2.9, 99),
  'b' = c('Hello', "'3'", 'EUF'),
  'c' = rep(FALSE, 3)
)
```

2. Transform this list into a `data.frame` and a `tibble`. Then apply `str()` to get information about the respective structure. How do the results differ?

```
ex_3_df <- as.data.frame(ex_3_list)
ex_3_tb <- tibble::as_tibble(ex_3_list)
str(ex_3_list)
```

```
List of 3
 $ a: num [1:3] 1.5 -2.9 99
 $ b: chr [1:3] "Hello" "'3'" "EUF"
 $ c: logi [1:3] FALSE FALSE FALSE
```

```
str(ex_3_df)
```

```
'data.frame':   3 obs. of  3 variables:
 $ a: num  1.5 -2.9 99
 $ b: chr  "Hello" "'3'" "EUF"
 $ c: logi  FALSE FALSE FALSE
```

```
str(ex_3_tb)
```

```
tibble [3 x 3] (S3: tbl_df/tbl/data.frame)
 $ a: num [1:3] 1.5 -2.9 99
 $ b: chr [1:3] "Hello" "'3'" "EUF"
 $ c: logi [1:3] FALSE FALSE FALSE
```

`str()` only differs with regard to the first line describing the type.

Exercise 4: Data frames and the study semester distribution at EUF

The package `DataScienceExercises` contains a data set called `EUFstudentsemesters`, which contains information about the distribution of study semesters of enrolled students at the EUF in 2021. You can shortcut the data set as follows:

```
euf_semesters <- DataScienceExercises::EUFstudentsemesters
```

1. What happens if you extract the column with study semesters as a vector and transform it into a `double`?

```
unique(euf_semesters[["Semester"]])
```

```
[1] "6"           "4"           "2"           "8"           "9 or higher"
[6] "7"           "5"           "3"           "1"
```

```
semesters <- as.double(euf_semesters[["Semester"]])
```

```
Warning: NAs introduced by coercion
```

```
unique(semesters)
```

```
[1]  6  4  2  8 NA  7  5  3  1
```

We see that the previous entry `"9 or higher"` has been transformed into `NA`.

2. What is the average study semester of those students being in their 8th or earlier semester?

```
mean(semesters, na.rm = TRUE)
```

```
[1] 4.177026
```

3. How many students are in their 9th or higher study semester?

```
sum(euf_semesters$Semester=="9 or higher")
```

```
[1] 469
```

4. What does `typeof(euf_semesters)` return and why?

```
typeof(euf_semesters)
```

```
[1] "list"
```

It returns `list`, because while `euf_semesters` is a `tibble`, `typeof()` always gives the underlying basic object type. For `tibble`s, this is `list`.