

Data types and structures

AUTHOR

Ralf Ansorg

PUBLISHED

18 September 2025

Exercise 1

Suppose you got the following data or data frame:

```
a <- c(1:5) # mouse ID
b <- c(5, 6, 3, 10, 7) # licks per hours
c <- c("17.18", "16.03", "15.9", "17.99", "14") # length in cm
d <- c("Female", "Male", "Male", "Female", "Maile") # sex of mouse
e <- c(TRUE, FALSE, TRUE, TRUE, FALSE) # healthy mouse
```

```
df <- data.frame(a=a, b=b, c=c, d=d, e=e)
```

1. Which vectors (or columns of the data frame) need some work-over? If so, please write the correct code. *Hint:* try to figure this out by looking at its structure; recall coercion and accessing elements of vectors, and factors. Think about on how to assign new a value to a particular element of a vector, e.g. my_example_vector[5]. **Solution:**

```
class(c)
```

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

```
c
```

```
[1] "17.18" "16.03" "15.9"  "17.99" "14"
```

```
c <- as.numeric(c)
c
```

```
[1] 17.18 16.03 15.90 17.99 14.00
```

```
class(c)
```

```
[1] "numeric"
```

```
d[5] <- "Male" # was misspelled as "Maile"
d <- as.factor(d)
d
```

```
[1] Female Male   Male   Female Male
Levels: Female Male
```

2. Add all vectors to a new data frame `df_clean` with proper/meaningful headers (instead of `a,...,e`). **Solution:**

```
df_clean <- data.frame(ID=a, licks=b, length=c, sex=d, healthy=e)
df_clean
```

```
  ID licks length  sex healthy
1  1     5  17.18 Female    TRUE
2  2     6  16.03  Male   FALSE
3  3     3  15.90  Male    TRUE
4  4    10  17.99 Female    TRUE
5  5     7  14.00  Male   FALSE
```

3. What is the mean of the length of the mice? *Hint: you could use the `mean()` function.* What is the median? **Solution:**

```
mean(df_clean$length)
```

```
[1] 16.22
```

```
median(df_clean$length)
```

```
[1] 16.03
```

The mean length of the mice is 16.22 and the median is 16.03.¹

4. Look at the structure **and** summary if your cleaned-up data is reasonable and briefly explain why. Bullet points only (use the `*` or `-` symbol), no essay. **Solution:**

```
str(df)
```

```
'data.frame':  5 obs. of  5 variables:
 $ a: int  1 2 3 4 5
 $ b: num  5 6 3 10 7
 $ c: chr  "17.18" "16.03" "15.9" "17.99" ...
 $ d: chr  "Female" "Male" "Male" "Female" ...
 $ e: logi  TRUE FALSE TRUE TRUE FALSE
```

```
summary(df)
```

```
      a      b      c      d
Min.   :1  Min.   : 3.0 Length:5 Length:5
1st Qu.:2  1st Qu.: 5.0 Class  :character Class  :character
Median :3  Median : 6.0 Mode   :character Mode   :character
Mean    :3  Mean    : 6.2
3rd Qu.:4  3rd Qu.: 7.0
Max.    :5  Max.    :10.0
      e
Mode  :logical
FALSE:2
TRUE :3
```

¹Check the `.qmd` file on how to have inline code!

```
str(df_clean)
```

```
'data.frame':  5 obs. of  5 variables:
 $ ID      : int  1 2 3 4 5
 $ licks   : num  5 6 3 10 7
 $ length  : num  17.2 16 15.9 18 14
 $ sex     : Factor w/ 2 levels "Female","Male": 1 2 2 1 2
 $ healthy: logi  TRUE FALSE TRUE TRUE FALSE
```

```
summary(df_clean)
```

	ID	licks	length	sex	healthy
Min. :	1	Min. : 3.0	Min. :14.00	Female:2	Mode :logical
1st Qu.:2		1st Qu.: 5.0	1st Qu.:15.90	Male :3	FALSE:2
Median :3		Median : 6.0	Median :16.03		TRUE :3
Mean :3		Mean : 6.2	Mean :16.22		
3rd Qu.:4		3rd Qu.: 7.0	3rd Qu.:17.18		
Max. :5		Max. :10.0	Max. :17.99		

Comments/Bullet points ()

- Variable length had to be converted to numerical (from character), especially for computations
- Variable sex is categorical, hence a factor (without any order)
- Factors and logicals are displayed differently: counts instead of the “big 5”²

Bonus

If you know that FALSE and TRUE are internally represented as 0 and 1. What is the number of healthy mice (computed)?

Solution

Use the `sum()` function to count all TRUEs in column `healthy` which returns 3 (compare it to the `summary()` output above).

²Min, 25% Qrt, (Median), Mean, 75% Qrt, Max