Exercise sheet: Day 1

Vangelis Theodorakis, Fatemeh Behjati, Julien Gagneur

11 October, 2020

Contents

1 Vectors

First, create three named numeric vectors of size 10, 11 and 12 respectively in the following manner:

- One vector with the "colon" approach: *from:to*
- One vector with the `seq()` function: *seq(from, to)*
- And one vector with the `seq()` function and the `by` argument: *seq(from, to, by)*

For easier naming you can use the vector `letters` or `LETTERS` which contain the latin alphabet in small and capital, respectively. In order to select specific letters just use e.g. `letters[1:4]` to get the first four letters. Check their types. What is the outcome? Where do you think the difference comes from?

Then combine all three vectors in a list. Check the attributes of the vectors and the list. What is the difference and why?

Hint: If list elements have no names, we can access them with the double brackets and an index, e.g. `my_list[[1]]`

2 Factors

```
f1 <- factor(letters)
levels(f1) <- rev(levels(f1))
f2 <- rev(factor(letters))
f3 <- factor(letters, levels = rev(letters))
```

The function `rev` reverses the order of an order-able object. What is the difference between f1, f2 and f3? Why?

3 Data tables

The purpose of this exercise is to get familiarize with data.table and try out some of its useful features.

3.1 Basic operations

Please follow the steps listed below:

1) Download the GTEx data (annotation v7) from the following link: https://storage.googleapis.com/gtex_analysis_v7/annotations/GTEx_v7_Annotations_SampleAttributesDS.txt
2) Read the file downloaded above and store it in a variable named: *data*.
3) Inspect *data* by checking properties such as: *class(data)*, *dim(data)*, *colnames(data)*, *data[1:3, 1:5]*, *unique(data$SMTS)*.
4) Count how many NA's exist in *data*.

3.2 More exciting operations

Continue from the previous part and perform the following actions:

3) Subset the data based on the *Brain* cell type sample and store the result in a variable called: *data_Brain*.

4) Inspect the *data_Brain* similar to the point 3 above.

5) Examine the range of values in *SMEXPEFF* field of *data_Brain*. How can you make it more meaningful?

6) For *data_Brain*, compute the average of the values stored in the "SMEXPEFF" column. Also, compute the min of values stored in "SME1MPRT".

7) Compute the correlation between the two columns mentioned above.

8) Remove the rows that are NA from the *data_Brain$SMEXPEFF*. Retry the correlation on the NA-removed *data_Brain_noNA*.

Hint: Use the is.na() function to find the rows that are NA.

4 Looping

- Initialize a variable called *counter* by 0.
- Using a for loop that iterates 10 times, increment *counter* by 1.
- Print the final value in *counter*.

Write a function named *get_counts* that takes a GTEx data table as input and outputs the total counts of rows that the sample tissue type (*SMTS*) is *Heart* and the sample analysis freeze (*SMAFRZE*) is *RNASEQ*. How about if you try the same but for *Blood*. If this task was too easy, can you modify your function such that instead of taking only one argument, it takes two additional ones, one for the *SMTS* and another for *SMAFRZE*. Iterate over all possible values of *SMTS* (**Hint:** *unique(data$SMTS)*) and call your function by providing the sample tissue type.