II. Working with data in R (exercises)

Data Science Lab, University of Copenhagen

16 May, 2022

Importing data

The data set used in these exercises was compiled from data downloaded from the website of the UK's national weather service, the *Met Office*. It is saved in the file **climate.xlsx**[1]. The spreadsheet contains data from five UK weather stations. The following variables are included in the data set:

Variable name	Explanation
station	Location of weather station
year	Year
month	Month
af	Days of air frost
rain	Rainfall in mm
sun	Sunshine duration in hours
device	Brand of sunshine recorder / sensor

1. You should have already imported the data set from the small exercises during the lecture. If not, import it now. You can also do this via the graphical interface by selecting **Import Dataset** from the **Environment** tab on the upper left window. Check that the data preview looks okay and click *Import*.
2. Write the name of the dataframe, i.e. `climate` and press enter to see the first lines of the dataset. You can also click on the `climate` object in the Environment panel.

Working with the data

Before you proceed with the exercises in this document, make sure to run the command `library(tidyverse)` in order to load the core **tidyverse** packages.

3. Select only months, i.e. rows, from the Oxford station where there were no days with airfrost. Assign this new dataset to the variable name `oxford_af`. If you have trouble with this, have a look at Tidyverse exercise B.
4. Compute the average rainfall over all stations and months by using the `summarize` function. You do not need to use any grouping variables. Use the tidyverse syntax, i.e. :

```
new_object <- dataset %>%
  summarize(...)
```

5. Now, calculate the standard deviation of the monthly rainfall as well as the total rainfall (the sum), in addition to the average rainfall as above. I.e. all three measures should be inside the same resulting table. Have a look at the tidyverse lecture if you have trouble with this.
6. Now, use `group_by` before `summarize` in order to compute group summary statistics (average, standard deviation, and sum) for the monthly rainfall observations from each of the five weather stations.

[1] Contains public sector information licensed under the Open Government Licence v3.0.

7. Include a column in the summary statistics which shows how many observations the data set contains for each station.
8. Sort the rows in the output in descending order according to annual rainfall.

The final questions require that you combine commands and variables of the type above.

9. Like in the previous question, compute group summary statistics (average, standard deviation, and sum) for the rainfall observations from all five weather stations. This time, sort the output in ascending order according to the stations' average monthly sunshine duration.
10. Identify the weather station for which the median number of monthly sunshine hours over the months April to September was largest.
11. For each weather station apart from the one in Armagh, compute the total rainfall and sunshine duration during the months with no days of air frost. Present the totals in centimetres and days, respectively.
12. For each of the months in the year 2016:
    a. Count how many stations recorded at least two days of air frost *or* more than 95 mm rain.
    b. Compute the average number of sunshine hours for these stations, for the month in question.
13. Go through your solutions again. Figure out where you could have used the pipe operator `%>%` to make your R-code more readable (instead of saving intermediate data sets). Solve these questions again, this time using the pipe operator.