II. Working with data in R (exercises)

Data Science Lab, University of Copenhagen

14 March, 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. Import the climate data to R from the Excel file **climate.xlsx**. You can, for example, do this via *File* → *Import Dataset*. Check that the data preview looks okay, copy the content of the code preview, and click *Import*.
2. Paste the code that you copied into your R script. Delete the line which starts with `View`.
3. Use the command `climate` to see the first lines of the dataset.

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.

4. Try out the following commands, and understand what they do.

```
filter(climate, station == "oxford")
select(climate, station, year, month, af)
mutate(climate, rain_cm = rain/10)
count(climate, station)
summarize(climate, total_sun = sum(sun))
arrange(climate, sun)
```

5. a. Assign the output of `filter(climate, station == "oxford", af == 0)` to `oxford_af`, that is, use the command

```
oxford_af <- filter(climate, station == "oxford", af == 0)
```

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

b. Type `oxford_af` to view the content of the new data set. What do the observations have in common?

c. Use the `count` function to count how many months with no days of air frost were recorded at the Oxf

6. Change the command `filter(climate, station == "oxford")` so that observations for the weather station in Camborne are included in the output, too. *Help:* Use `|` (vertical bar) for the logical *or*; or use `%in%` for *included in*.

7. Add a variable called **sqrtSun** to the dataset which contains the squareroot of the sunshine duration. Assign this extended dataset to a dataset called **climate2**. *Help:* The squareroot function in R is called `sqrt`.

8. a. Use the `$` operator to extract the rainfall observations from the climate data set, and assign the output to `rain_vector`.

   b. Try out the following commands, and understand the output.

```
rain_vector
rain_vector[1:6]
rain_vector[5]
rain_vector[c(2,4,6)]
```

c. Use the `mean` and `sd` functions to compute the average and standard deviation of the monthly rainfa

d. Use the `sum` function to compute the total annual rainfall recorded in 2016 (total over all five sta

9. Compute the same values as in the previous question, this time by applying the `summarize` function to the climate data set (don't use any grouping variables). This can be done in three lines of code, but try to do it in one line. *HINT:* assign the output of `mean`, `sd` and `sum` a name.

10. a. Use `group_by` and `summarize` to compute group summary statistics (average, standard deviation, and sum) for the monthly rainfall observations from each of the five weather stations. NB: This is done most elegantly with the pipe operator (`%>%`).

    b. Include a column in the summary statistics which shows how many observations the data set contains for each station.

    c. 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.

11. 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.

12. Identify the weather station for which the median number of monthly sunshine hours over the months April to September was largest.

13. 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.

14. 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.

15. 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.