

II. Working with data in R (solution)

Data Science Lab, University of Copenhagen

2026-05-06

Table of contents

Importing data	1
Working with the data	2

Importing data

Loading the core `tidyverse` packages, as well as the `readxl` package for importing data from `.xlsx`

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr      1.2.0      v readr      2.1.6
v forcats    1.0.1      v stringr    1.6.0
v ggplot2    4.0.2      v tibble     3.3.1
v lubridate  1.9.5      v tidyr      1.3.2
v purrr      1.2.1
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()     masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

```
library(readxl)
```

1. (and 2 and 3) Use the *Import data facility*.

We have already loaded the `readxl` package, so we can use the `read_excel()` function from that package.

```
climate <- read_excel("climate.xlsx")
climate
```

```
# A tibble: 60 x 7
  station year month   af rain  sun device
  <chr>   <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
1 armagh  2016     1     5 132.  44.5 Campbell Stokes
2 armagh  2016     2    10  62.6  71.3 Campbell Stokes
3 armagh  2016     3     4  43.8 117. Campbell Stokes
4 armagh  2016     4     5  54    140. Campbell Stokes
5 armagh  2016     5     0  41.4 210. Campbell Stokes
6 armagh  2016     6     0  75.1 114. Campbell Stokes
7 armagh  2016     7     0  80.6 113. Campbell Stokes
8 armagh  2016     8     0  52.5 135. Campbell Stokes
9 armagh  2016     9     0  65.4  91.1 Campbell Stokes
10 armagh 2016    10     0  37.1  89.8 Campbell Stokes
# i 50 more rows
```

Working with the data

Below we will use functions from the `tidyverse` package, which was loaded using `library(tidyverse)` above.

4. Trying out some basic commands.

```
# Only data from Oxford
filter(climate, station == "oxford")
```

```
# A tibble: 12 x 7
  station year month   af rain  sun device
  <chr>   <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
1 oxford  2016     1     5  83.9  59.1 Campbell Stokes
2 oxford  2016     2     6  47.6 113. Campbell Stokes
3 oxford  2016     3     4  74.2 124. Campbell Stokes
4 oxford  2016     4     1  53.1 164. Campbell Stokes
5 oxford  2016     5     0  86.1 203. Campbell Stokes
6 oxford  2016     6     0  95.7 100. Campbell Stokes
7 oxford  2016     7     0   3.4 228. Campbell Stokes
8 oxford  2016     8     0  41.2 204. Campbell Stokes
9 oxford  2016     9     0  44.6 113. Campbell Stokes
10 oxford 2016    10     0  26.5 112. Campbell Stokes
11 oxford 2016    11     3  76.1  88.3 Campbell Stokes
12 oxford 2016    12    10  25.8  62.3 Campbell Stokes
```

```
# Only certain variables
select(climate, station, year, month, af)
```

```
# A tibble: 60 x 4
  station year month   af
  <chr>   <dbl> <dbl> <dbl>
1 armagh  2016     1     5
2 armagh  2016     2    10
3 armagh  2016     3     4
4 armagh  2016     4     5
5 armagh  2016     5     0
6 armagh  2016     6     0
7 armagh  2016     7     0
8 armagh  2016     8     0
9 armagh  2016     9     0
10 armagh 2016    10     0
# i 50 more rows
```

```
# New variable with rain measured in cm
mutate(climate, rain_cm = rain/10)
```

```
# A tibble: 60 x 8
  station year month   af rain  sun device      rain_cm
  <chr>   <dbl> <dbl> <dbl> <dbl> <dbl> <chr>      <dbl>
1 armagh  2016     1     5 132.  44.5 Campbell Stokes  13.2
2 armagh  2016     2    10  62.6  71.3 Campbell Stokes   6.26
3 armagh  2016     3     4  43.8 117. Campbell Stokes   4.38
4 armagh  2016     4     5  54   140. Campbell Stokes    5.4
5 armagh  2016     5     0  41.4 210. Campbell Stokes   4.14
6 armagh  2016     6     0  75.1 114. Campbell Stokes   7.51
7 armagh  2016     7     0  80.6 113. Campbell Stokes   8.06
8 armagh  2016     8     0  52.5 135. Campbell Stokes   5.25
9 armagh  2016     9     0  65.4  91.1 Campbell Stokes   6.54
10 armagh 2016    10     0  37.1  89.8 Campbell Stokes   3.71
# i 50 more rows
```

```
# No of obs at each station
count(climate, station)
```

```
# A tibble: 5 x 2
  station      n
  <chr>    <int>
1 armagh    12
```

```
2 camborne      12
3 lerwick       12
4 oxford        12
5 sheffield     12
```

```
# Total sum over all stations
summarize(climate, total_sun = sum(sun))
```

```
# A tibble: 1 x 1
  total_sun
  <dbl>
1      6829.
```

```
# Sort after sun
arrange(climate, sun)
```

```
# A tibble: 60 x 7
  station year month   af rain sun device
  <chr>   <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
1 lerwick 2016    12     0 159.  11.5 Kipp Zonen
2 lerwick 2016     1     7 187.  34.6 Kipp Zonen
3 sheffield 2016     1     3 84.8  40.3 Kipp Zonen
4 armagh   2016     1     5 132.  44.5 Campbell Stokes
5 camborne 2016    12     0 58.4  47.3 Kipp Zonen
6 camborne 2016     1     0 222.   48 Kipp Zonen
7 lerwick 2016    11     1 133.   48 Kipp Zonen
8 armagh   2016    12     1 51.4  50.5 Campbell Stokes
9 camborne 2016    11     0 137.  56.5 Kipp Zonen
10 sheffield 2016    12     3 31.8  57.9 Kipp Zonen
# i 50 more rows
```

5. Assigning the data set of observations from Oxford with 0 days of air frost to `oxford_af`.

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

```
# A tibble: 6 x 7
  station year month   af rain sun device
  <chr>   <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
1 oxford 2016     5     0 86.1 203. Campbell Stokes
2 oxford 2016     6     0 95.7 100. Campbell Stokes
3 oxford 2016     7     0  3.4 228. Campbell Stokes
4 oxford 2016     8     0 41.2 204. Campbell Stokes
5 oxford 2016     9     0 44.6 113. Campbell Stokes
6 oxford 2016    10     0 26.5 112. Campbell Stokes
```

Counting the number of observations assigned to `oxford_af`.

```
count(oxford_af)
```

```
# A tibble: 1 x 1
      n
  <int>
1     6
```

Alternative solution without using `oxford_af`. Preferably done using the pipe operator `%>%`.

```
climate %>%
  filter(station == "oxford", af == 0) %>%
  count()
```

```
# A tibble: 1 x 1
      n
  <int>
1     6
```

6. Dataset with observations from Camborne and Oxford.

```
filter(climate, station == "camborne" | station == "oxford")
```

```
# A tibble: 24 x 7
  station year month   af rain  sun device
  <chr>   <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
1 camborne 2016     1     0 222.   48 Kipp Zonen
2 camborne 2016     2     0 162.  64.1 Kipp Zonen
3 camborne 2016     3     0  88.4 140. Kipp Zonen
4 camborne 2016     4     0  81.4 184. Kipp Zonen
5 camborne 2016     5     0  45.6 206. Kipp Zonen
6 camborne 2016     6     0  65.8 132. Kipp Zonen
7 camborne 2016     7     0  23.2 161. Kipp Zonen
8 camborne 2016     8     0  57.4 171. Kipp Zonen
9 camborne 2016     9     0 154.  103. Kipp Zonen
10 camborne 2016    10     0  53.2 125. Kipp Zonen
# i 14 more rows
```

```
# Or: filter(climate, station %in% c("camborne", "oxford"))
```

7. Make the new variable and give the new dataset a name:

```
climate2 <- mutate(climate, sqrtSum=sqrt(sun))
```

8. Assigning the rainfall observations to the vector `rain_vector`.

```
rain_vector <- climate$rain
```

Extracting elements from the rainfall vector.

```
rain_vector
```

```
[1] 131.9  62.6  43.8  54.0  41.4  75.1  80.6  52.5  65.4  37.1  40.8  51.4
[13] 221.6 161.6  88.4  81.4  45.6  65.8  23.2  57.4 153.7  53.2 137.0  58.4
[25] 187.0 119.6  79.2  55.1  46.0  48.4 115.2 108.4 110.0  56.2 133.4 159.4
[37]  83.9  47.6  74.2  53.1  86.1  95.7   3.4  41.2  44.6  26.5  76.1  25.8
[49]  84.8  68.6  87.2  65.8  58.2 130.4  30.0  62.0  44.2  29.2  95.6  31.8
```

```
# First six elements
```

```
rain_vector[1:6]
```

```
[1] 131.9  62.6  43.8  54.0  41.4  75.1
```

```
# Element number five
```

```
rain_vector[5]
```

```
[1] 41.4
```

```
# Element numbers 2, 4 and 6
```

```
rain_vector[c(2,4,6)]
```

```
[1] 62.6 54.0 75.1
```

Some summary statistics for the rainfall observations.

```
mean(rain_vector)
```

```
[1] 75.79667
```

```
sd(rain_vector)
```

```
[1] 43.18915
```

```
sum(rain_vector)
```

```
[1] 4547.8
```

9. Summary statistics for the rainfall observations computed using `summarize`. It is most easily made with the pipe operator, but can also be done in several steps (not shown).

```
summarize(climate,
  avg_rain = mean(rain),
  sd_rain = sd(rain),
  sum_rain = sum(rain))
```

```
# A tibble: 1 x 3
  avg_rain sd_rain sum_rain
  <dbl>    <dbl>    <dbl>
1    75.8    43.2   4548.
```

10. Summary statistics for rainfall data by weather station, including number of observations for each station, and sorted in descending order according to annual rainfall.

```
climate %>%
  group_by(station) %>%
  summarize(avg_rain = mean(rain),
    sd_rain = sd(rain),
    sum_rain = sum(rain),
    n=n()) %>%
  arrange(desc(sum_rain))
```

```
# A tibble: 5 x 5
  station avg_rain sd_rain sum_rain    n
  <chr>    <dbl>    <dbl>    <dbl> <int>
1 lerwick    101.    45.6    1218.    12
2 camborne    95.6    59.4    1147.    12
3 sheffield    65.6    30.5     788.    12
4 armagh     61.4    26.1     737.    12
5 oxford     54.8    28.5     658.    12
```

11. Summary statistics for rainfall data by weather station, sorted in ascending order according to average monthly sunshine duration.

```
climate %>%
  group_by(station) %>%
  summarize(avg_rain = mean(rain),
    sd_rain = sd(rain),
    sum_rain = sum(rain),
    avg_sun = mean(sun)) %>%
  arrange(avg_sun)
```

```
# A tibble: 5 x 5
  station avg_rain sd_rain sum_rain avg_sun
  <chr>    <dbl>    <dbl>    <dbl>    <dbl>
```

1	lerwick	101.	45.6	1218.	101.
2	armagh	61.4	26.1	737.	104.
3	sheffield	65.6	30.5	788.	113.
4	camborne	95.6	59.4	1147.	120.
5	oxford	54.8	28.5	658.	131.

12. Weather station with largest median number of monthly sunshine hours over the months April to September was Oxford:

```
climate %>%
  filter(month %in% 4:9) %>%
  group_by(station) %>%
  summarize(med_sun = median(sun)) %>%
  arrange(desc(med_sun))
```

```
# A tibble: 5 x 2
  station med_sun
  <chr>    <dbl>
1 oxford    183.
2 camborne   166
3 sheffield  160.
4 lerwick   132.
5 armagh    124.
```

13. For each weather station apart from Armagh, the total rainfall (in cm) and duration of sunshine (in days) in the months with no days of air frost.

```
climate %>%
  group_by(station) %>%
  filter(station != "armagh", af == 0) %>%
  summarize(total_rain = sum(rain)/10,
            total_sun = sum(sun)/24)
```

```
# A tibble: 4 x 3
  station total_rain total_sun
  <chr>    <dbl>    <dbl>
1 camborne    115.     59.9
2 lerwick     64.4     33.3
3 oxford      29.8     40.0
4 sheffield   35.4     35.0
```

14. For each month: Number of stations with at least two days of air frost or more than 95 mm rain, and the average sunshine duration for these stations.


```
climate %>%
  filter(af >= 2 | rain > 95) %>%
  group_by(month) %>%
  summarize(count = n(), avg_sun = mean(sun))
```

```
# A tibble: 10 x 3
  month count avg_sun
  <dbl> <int>   <dbl>
1     1     5    45.3
2     2     5    86.2
3     3     4   106.
4     4     2   148.
5     6     2   103.
6     7     1    80.5
7     8     1    92.2
8     9     2   105.
9    11     5    68.0
10   12     3    43.9
```

15. Many solutions were already made with the pipe operator.
16. Computation of station-wise averages and standard deviations averages of both rain and sun:

```
climate %>% group_by(station) %>%
  summarize(across(c(rain,sun), list(avg=mean,sd=sd)))
```

```
# A tibble: 5 x 5
  station rain_avg rain_sd sun_avg sun_sd
  <chr>      <dbl>   <dbl>   <dbl> <dbl>
1 armagh     61.4    26.1    104.  45.1
2 camborne    95.6    59.4    120.  55.8
3 lerwick    101.     45.6    101.  57.5
4 oxford     54.8    28.5    131.  56.1
5 sheffield   65.6    30.5    113.  49.2
```

End of solution