Team B : Rainfall Data Reporting

Barry Rowlingson

October 29, 2021

Task Outline

You have been given a data file of rainfall readings, and a data file of weather station information. The readings are monthly rainfall measurements in inches at a small number of sites in Lancaster, and cover the 50-year period from 1851 to 1900.

Your task is to produce a neatly formatted short report of the rainfall in that period. This briefing document will explain the outputs required in the document, but you should also make sure you write explanatory text.

Your report shouldn't show any code, since its intended for a general audience.

Meanwhile Team A are preparing another similar pair of files from a different location, so you should try and make your report and code flexible enough to cope with a possibly slightly different data set!

Within the text various R functions will appear in brackets as guides for how to do something. These are just a hint and there may well be other ways – even better ways – to do something.

Data File Format

The rainfall readings are in a file called `readings.csv` and the first few lines look like this:

```
"stationID","month","year","rain"
 "3049","January",1893,1.1
 "3049","February",1893,4.44
 "3049","March",1893,1.19
 "3049","April",1893,0.45
```

The weather station data is in a file called `stations.csv` and the first few lines look like this:

```
"stationID","name","long","lat","elev"
 "RR1669","CATON SCHOOL HOUSE",-2.7004,54.073,170
 "3050/5","LANCASTER ESCOWBECK",-2.727349,54.070631,150
 "3049","LANCASTER GREG OBSERVATORY",-2.78492,54.044204,312
 "RR81","HEST BANK LANCASTER",-2.807,54.0916,58
```

To avoid ambiguity, we'll say that a "month" is any of the 12 calendar months, and a "date" is a month in a given year, possibly using the first of that month as a day when we need a single day to represent the month. So the rainfall for April 1846 would be 1946-April-01.

Data Reading and Preparation

The first thing to do is read the data in (`read.csv`) and see what you get (`summary`).

Convert the month column to a factor with levels set to the correct order of the month names (`factor`, `month.name`).

Create a year-month-day column as a text character column from the year and month with the day set to 1 (`paste0`). Convert this to a `Date` class (`as.Date`). The code for a long month name is `%B`, so your conversion format should be `%Y-%B-%d` – for example:

```
as.Date("1966-March-31", format="%Y-%B-%d")

[1] "1966-03-31"
```

Check using `summary` again that your data has all expected columns and there are not odd values in there.

Months have different numbers of days, so comparisons of total monthly rainfalls should consider this, but we'll ignore it.

Read the station data from the file. Check that the station IDs are unique in the station data (`any`, `duplicated`), and that every station ID in the readings appears in the station data (`all`, `%in%`).

You might want to put all this reading and cleaning into a function which takes two arguments and returns a list of readings and stations. In the event of the ID tests not passing, stop the code with an error message (`stop`). Your function would look something like this in outline:

```r
read_readings_stations <- function(readings, stations){

### read the files...

### clean, modify, and check the data...

    if( something_not_right ){
        stop("Useful error message here")
    }

    return(
        list(
            readings=readings,
            stations=stations
        )
    )
}
```

Check that the function works properly at the R command line before starting a new document and putting it into a code block.

Data Aggregation

For the charts and tables in the report, create some aggregated data frames (`aggregate`, `dplyr::`, `data.table`, etc).

1. Some months have multiple rainfall readings from different stations. Create a data frame averaging the readings for each date. It should have columns for the date and the rainfall. Add the month and year as extra columns (`month`, `year`) by extracting these values from the date column.

2. To investigate the wettest and driest months, create a data frame of average rainfall for each month over all years and stations. This should only have 12 rows, one for each month.

3. To investigate station annual extremes, sum the 12 monthly rainfall readings for each station for each year. The resulting table should have station ID, year, and rainfall columns.

Report Contents

Rainfall Graph

Plot the mean rainfall over all dates as a line graph.

Year-Month Chart

The line graph can be hard to interpret because the information is concentrated in that one dimension. Instead we can exploit the monthly structure to display the same data in two dimensions.

One way to do this is to reformat the data into a matrix with 12 rows (one for each month) and 50 columns (one for each year) (`image`), the other is to treat the long data as coordinates and values as if it was a scatterplot, colouring or scaling the points by the rainfall (`plot`, `geom_raster`).

The resulting output should be a colour-shaded grid of 12 rows and 50 columns, labelled as years on the X-axis and months on the Y-axis.

Monthly Rainfall Pattern

Plot the twelve values of mean rainfall for each month. What would be the best way to display this? Various options are possible (`plot`, `barplot`, `pie`). Display the twelve numeric values as well, rounded to a small number of decimal places (`text`, `round`, `as.character`).

Extreme Tables

Highest Individual Readings

Merge the readings data with the stations (`merge`, `dplyr::left_join`) to get a dataset with the station data on each reading. Next sort this data frame (`order`) by decreasing rain and show the five highest monthly readings. The table should have station name, month, year, and rainfall record.

Highest and Lowest Annual Station Records

Give a table of the five stations that recorded the highest yearly total rainfall measurements with name, year, and total amount.

Which stations had the lowest 5 total annual rainfall in any year? Show this in a table with the absolute lowest first.

Wettest Month Records

How often is each month the wettest in a year?

There's a number of ways of doing this. Here's one:

- Create a sorted version of the month-year rainfall table by decreasing rainfall (`order`).
- The highest rainfall in any year is then the first occurrence of that year in the year column.
- The `duplicated` function is `TRUE` for the first occurrence of any value in a vector.
- Hence a subset of the sorted data frame, by the *un*-duplicated records will be the highest rainfall values for each year.

Other methods include using `split` to break the data frame into parts and then finding the maximum value in each part before combining them again, or similar operations using `aggregate` or `dplyr` functions. The `which.max` function may be useful here.