Wrangling data with the tidyverse

2031 - Introduction to Statistical Computing

Sam Clifford

Session 9

Introduction

About this practical session

In the lecture session we introduced data wrangling with the tidyverse as an alternative to using base R for common tasks.

This prac will investigate some simple tasks for facilitating exploratory analysis of a data set on birth weight, date, and gestational period collected as part of the Child Health and Development Studies in 1961 and 1962. Information about the baby's parents — age, education, height, weight, and whether the mother smoked is also recorded.

The study was designed to investigate the relationship between smoking status and birth weight and common confounders are included which may moderate the biology of the baby directly or through environmental or social factors.

You are not required to use the pipe operator, `%>%`, from magrittr, but you are welcome to attempt to if you feel comfortable with piping in UNIX systems or feel like you have a reasonable level of R skills.

- Assumed skills
  - Writing R code into a script file
  - Familiarity with data frames and spreadsheet-style data
  - Understanding of summary statistics
- Learning objectives
  - reshaping data
  - calculating summary statistics for grouped data without using loops

A reminder of expectations in the prac:

- Keep a record of the work being completed with a well-commented R script
- Allow everyone a chance to participate in the learning activities, keeping disruption of other students to a minimum while still allowing for fruitful discussion
- All opinions are valued provided they do not harm others
- Everyone is expected to do the work, learning seldom occurs solely by watching someone else do work

Activity 1 - Quick look at the data

We will be looking at the Gestation data set as found in the mosaicData package (Pruim, Kaplan, and Horton 2020; Nolan and Speed 2001). This data has been collected from the USA and contains records on 1236 single births between 1961 and 1962.

Exercise: Copy and paste the code below to load the data set and the tidyverse package.

```
library(tidyverse)
library(mosaicData)
data(Gestation)
```

Exercise: Copy and paste the code below to count the number of observations in the data set

```
count(Gestation)
```

Exercise: Copy and paste the code below to count the number of observations in the data set in each level of the `race` variable, as in the slides

```
count(Gestation, race)
```

Exercise: Copy, paste and modify the code below to count the number of observations in the data set in each level of the `race` and `ed` (educational attainment) variables. Store and then print the result.

```
Gestation_n_race_ed <- count(Gestation, ...)
Gestation_n_race_ed
```

Activity 2 - Further summary statistics

Now that we are familiar with our data frame and that we can count the number of entries, we will focus on some more useful summaries of the data.

Exercise: Calculate the mean age of mothers in the data set

Exercise: Calculate the mean age of mothers in the data set again, but this time give a human friendly name for the calculated column. Here the backticks are necessary to allow us to put punctuation (a space) in our column name.

Exercise: By adding an additional line to the calculation of summary statistics, calculate the mean birth weight in the data set.

Activity 3 - Grouped summaries

We want to calculate multiple summary statistics for each level of the `race` variable in the data set. In the lecture, we used `summarise_at()` to apply multiple functions to multiple variables. Here, we will reshape the data frame so that we have a key-value representation of the data.

Exercise: Make a new data frame containing only the `id`, `age` and `race` variables.

Exercise: Calculate the mean age for each race group in this data frame

Activity 4 - Extension activities

The following activities increase in level of difficulty. You are not expected to do all of them, but it's suggested that you attempt at least one of them during class time.

Activity 4a - Correlation

Exercise: Calculate the correlation between age and weight for all data using the `summarise` function.

Exercise: Calculate the correlation between age and weight for each race group.

Activity 4b - Multiple summary statistics

Exercise: Calculate the sample mean of the ages and weights of the mothers in each race group. You may wish to modify the solution to the final exercise in Activity 2 to do this.

Activity 4c - Pivoting wider

Exercise: Make a table from the summaries stored in `Gestation_n_race_ed` such that each row is an education level and each column is a different level of the `race` variable. Hint: Look at the help file for `pivot_wider` for an argument that allow you to specify what value to fill any missing cells with and set it to 0

Activity 4d - Multiple summary statistics

Exercise: Either by specifying separately or using `summarise_at()`, calculate the mean, standard deviation, minimum, maximum and proportion missing for the age values for each race group.

Tidy up

Make sure you save your R script.

Further reading

- More help on the tidyverse is available
- The `#r4ds` community have TidyTuesday which makes use of the ideas in the R for Data Science book (Wickham and Grolemund 2017)
- Wickham (2014) on what tidy data is
- Wickham et al. (2019) for an explanation as to what the tidyverse is

References

Nolan, Deborah, and Terry P Speed. 2001. *Stat Labs: Mathematical Statistics Through Applications*. Springer Science & Business Media.

Pruim, Randall, Daniel Kaplan, and Nicholas Horton. 2020. *mosaicData: Project MOSAIC Data Sets*. https://CRAN.R-project.org/package=mosaicData.

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (1):1–23. https://doi.org/10.18637/jss.v059.i10.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43):1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and Garrett Grolemund. 2017. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media. http://r4ds.had.co.nz.