Exploratory data analysis

2491 - Data Challenge

Sam Clifford

Session 3 2022-01-13

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, and visualisation with ggplot2.

This prac will investigate some simple tasks for facilitating exploratory analysis of two data sets. The first is 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.

The second data set is repeated measurement of the forced expiratory volume in the first second of breath, FEV1, a measure of lung function. The data are from the "Six Cities" study, which aimed to investigate the relationship between air pollution and mortality described in Dockery et al. (1983). Re-analysis of this data is detailed in Krewski, Burnett, Goldberg, Hoover, Siemiatycki, Abrahamowicz, and White (2005) and Krewski, Burnett, Goldberg, Hoover, Siemiatycki, Abrahamowicz, Villeneuve, et al. (2005), and a discussion of the original study and the replicability of its results is found in Choirat, Braun, and Kioumourtzoglou (2019)

Getting help: You may find it useful to look at the RStudio cheatsheets (RStudio 2021) and ggplot2 documentation for hints on how to implement particular graphical ideas. Another good resource is "R for Data Science" (Wickham and Grolemund 2017), particularly Chapter 3, "Data Visualisation," and Chapter 7, "Exploratory Data Analysis."

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.

Download the material you need for the prac by browsing to the git repository at https://github.com/samclifford/2491_eda and clicking the "Fork" button in the top right corner. This will give you a copy of the repository under your own account that you'll have write access to. Once you have done this, clone the github repository either via RStudio or command line git.

```
git clone https://github.com/<your user name>/2491_eda
```

Work through the `eda_gestation.R` script adding your solutions under the comments that describe the questions. Ensure you regularly commit and push your changes, with descriptive comments, after finishing each activity.

- 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
  - Creating a graph using a layered grammar of graphics

    - Being able to critique a graph that you have created
- Professional skills
    - Using a git version control system
    - Creating exploratory graphics which are reproducible and clear
    - Documenting code by commenting

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.

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.

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.

Activity 4 - Extension activities

These 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. Feel free to discuss your approach with other students in the group.

Activity 5 - Visualisation of the FEV1 data

Choose one person to add the others as collaborators on their cloned GitHub repository. Have them fork the GitHub repository again, this time giving it a different name (e.g. `2491_eda_collab`) and then add each of the other team members as collaborators, by going to `https://github.com/<username>/<reponame>/settings/access` and clicking "Add people." Add each person by their GitHub user name. Once this is done, have them clone the repository to their local machine (they do not need to fork it, as they'll have direct write access).

Have each person open the `vis_fev1.R` script in the `R` folder of your cloned repository.

On one person's computer, build a plot that shows the relationship between FEV1 and age. Commit and push the results up. Have all others in the group pull the changes from the repository.

Question: Given the strength of the linear association between these two variables, do you think a linear trend would be an appropriate model?

Activity 6 - Improving the plot

You may wish to save the plot object in Activity 5 and add to it, continuing to do so from here on, or copy-paste the code and modify it.

On another person's computer, add meaningful labels for the x and y axes, including units, and change the plot's colour theme from the default. You may want to consult the suggested reading or search online for the included ggplot2 theme choices.

Add a smooth line of best fit to the plot. You may wish to change its colour, turn off the standard error ribbon, or make other changes to it to help show the data and improve contrast with the background colour of your plot. The default behaviour is to use a LOESS smoother (Cleveland, Grosse, and Shyu 1992) which can be set with `method = 'loess'` as an argument to `geom_smooth()`. You could also use a generalised additive model (Wood 2017) with `method = 'mgcv'`.

Push these changes up and have everyone else pull them down.

Activity 7 - Collaborative activities

Between you and your group, divide up the following activities so that each of you is working on one (you may need to double up).

Activity 7a - Showing structure

We have repeat measurements on 20 individuals. Through either small multiples, geometry grouping, or other aesthetic options, determine a way to highlight which observations belong to the same individual.

Activity 7b - How many observations per individual?

Many of the 300 individuals in the downloaded data set have been measured multiple times over the years. Count the number of times that each `id` is measured and make a bar plot to summarise the proportion of individuals who have 1, 2, etc. measurements.

Activity 7c - Incorporating height

Make a plot that shows both FEV1 and age but also includes height. There are a number of ways to do this.

Activity 7d - additional summaries with skimr

The skimr package (Waring et al. 2020) provides a way of extending the five number summary provided by `summary()` to a seven number summary, histogram of values, the number of observations missing, and corresponding completeness rate. Use `skim()` to generate a summary table of the data and investigate what information is available to you.

You may also wish to experiment with passing in a grouped data frame (e.g. by age in whole years at enrolment). As `skim()` returns an object which is of class `data.frame` (and `tbl_df` among others) we can filter it, pivot from wide to long format, etc.

Activity 7e - Additional summary tools - GGally

The GGally package contains many functions built using ggplot2 that allow further exploratory analysis. In particular, pairs plots (with `ggpairs()`) show pairwise comparisons for all variables passed in to them and can be a useful way to visualise many relationships at once.

Generate a pairs plot with `ggpairs()` but ensure you specify all columns except `id`, as this will result in an incomprehensible mess. You can pass additional options to `ggpairs()` (check the help file) to control aesthetics, what goes in the upper and lower triangles of the grid, whether to use density plots, histograms, etc. for the univariate summaries on the main diagonal of the plot, and many other things.

NB: Because `ggpairs()` returns an object of class `gg`, you can add to it using the grammar of graphics.

Activity 7f - Accounting for repeat measurement

This is the most difficult task, should probably only be attempted by discussion within the group and requires the use of the mgcv package (Wood 2017).

Build a regression model for the change in FEV1 with age that accounts for repeat measurement of individuals. Making use of the mgcv package in R (Wood 2017), we can fit a mixed effects model that uses a spline for the effect of age and has a random effects mean to account for the differences in baseline FEV1 across individuals using the `gamm` function (Wood 2004). You may wish to use one `geom_line()` for the data and another `geom_line()` for the predicted values. When building your prediction data frame, make sure that you give the predicted values the name `FEV1` so you can reuse the aesthetics from the base plot.

Tidy up

Make sure you have saved your R script. Ensure that you have committed and pushed all your changes up to the shared remote repository and that you have pulled others' changes.

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

A lot of the key ideas in data visualisation that we will investigate arose with Tufte (1983), and are summarised by Pantoliano (2012). Some of the history of data visualisation is summarised well by Friendly (2005) and Friendly (2006). Tufte's website is well worth exploring, particularly the discussion on how the visual presentation of information could have helped avert the *Challenger* disaster (Tufte 1997). For some more guidance on using ggplot2 for data visualisation, check Chapter 3 of Wickham and Grolemund (2017), the RStudio cheatsheets (RStudio 2021), and Chang (2017).

References

Chang, Winston. 2017. *R Graphics Cookbook: Practical Recipes for Visualizing Data*. 2nd ed. O'Reilly Media. http://www.cookbook-r.com/Graphs/.

Choirat, Christine, Danielle Braun, and Marianthi-Anna Kioumourtzoglou. 2019. "Data Science in Environmental Health Research." *Current Epidemiology Reports* 6 (3): 291–99. https://doi.org/10.1007/s40471-019-00205-5.

Cleveland, WS, E Grosse, and WM Shyu. 1992. "Local Regression Models." In *Statistical Models in s*, edited by JM Chambers and TJ Hastie. Wadsworth & Brooks/Cole, Pacific Grove, CA.

Dockery, DW, CS Berkey, JH Ware, FE Speizer, and BG Ferris Jr. 1983. "Distribution of Forced Vital Capacity and Forced Expiratory Volume in One Second in Children 6 to 11 Years of Age." *American Review of Respiratory Disease* 128 (3): 405–12.

Friendly, M. 2005. "Milestones in the History of Data Visualization: A Case Study in Statistical Historiography." In *Classification: The Ubiquitous Challenge*, edited by C. Weihs and W. Gaul, 34–52. New York: Springer. http://www.math.yorku.ca/SCS/Papers/gfkl.pdf.

———. 2006. “A Brief History of Data Visualization.” In *Handbook of Computational Statistics: Data Visualization*, edited by C. Chen, W. Härdle, and A Unwin. Vol. III. Heidelberg: Springer-Verlag. http://www.datavis.ca/papers/hbook.pdf.

Krewski, D., R. T. Burnett, M. Goldberg, K. Hoover, J. Siemiatycki, M. Abrahamowicz, P. J. Villeneuve, and W. White. 2005. “Reanalysis of the Harvard Six Cities Study, Part II: Sensitivity Analysis.” *Inhalation Toxicology* 17 (7-8): 343–53. https://doi.org/10.1080/08958370590929439.

Krewski, D., R. T. Burnett, M. Goldberg, K. Hoover, J. Siemiatycki, M. Abrahamowicz, and W. White. 2005. “Reanalysis of the Harvard Six Cities Study, Part i: Validation and Replication.” *Inhalation Toxicology* 17 (7-8): 335–42. https://doi.org/10.1080/08958370590929402.

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

Pantoliano, Mike. 2012. “Data Visualization Principles: Lessons from Tufte.” 2012. https://moz.com/blog/data-visualization-principles-lessons-from-tufte.

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

RStudio. 2021. “RStudio Cheat Sheets.” 2021. https://www.rstudio.com/resources/cheatsheets/.

Tufte, Edward R. 1983. *The Visual Display of Quantitative Information*. Graphics Press.

———. 1997. “Visual and Statistical Thinking.” In *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press. https://www.edwardtufte.com/tufte/books_textb.

Waring, Elin, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu, and Shannon Ellis. 2020. *Skimr: Compact and Flexible Summaries of Data*. https://CRAN.R-project.org/package=skimr.

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.

Wood, S. N. 2004. “Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models.” *Journal of the American Statistical Association* 99 (467): 673–86.

———. 2017. *Generalized Additive Models: An Introduction with r*. 2nd ed. Chapman; Hall/CRC.