Data Visualization with R

Instructor's Guide

Table of Contents

Dataset

The data used for this lesson are in the figshare repository at: https://figshare.com/articles/SAFI_Survey_Results/6262019.

This lesson uses `SAFI_clean.csv`. The direct download link for this file is: https://ndownloader.figshare.com/files/11492171.

When time comes in the lesson to use this file, we recommend that the instructors place the `download.file()` command in the Etherpad, and that the learners copy and paste it in their scripts to download the file directly from figshare in their working directory. If the learners haven't created the `data/` directory and/or are not in the correct working directory, the `download.file()` command will produce an error. Therefore, it is important to use the stickies at this point.

RStudio and Multiple R Installs

Some learners may have previous R installations. On Mac, if a new install is performed, the learner's system will create a symbolic link, pointing to the new install as 'Current.' Sometimes this process does not occur, and, even though a new R is installed and can be accessed via the R console, RStudio does not find it. The net result of this is that the learner's RStudio will be running an older R install. This will cause package installations to fail. This can be fixed at the terminal. First, check for the appropriate R installation in the library:

```
ls -l /Library/Frameworks/R.framework/Versions/
```

We are currently using R >=3.2. If it isn't there, they will need to install it. If it is present, you will need to set the symbolic link to Current to point to the R >=3.2 directory:

```
ln -s /Library/Frameworks/R.framework/Versions/3.x.y
/Library/Frameworks/R.framework/Version/Current
```

Then restart RStudio.

Narrative

Before we start

- The main goal here is to help the learners be comfortable with the RStudio interface. We use RStudio because it helps make using R more organized and user friendly.
- Go very slowly in the "Getting set up" section. Make sure everyone is following along (remind learners to use the etherpad). Plan with the helpers at this point to go around the room, and be available to help. It's important to make sure that learners are in the correct working directory, and that they create a `data` (all lowercase) subfolder.

Visualizing data with ggplot2

- This lesson is a broad overview of **ggplot2** and focuses on (1) getting familiar with the layering system of ggplot2, (2) using the argument group in the aes() function, (3) basic customization of the plots.

Solutions to Exercises

Before we start

1. Use both the Console and the Packages tab to confirm that you have the tidyverse installed.

Visualizing data with ggplot2

Set 1

1. Use what you just learned to create a scatter plot of `rooms` by `village` with the `respondent_wall_type` showing in different colors. Does this seem like a good way to display the relationship between these variables? What other kinds of plots might you use to show this type of data?

```
interviews_plotting %>%
    ggplot(aes(x = village, y = rooms)) +
    geom_jitter(aes(color = respondent_wall_type),
                alpha = 0.5,
                width = 0.2,
                height = 0.2)
```

Set 2

1. Boxplots are useful summaries, but hide the *shape* of the distribution. For example, if the distribution is bimodal, we would not see it in a boxplot. An alternative to the boxplot is the violin plot, where the shape (of the density of points) is drawn.
    - Replace the box plot with a violin plot; see `geom_violin()`.

```
interviews_plotting %>%
  ggplot(aes(x = respondent_wall_type, y = rooms)) +
  geom_violin(alpha = 0) +
  geom_jitter(alpha = 0.5, color = "tomato")
```

2. So far, we've looked at the distribution of room number within wall type. Try making a new plot to explore the distribution of another variable within wall type.
    - Create a boxplot for `liv_count` for each wall type. Overlay the boxplot layer on a jitter layer to show actual measurements.

```
interviews_plotting %>%
    ggplot(aes(x = respondent_wall_type, y = liv_count)) +
    geom_boxplot(alpha = 0) +
    geom_jitter(alpha = 0.5, width = 0.2, height = 0.2)
```

3. Add color to the data points on your boxplot according to whether the respondent is a member of an irrigation association (`memb_assoc`).

```
interviews_plotting %>%
  ggplot(aes(x = respondent_wall_type, y = liv_count)) +
  geom_boxplot(alpha = 0) +
  geom_jitter(aes(color = memb_assoc), alpha = 0.5, width = 0.2, height = 0.2)
```

Set 3

1. Create a bar plot showing the proportion of respondents in each village who are or are not part of an irrigation association (`memb_assoc`). Include only respondents who answered that question in the calculations and plot. Which village had the lowest proportion of respondents in an irrigation association?

```
percent_memb_assoc <- interviews_plotting %>%
  filter(!is.na(memb_assoc)) %>%
  count(village, memb_assoc) %>%
  group_by(village) %>%
  mutate(percent = (n / sum(n)) * 100) %>%
  ungroup()

percent_memb_assoc %>%
   ggplot(aes(x = village, y = percent, fill = memb_assoc)) +
    geom_bar(stat = "identity", position = "dodge")
```

Set 4

1. Experiment with at least two different themes. Build the previous plot using each of those themes. Which do you like best?

Set 5

1. With all of this information in hand, please take another five minutes to either improve one of the plots generated in this exercise or create a beautiful graph of your own. Use the RStudio ggplot2 cheat sheet for inspiration. Here are some ideas:

    - See if you can make the bars white with black outline.
    - Try using a different color palette (see http://www.cookbook-r.com/Graphs/Colors_(ggplot2)/).

Technical Tips and Tricks

Show how to use the 'zoom' button to blow up graphs without constantly resizing windows.

Sometimes a package will not install. You can try a different CRAN mirror:

- Tools > Global Options > Packages > CRAN Mirror

Alternatively you can go to CRAN and download the package and install from ZIP file:

- Tools > Install Packages > set to 'from Zip/TAR'

It is important that R, and the R packages be installed locally, not on a network drive. If a learner is using a machine with multiple users where their account is not based locally this can create a variety of issues (this often happens on university computers). Hopefully the learner will realize these issues beforehand, but depending on the machine and how the IT folks that service the computer have things set up, it may be very difficult to impossible to make R work without their help.

If learners are having issues with one package, they may have issues with another. It's often easier to make sure they have all the needed packages installed at one time, rather than deal with these issues over and over. Here is a list of all necessary packages for these lessons.

`|` **character on Spanish keyboards:** The Spanish Mac keyboard does not have a `|` key. This character can be created using:

`` `alt` + `1` ``

Other Resources

If you encounter a problem during a workshop, feel free to contact the maintainers by email or open an issue.

For a more in-depth coverage of topics of the workshops, you may want to read "R for Data Science" by Hadley Wickham and Garrett Grolemund.

Adapted from The Carpentries' Instructors Guide.