R Extension Workshop: Visualising Data

February 2019, Macquarie University

# Link to this document:

<http://bit.ly/MQ-r-plot-Feb19>

# Helpful links

* **Workshop webpage:** <https://mq-software-carpentry.github.io/2019-02-22-plotting-r/>
* **R Lesson:** <https://humburg.github.io/r-socialsci-git/>
* **Git:**
  + Workshop GitHub repository: <https://github.com/humburg/r-ggplot-project>
* **ggplot reference:** <https://ggplot2.tidyverse.org/reference/index.html>
* **RStudio Cheat Sheets:** <https://www.rstudio.com/resources/cheatsheets/>
  + ggplot cheat sheet: <https://github.com/rstudio/cheatsheets/raw/master/data-visualization-2.1.pdf>
* **ggplot Extensions:** <https://www.ggplot2-exts.org>
* **Stackoverflow:** <https://stackoverflow.com/questions/tagged/r>

# 

# Version control with Git

## Setting up Git

#### User

$ git config **--global** user.name "Vlad Dracula"  
$ git config **--global** user.email "[vlad@tran.sylvan.ia](about:blank)"

#### Text Editor

$ git config **--global** core.editor "nano -w"

#### Line Endings for Mac

$ git config **--global** core.autocrlf input

#### Line Ending for Windows:

$ git config **--global** core.autocrlf true

# Git Hub

### Sign Up

<https://github.com>

Register with your university email address and request an education discount here:

<https://education.github.com/>

GitHub requires you to verify your email address before you can use it. If you didn’t receive a confirmation email it may be stuck in the spam filter. Go to Mime Cast to release it:

[**Mimecast.com**](https://www.mimecast.com/) **-> login -> OneID**

# Questions, Comments, and Feedback

# 

# Code of Conduct

<https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html>

To summarise:

All participants in our events and communications are expected to show respect and courtesy to others. All interactions should be professional regardless of platform: either online or in-person. In order to foster a positive and professional learning environment we encourage the following kinds of behaviours in all Carpentries events and platforms:

* Use welcoming and inclusive language
* Be respectful of different viewpoints and experiences
* Gracefully accept constructive criticism
* Focus on what is best for the community
* Show courtesy and respect towards other community members

# 

# Schedule

* 9:00 Introductions
  + Who are you?
  + What are your interests / potential applications from this course?
* 9:10 Using Git and GitHub with RStudio
* 9:40 Introduction to ggplot
* 10:30 Catered Coffee
* 10:45 Visualising Data Part 1
* 12:15 Catered Lunch break
* 13:00 Visualising Data Part 2
* 14:30 Catered Coffee
* 14:45 Customising ggplot output
* 16:15 Wrap-up

# 

# Lessons

## Data

This curriculum uses the same dataset throughout.

**Studying African Farmer-led Irrigation (SAFI) dataset**

The [SAFI Project](http://www.safi-research.org/) is a research project looking at farming and irrigation methods used by farmers in Tanzania and Mozambique. This dataset is composed of survey data relating to households and agriculture in Tanzania and Mozambique. The survey form was created using the ODK (Open Data Kit) software via an Excel spreadsheet. This is used to create a form which can be downloaded and displayed (and completed) on an Android smartphone. The results are then sent back to a central server. The server can be used to produce the collected data in both JSON and CSV formats. We will use a sample of the collected data in CSV format throughout this workshop. The data can be [downloaded from Figshare](https://figshare.com/articles/SAFI_Survey_Results/6262019).

[More information on this dataset](https://datacarpentry.org/socialsci-workshop/data)

### Data visualisation with ggplot2

#### Plotting with ggplot

##### Exercise 1

We would like to inspect the data for a potential relationship between the number of people living in a household (no\_members) and the number of items owned by that household (number\_items). What would be a good way to visualise these data?

#### Building your plots iteratively

##### Exercise 2

Because many households share the same values for size and number of items a lot of points in our plot end up on top of each other. That makes it hard to see how common a given combination is. How could we modify the plot to avoid this problem?

##### Exercise 3 :)

It may be possible to enhance the appearance of the plot further by changing the size of the points. Use the size argument to change the point size. Combine this with the width and height arguments to adjust the jitter.

Try a few different values. Which one do you like best?

##### Exercise 4

Add separate regression lines for each village.

**Hint:** Asking *ggplot* to map village to an aesthetic when drawing the regression lines will produce one line per village.

##### Exercise 5

Commit this revised version of the plot to git and push your changes to your GitHub repository.

Go to GitHub and find the your last commit. Examine the differences between the versions of your plot in the diff viewer.

##### Exercise 6

Use what you just learned to create a scatter plot of rooms by village with the respondent\_wall\_type showing in different colors. Is this a good way to show this type of data?

#### Histograms and density plots

##### Exercise 7

It might be more interesting to look at household size by village. Create a plot with separate histograms for each village.

**Hint:** Use fill to get different colored bars.

#### Barplots

##### Exercise 8

So far you’ve looked at how the different wall types are distributed across the villages. A more interesting question might be what the makup of wall types within each of the villages is and how they compare to each other.

Create a plot that shows the proportion of wall types by village. What differences do you notice?

#### Boxplot

##### Exercise 9

Consider the *cement* buildings in the boxplot above. Judging from this plot alone, can you tell how many buildings with cement walls there are in the dataset?

##### Exercise 10

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.
* Add color to the data points on your boxplot according to whether the respondent is a member of an irrigation association (memb\_assoc).

##### Exercise 11

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 of rooms by wall type with a violin plot; see geom\_violin().

#### Customising plots

#### Exercise 12

Adjust the legend title and wall type labels as you did before. Which scale\_\*() function do you have to use to adjust the labels now?

#### Exercise 13

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

#### Exercise 14

You already know how to customise things like axis labels and point sizes. Change your plot to incorporate those changes.

#### Exercise 15

Take a look at the [**ggplot2** cheat sheet](https://github.com/rstudio/cheatsheets/raw/master/data-visualization-2.1.pdf), or the [**ggplot2** reference](https://ggplot2.tidyverse.org/reference/index.html). Can you figure out how to change the font size.

#### Exercise 16

With all of this information in hand, take some time to explore the dataset. Can you identify an aspect of the data that may be interesting to visualise?

Create an informative and visually appealing plot that showcases this aspect of the data. The [ggplot cheat sheet](https://github.com/rstudio/cheatsheets/raw/master/data-visualization-2.1.pdf) and the [R graph gallery](https://www.r-graph-gallery.com/) may provide some inspiration.

Here are some ideas you might explore:

* Take another look at the scatterplot of household size and number of items owned you created this morning. Can you use geom\_count() to improve this plot?
* Create a plot showing how often respondents have been involved in conflicts with other irrigators (affect\_conflicts). Does this differ between those that are members of an irrigation association and those that aren’t?
  + What type of plot is best suited for this?
  + Is the order in which ggplot() presents the factor levels appropriate? How could you change that?
* To what extent does the number of months a household has not had sufficient food during the last year (number\_month\_lack\_food) affect the number of meals members of the household have per day?
  + Does this differ between villages?
  + Note that the responses for number\_month\_lack\_food are always recorded as whole months. Can you adjust the grid lines in the plot to only occur at values that are valid responses?

**Note:** Feel free to transform the data or compute derived variables as necessary.