Data Visualization in Practice

Karl Ho

2021-01-11

Contents

Preface: Prerequisites for course

This book provides training materials for data visualization and creation of professional data charts using open source software. It requires no prior experience in data programming. Yet, if you have some programming experience in statistical software (e.g. R, SAS, SPSS and Stata), it will be helpful. R is the main language used in this course and the primary IDE (Integrated Development Environment) is RStudio. Students are encouraged to install the open-sourced software on own computer on MacOS, Linux or Windows operating systems. Mobile operating systems are not supported. Alternatively, they can use the cloud version of RStudio (https://Rstudio.cloud) using a Google or GitHub account.

0.1 Recommended software and IDE's

1. R version 4.x (https://cran.r-project.org)
2. RStudio version 1.3.x (https://www.rstudio.com)

0.2 Course book

An online book is created to provide materials and sample programs for this course. The link is https://datageneration.org/datavisualizationinpractice/book. It will be updated by module/day progress of the course.

0.3 GitHub

1. GitHub account (https://github.com/join)

GitHub is a repository hosting service for version control and most often nowadays hosting and managing development projects. Developers and educators use GitHub to share program codes and newly developed programs. It is free and has many features that facilitate the learning process. Users can create free account and clone projects for practices or co-development. In this course, the instructor will share codes or sample programs in class GitHub at Data Visualization in Practice.

0.4 Cloud websites/accounts

1. RStudio Cloud (https://rstudio.cloud)

The class demonstrations for hands-on workshops will be mostly using RStudio Cloud. Some advanced programs will be shown in RStudio installed locally. It is highly recommended to have RStudio installed on your local computer and users will switch between applications for practices. To do so, hold the Command key (MacOS) or Alt key (Windows) then then Tab key to perform switching.

0.4.1 Programming in RStudio Cloud

To start using RStudio Cloud,

1. Use Google or GitHub account to login to the RStudio Cloud website
2. RStudio Cloud employs own cloud services and storage for users. The free plan is limited to 15 projects and 15 project hours.
3. The advantages of using RStudio Cloud include sharing projects, cloud resources and most importantly it is on common platform (controlling for differences account different versions and operating systems)
4. Add a new space to start RStudio Cloud and you are ready to run codes and visualize data!

Chapter 1

Introduction

1.0.1 What is Data Visualization?

Data visualization is to deliver a message from your data. It is like telling a story using the chart or data applications. Sometimes the data is huge or the story to too long to tell. Visualization provides an ability to comprehend huge amounts of data. The important information from more than a million measurements is immediately available.

Visualization often enables problems with the data to become immediately apparent. A visualization commonly reveals things not only about the data itself but also about the way it is collected. With an appropriate visualization, errors and artifacts in the data often jump out at you. For this reason, visualizations can be invaluable in quality control.

Visualization facilitates understanding of both large-scale and small-scale features of the data. It can be especially valuable in allowing the perception of patterns linking local features.

Visualization facilitates hypothesis formation, inviting further inquiries into building a theory (Colin Ware 2012). It is exploratory data analysis (EDA) but can also provide the tools for hypothesis confirmation.

1.0.2 Learn to read data

Edward Tufte is one of the earliest data scientists emphasizing visual thinking. He postulates that one should first learn to read data, before moving on to visualize. He suggests training the visual thinking, then preparing the educated eyes. His newest book is titled SEEING WITH FRESH EYES: MEANING, SPACE, DATA, TRUTH, vividly testifying his philosophy of connecting the human perception with the data message.


(Source: Keegan Peterzell, CC BY-SA 4.0 https://creativecommons.org/licenses/by-sa/4.0, via Wikimedia Commons)

For Tufte, number one thing to learn about data visualization is to discard the default.

"If you're not doing something different, you're not doing anything at all." - Edward Tufte

Example: Multiple dimensions of data

1.1 Know your data

1. Data Literacy
2. Data types
3. Visual Vocabulary

1.1.1 Data Literacy

Before generating a chart or a data production for visualization, it is imperative you build your data literacy. This can be developed in three areas:

1. Data generaton process
2. Grammer of Graphics
3. Statistical Judgment

This course focuses on hands-on applications using real world data. It is advised students bring own dataset for practices. Sample programs will demonstrate how to import the data into R using base or add-on packages (e.g. foreign, haven, vroom).

1.1.2 Hands-on workshop: Data programming

1.1.2.1 Data Programming

This session starts with basic principles for data programming or coding involving data. Data programming is a practice that works and evolves with data. Data programming or coding allows the user to manage and process data in more effective manner. Programs are designed to be replicated or replicable by user and collaborators. A data program can be developed and updated iteratively and incrementally. In other words, it is building on the culminated works without repeating the steps. It takes debugging, which is the process of identifying problems (bugs) but, in fact, updating the program in different situations or with different inputs when used in different contexts, including the programmer himself or herself working in future times.

1.1.2.2 Getting started with RStudio Cloud

RStudio is the most popular IDE for programming in R. It is very powerful and versatile, providing an environment for data science functions including managing, modeling and visualizing data. Above all those, RStudio is also equipped with tools to interface with other languages such as Python, Stan and SQL and creating a variety of output products such as LaTeX documents for press-ready publication and web-ready html files for web publication. This document is also prepared and rendered inside RStudio.

RStudio Cloud allows processing R programs in cloud using a browser only. This will free users from using local installation, which could sometimes be subject to hardware and/or operating system limits.

Steps to start RStudio Cloud:

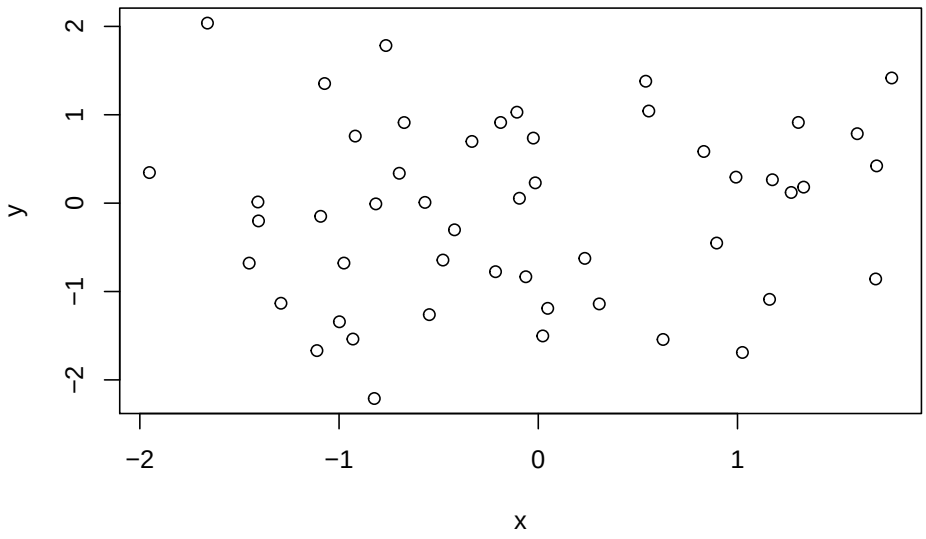
1. Use Google or GitHub account to login to the RStudio Cloud website
2. At left menu, click on Your Workspace under Spaces.
3. Start a New Project (blue button on right) Project. Be sure to name your new project. A new R session will open.
4. You are ready to code and visualize data!

1.1.2.3 Data programming in R

R basics

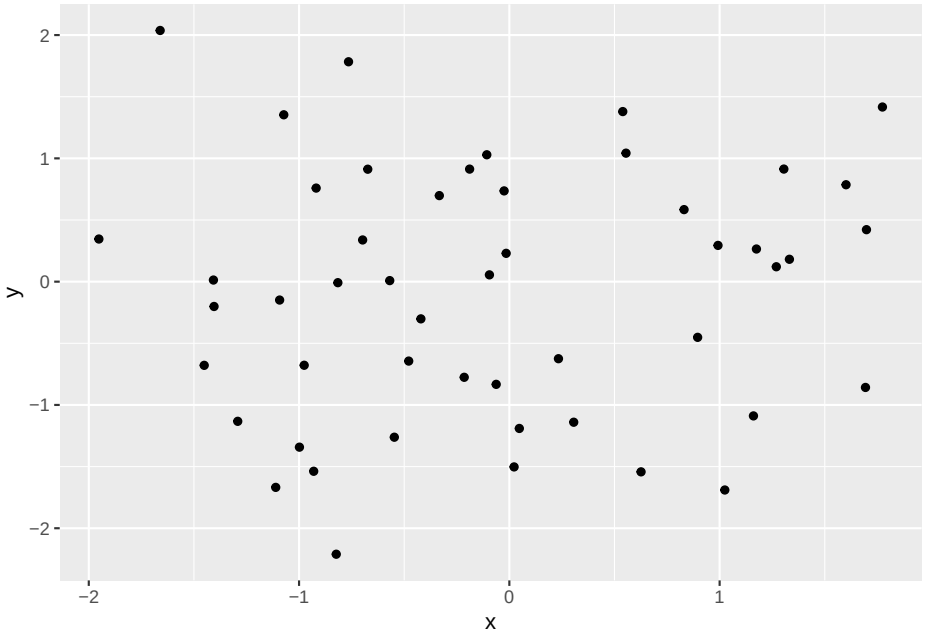
```
# Create variables composed of random numbers using the rnorm function
x <-rnorm(50)
y = rnorm(x)
```

```
# Plot the points in the plane
plot(x, y)
```



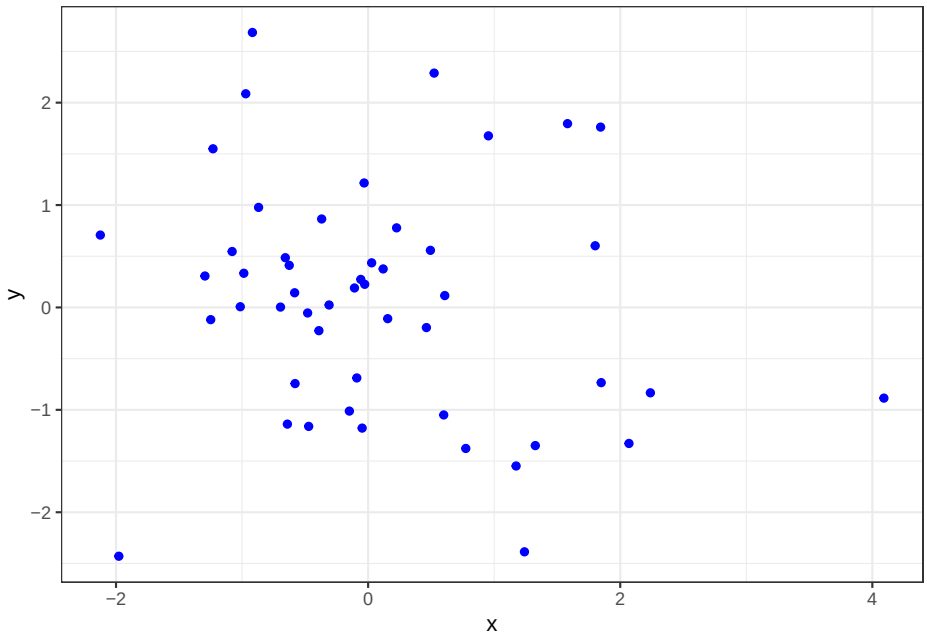
1.1.2.4 Using R packages

```
# Plot better, using the ggplot2 package
## Prerequisite: install and load the ggplot2 package
## install.packages("ggplot2")
library(ggplot2)
qplot(x,y)
```



1.1.3 ggplot2 starter

```
# Plot better better with ggplot2
x <- rnorm(50)
y = rnorm(x)
ggplot(,aes(x,y)) + theme_bw() + geom_point(col="blue")
```



1.2 Gallery

1.2.0.1 Recommended R Resources:

- The R Journal
- Introduction to R by W. N. Venables, D. M. Smith and the R Core Team
- Introduction to R Seminar at UCLA
- Getting Started in Data Analysis using Stata and R by Data and Statistical Services, Princeton University

Chapter 2

Functional approach

In this module, we will emphasize on hands-on applications, including building visual vocabulary, deciding chart types by function and data types and building charts using sample programs.

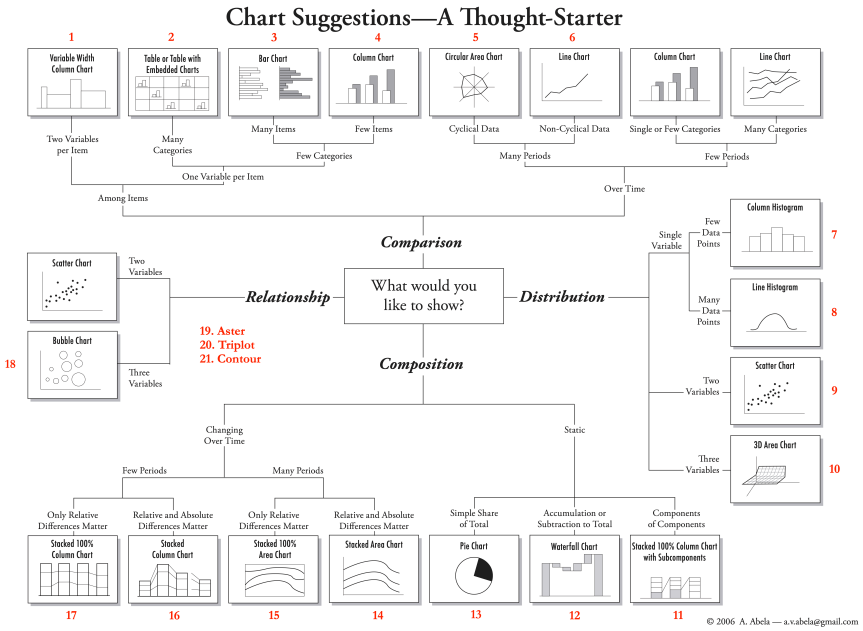


Figure 2.1: Starter: What to plot and How

Chapter 3

Interactive Charts

We describe building interactive and reactive data visualization for web publication.

3.1 ggiraph package

3.2 Leaflet for interactive map

3.3 Dash

3.4 Shiny

Chapter 4

Summary

In this brief three-day course, we have covered:

- Basics of data visualization
- Building charts by functions (and data types)
- Generating interactive data products