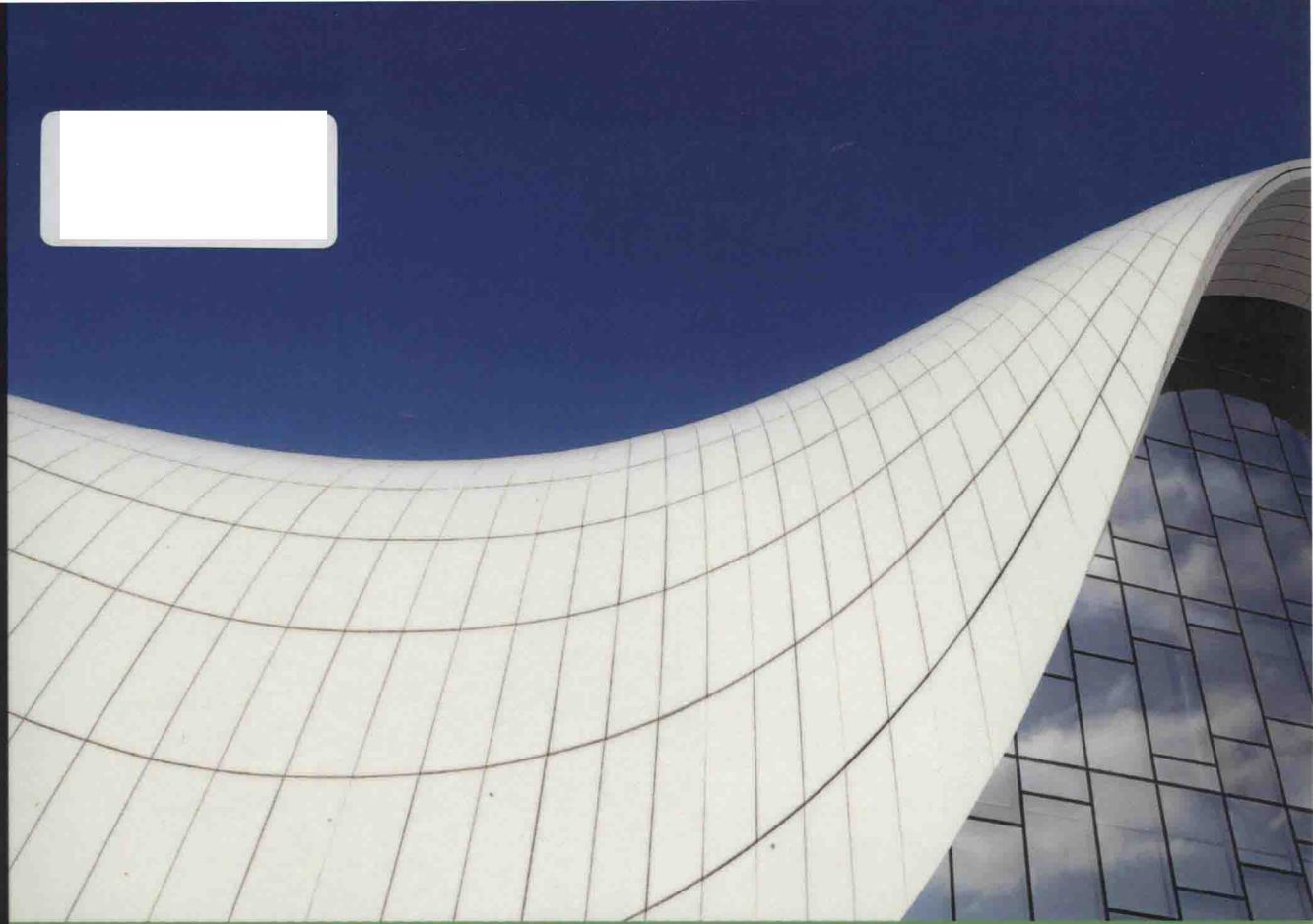

Quick answers to common problems

# R Data Visualization Cookbook

Over 80 recipes to analyze data and create stunning visualizations with R

Atmajitsinh Gohil

[PACKT] PUBLISHING open source* community experience distilled

# R Data Visualization Cookbook

Over 80 recipes to analyze data and create stunning visualizations with R

**Atmajitsinh Gohil**



[PACKT] PUBLISHING
open source
community experience distilled

BIRMINGHAM - MUMBAI

# R Data Visualization Cookbook



First published: January 2015

Production reference: 1240115

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham B3 2PB, UK.

ISBN 978-1-78398-950-8

www.packtpub.com

# Credits

**Author**

Atmajitsinh Gohil

**Reviewers**

Sharan Kumar Ravindran

Kannan Kalidasan

Erik M. Rodríguez Pacheco

Arun Padmanabhan

Juan Pablo Zamora

Patric Zhao

**Commissioning Editor**

Kartikey Pandey

**Acquisition Editor**

Neha Nagwekar

**Content Development Editor**

Arun Nadar

**Technical Editors**

Rohit Kumar Singh

Mitali Somaiya

**Copy Editors**

Nithya P

Shambhavi Pai

Rashmi Sawant

**Project Coordinator**

Neha Bhatnagar

**Proofreaders**

Simran Bhogal

Stephen Copestake

Paul Hindle

Joanna McMahon

**Indexer**

Priya Sane

**Graphics**

Disha Haria

Abhinash Sahu

**Production Coordinator**

Nilesh R. Mohite

**Cover Work**

Nilesh R. Mohite

# About the Author

**Atmajitsinh Gohil** works as a senior consultant at a consultancy firm in New York City. After graduating, he worked in the financial industry as a Fixed Income Analyst. He writes about data manipulation, data exploration, visualization, and basic R plotting functions on his blog at `http://datavisualizationineconomics.blogspot.com`.

He has a master's degree in financial economics from the State University of New York (SUNY), Buffalo. He also graduated with a master of arts degree in economics from University of Pune, India. He loves to read blogs on data visualization and loves to go out on hikes in his free time.

---

This book would not have been possible without the help from numerous data visualizers and data scientists around the globe who bring into existence new and innovative ways to transform data into beautiful stories. I would like to sincerely thank the developers of R and R packages who have contributed so generously to the growing R open source community.

I would like to thank Jer Thorpe and Hans Rosling for their inspiring Ted videos on data visualization.

I would also like to thank all the economists and statisticians who have so often inspired me.

I would like to thank my publisher, Packt Publishing, for giving me the opportunity to work on this book. I would also like to thank all the technical reviewers and content development editors at Packt Publishing for their informative comments and suggestions.

Finally, I would like to thank my amazing family and magnificent friends for always encouraging and supporting me.

---

# About the Reviewers

**Sharan Kumar Ravindran** is a lead data scientist in the fastest growing big data start-up based in Bangalore. His primary interests lie in statistics and machine learning. He has over 4 years of experience and has worked in the domains of e-commerce and IoT.

He has solved several problems on Kaggle and is among the top 10 percent of experts on Kaggle. His blog and social profiles can be found at `www.rsharankumar.com`.

He works for Flutura, which is ranked among the top 20 most promising big data companies across the globe by the leading analyst magazine, *CIO Review*. Flutura also featured on Gigaom reports on big data and M2M in the energy sector. Flutura was also the winner at TechSparks, where 800 innovative start-ups were evaluated.

**Kannan Kalidasan** is a software developer by profession, an autodidact, and open source evangelist. He has a decade's experience in database computing, data management, open source, and distributed computing. He holds a bachelor's of technology degree in computer science from Pondicherry University. He has played different roles in his career, such as a developer, an architect, a team lead, and a DBA. He currently holds the position of a BI Engineer at Orbitz Worldwide.

He started his own start-up back in 2005 on a part-time basis during his college days, worked with other companies in different open source projects, and provided training. He is passionate about technology and an entrepreneur at heart, and he likes to mentor fellow enthusiasts. His inherent curiosity keeps him occupied with learning new technologies and trying new things. He always believes that "our dreams can be delayed but will never fail if we work hard."

He blogs at `www.kannandreams.wordpress.com` and you can follow him on Twitter at `@kannanpoem`. He loves to take long walks alone, write Tamil poems, paint, and read books.

---

A big thank you to all who believed in me and supported me. I would like to thank my strong soul for pushing me to achieve my dreams. I would like to express my deepest gratitude to Packt Publishing for giving me this opportunity.

---

**Erik M. Rodríguez Pacheco** works as a manager in the Business Intelligence Unit at Banco Improsa in San José, Costa Rica. He has 11 years of experience in the financial industry. He is currently a professor of the Business Intelligence Specialization Program at the Instituto Tecnológico de Costa Rica's Continuing Education Program. Erik is an enthusiast of new technologies, particularly those related to business intelligence, data mining, and data science. He holds a Bachelor's degree in business administration from Universidad de Costa Rica, and has specialized in business intelligence from the Instituto Tecnológico de Costa Rica, data mining from Promidat (Programa Iberoamericano de Formación en Minería de Datos), and business intelligence and data mining from Universidad del Bosque, Colombia. He is currently enrolled in a specialization program in data science from Johns Hopkins University. He can be reached at `cr.linkedin.com/in/erikrodriguezp/`.

**Arun Padmanabhan** has about 4 years of experience in developing products including mobile, enterprise, statistical, and data mining applications. He graduated with a master's degree in computer applications in 2010. Currently, he is a data scientist at Flutura Decision Science and Analytics, where he is working at saving the world, one data product at a time.

**Juan Pablo Zamora** holds a bachelor's degree in statistics from the University of Costa Rica (UCR) in 2007. He is currently working on his dissertation in the field of predictive analytics and will obtain an MSc degree in statistics from the University of Costa Rica.

He enjoys teaching and was a tutor of statistics courses at the Business School of UNED of Costa Rica during 2010-2012. He also mentored others in the areas of data processing and analytics as well as the use of statistical analysis tools, to name a few.

Juan has over 7 years of experience in the banking industry, primarily in the credit card business for Central America and Mexico. He began as an analyst, eventually becoming the leader of a team of analysts for Central America's largest credit card issuer and acquirer. During this period, he participated in several predictive analytics projects related to credit risk, account retention, and profitability.

Juan recently joined a large multinational company in the retail sector with the task of building an analytics program to identify and prevent high-risk issues and/or threats to the business in Latin America.

His current interests are R, data visualization, business intelligence, predictive modeling, and big data. He can be reached at `cr.linkedin.com/in/datasciencezamora` or data.science.zamora@gmail.com.

**Patric Zhao** is a senior GPU architect in the High Performance Computing (HPC) field at Nvidia. He has experience in developing scientific and engineering applications and focuses on parallelization, performance modeling, and architecture-specific tuning. Patric is currently working on big data and machine learning areas, including regression, neural network, recommending system design, and implementation in CPU and GPU architectures. Patric has also contributed to accelerate R's applications by CUDA in the GPU ecosystem. You can find related articles on Nvidia's blog at `http://devblogs.nvidia.com/parallelforall/author/patricz/` or write to him at `patric.zhao@gmail.com`.

---

I would like to really thank my wife Yan Li J for always supporting and encouraging me.

---

# www.PacktPub.com

## Support files, eBooks, discount offers, and more

For support files and downloads related to your book, please visit www.PacktPub.com.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.PacktPub.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at service@packtpub.com for more details.

At www.PacktPub.com, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.


https://www2.packtpub.com/books/subscription/packtlib

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can search, access, and read Packt's entire library of books.

## Why Subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

## Free Access for Packt account holders

If you have an account with Packt at www.PacktPub.com, you can use this to access PacktLib today and view 9 entirely free books. Simply use your login credentials for immediate access.

# Table of Contents

# Preface

Our ability to generate data has improved tremendously with the advent of technology. The data generated has become more complex with the passage of time. The complexity in data forces us to develop new tools and methods to analyze it, interpret it, and communicate with the data. Data visualization empowers us with the necessary skills required to convey the meaning of underlying data. Data visualization is a remarkable intersection of data, science, and art, and this makes it hard to define visualization in a formal way; a simple Google search will prove me right. The Merriam-Webster dictionary defines visualization as *"formation of mental visual images"*. In reality, the term visualization goes beyond the limits of providing visual images by assisting humans in data recording, revealing pattern, exploration of data, and spreading information in a meaningful way.

Jer Thorpe in an interview with Mashable.com (`http://mashable.com/2012/12/11/data-visualization-jer-thorp/`) introduces the idea of humanizing data:

> *"...And I think that there's a huge possibility for humans, society as a whole—if we could share that data more usefully, for science and for the construction of cities, and for all these kinds of things, then it becomes much more useful. So in my work, I'm really thinking about how we can give people glimpses into that type of future. Giving people an opportunity to think about data ownership or giving people a visualization so that they can see the kinds of things that can be done with data".*

R is an open source platform used to analyze data. It has been widely used as a statistical tool in the past. An individual does not necessarily have to be a programmer to use R. A beginner can use basic R functionalities to manipulate and extract data and create very simple and quick visualizations using the basic graphic tools. An intermediate R user can implement interactive visualizations, perform predictive modeling, or even create animated applications using packages developed by the R community. R will present you with the tools you need to process, manipulate, and communicate with your data, and it is not just limited to statistical analysis.

In this book, you will learn how to generate basic visualizations, understand the limitations and advantages of using certain visualizations, develop interactive visualizations and applications, understand various data exploratory functions in R, and finally learn ways of presenting the data to our audience. This book is aimed at beginners and intermediate users of R who would like to go a step further in using their complex data to convey a very convincing story to their audience.

# What this book covers

*Chapter 1, A Simple Guide to R*, is a quick tutorial on getting started with R. You will learn how to install packages, access help on R, construct and edit matrices, create and manipulate data frames, and write and save plots.

*Chapter 2, Basic and Interactive Plots*, introduces some of the basic R plots, such as scatter, line, and bar charts. We will also discuss the basic elements of interactive plots using the googleVis package in R. This chapter is a great resource for understanding the basic R plotting techniques.

*Chapter 3, Heat Maps and Dendrograms*, starts with a simple introduction to dendrograms and further introduces the concept of clustering techniques. The second half of this chapter discusses heat maps and integrating heat maps with dendrograms to get a more complete picture.

*Chapter 4, Maps*, discusses the importance of spatial data and various techniques used to visualize geographic data in R. You will learn how to generate static as well as interactive maps in R. The chapter discusses the topic of shape files and how to use them to generate a cartogram.

*Chapter 5, The Pie Chart and Its Alternatives*, is a detailed discussion on how to generate pie charts in R. You will also learn about the various criticisms of pie charts and how the pie chart is transformed to overcome them. The chapter also provides you with various alternatives used by data scientists and visualization artists to overcome the limitation of a pie chart.

*Chapter 6, Adding the Third Dimension*, dives into constructing 3D plots. This chapter also introduces packages such as rgl and animation, which are used to create interactive 3D plots.

*Chapter 7, Data in Higher Dimensions*, demonstrates the use of visualizations that are used to display data in higher dimension. You will learn the techniques to generate sunflower plots, hexbin plots, Chernoff faces, and so on. This chapter also discusses the usefulness of network, radial, and coxcomb plots, which have been widely used in news.

*Chapter 8, Visualizing Continuous Data*, illustrates the use of visualizations to display time series data. The chapter also discusses some general concepts related to visualizing correlations, the shape of the distribution, and detection of outliers using box and whisker plots.

*Chapter 9, Visualizing Text and XKCD-style Plots*, illustrates the use of text in creating effective visualizations. This chapter focuses mainly on techniques to create word clouds, phase tree, and comparison clouds in R. You will also learn how to use the XKCD package to introduce humor in visualizations.

*Chapter 10, Creating Applications in R*, shows you the techniques to create presentations and R markdown documents for publishing on a blog or a website. The chapter further discusses the XML package used to extract and visualize data as well as using shiny package used to create interactive applications.

# What you need for this book

You need to download R to generate the visualizations. You can download and install R using the CRAN website available at `http://cran.r-project.org/`. All the recipes were written using RStudio. RStudio is an integrated development environment (IDE) for R and can be downloaded from `http://www.rstudio.com/products/rstudio/`. Many of the visualizations are created using R packages and they are discussed in their respective recipes.

In few of the recipes, I have introduced users to some other open source platforms such as ScapeToad, ArcGIS, and Mapbox. Their installation procedures are outlined in their respective recipes.

# Who this book is for

Having studied economics, I am not a software programmer myself and have written this book for readers new to R and visualization. This book does not delvento complex R code or complex data manipulating techniques, and it is written keeping in mind new and intermediate R users interested in learning about data visualization and data exploration techniques.

The book aims at teaching you the implementation of interactive and animated data visualizations and not just the basic R techniques. However, I have introduced some basic functionalities in *Chapter 1, A Simple Guide to R* and *Chapter 2, Basic and Interactive Plots*.

Wherever possible, I have provided references to websites, blogs, and journals, which can be explored to learn more about specific functions, graphics, animations, or even basic functionalities in R.

# Sections

In this book, you will find several headings that appear frequently (Getting ready, How to do it, How it works, There's more, and See also).

To give clear instructions on how to complete a recipe, we use these sections:

### Getting ready

This section tells you what to expect in the recipe, and describes how to set up any software or any preliminary settings required for the recipe.

### How to do it...

This section contains the steps required to follow the recipe.

### How it works...

This section usually consists of a detailed explanation of what happened in the previous section.

### There's more...

This section consists of additional information about the recipe in order to make the reader more knowledgeable about the recipe.

### See also

This section provides helpful links to other useful information for the recipe.

## Conventions

In this book, you will find a number of styles of text that distinguish between different kinds of information. Here are some examples of these styles, and an explanation of their meaning.

Code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles are shown as follows: "We have used the `png()` function to save the plot as a PNG."

Any command line code is written as:

```
k = matrix((1:4),2,2)
l = matrix((5:10),2,3)
dim(k)
dim(l)
```

In R it is a general practice to use `<-` for assignment instead of the `=` sign. In all the recipes, I have followed the `=` sign for assignment. You should note that if you refer to blogs or websites related to R, you may encounter the `<-` sign in the code files.