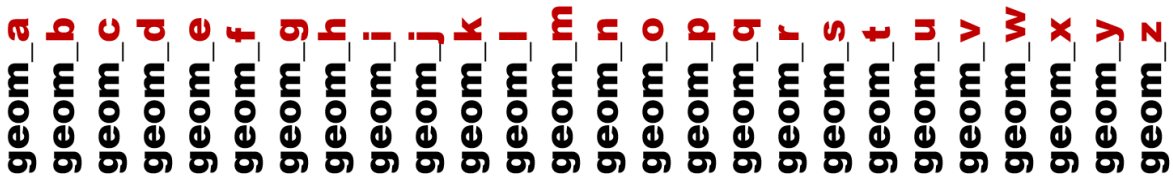
Data Visualisation Geometries Encyclopedia

Geoms in the Grammar of Graphics: All Types of Plots

Thiyanga S. Talagala

Table of contents

Welcome



Welcome to the "Data Visualisation `geom` Encyclopedia", an encyclopedia of geometrics use in data visualisation. This encyclopedia is a curated collection of `geom` available in different R programming software packages. This book can also be considered as an "Encyclopedia of Plots".

This book is a work in progress. The accompanying R package for drone is available on CRAN: https://github.com/thiyangt/drone

Accompanying R package: drone: Data foR geOm eNcyclopaEdia

Why choose “drone” as the package name?

This encyclopedia provides a comprehensive overview of `geoms` within the grammar of graphics framework, much like a drone offers a bird’s-eye view of a landscape.

To load this Data Visualisation Geom Encyclopedia via `drone` package use the following code.

```
install.packages(drone)
library(drone)
load_encyclopedia()
```

Preface

What can I find in this geom encyclopedia?

Let's begin by looking at what we mean by geom in data visualization. In this context, geoms (short for geometries in ggplot2 package) are the visual elements used to represent data in a plot. They define the type of chart such as point chart, line chart, bar chart, etc.

To illustrate the idea I use the following dataset which contains information related to 82 countires. The variable description is as follows (Table **??**):

The first eight rows of the dataset as well as the R code to load the dataset is given below:

```
library(drone)
library(tidyverse)
data("worldbankdata")
worldbankdata2021 <- worldbankdata |> filter(Year == 2021) |>
  filter( Income == "LM" | Income =="L") |>
  select( Country, Income, Electricity)
worldbankdata2021 |> head(8)
```

```
# A tibble: 8 x 3
  Country      Income Electricity
  <fct>        <fct>        <dbl>
1 Afghanistan  L             97.7
2 Angola       LM            48.2
3 Burundi      L             10.2
4 Benin        LM            42.0
5 Burkina Faso L             19.0
```

Table 1: Variable Description

Variable	Description
Country	Country name
Income	Income category in 2021: lower income (L), lower middle income (LM)
Electricity	Percentage of people access to electricity

```
6 Bangladesh   LM            99.0
7 Bolivia      LM            98.6
8 Bhutan       LM           100
```

Now, I want to visualize the relationship between Electricity and Income variables on the cartesian coordinate plane shown in Figure **??**

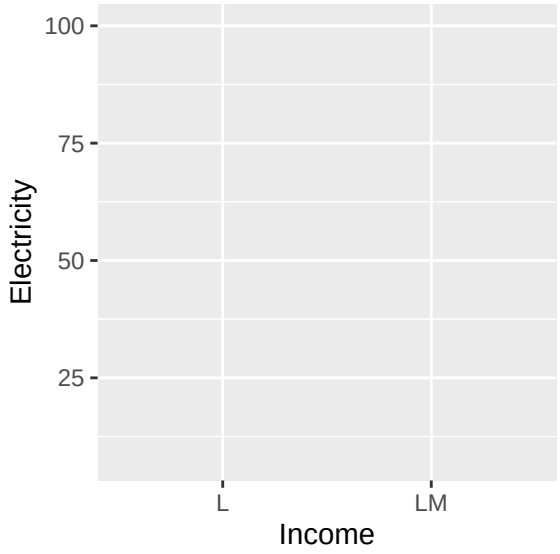


Figure 1: Cartesian coordinate plane showing Income (x-axis) and Electricity (y-axis).

Figure **??** shows 6 plotting types that I created to visualise the relationship between the variables `Electricity` and `Income` on the Cartesian plane Figure **??**.

In all panels of Figure **??** (a–h), the same dataset, variables, and Cartesian coordinate system are used. However, the chart types different.

Could you list the differences you observe in the charts?

1. The type of chart or shape used to depict data.
2. The statistics computed on the data to visualize on the chart.

When creating a chart, we first decide on the statistic we want to visualize. Next, we compute this statistic from the data, and finally, we use a suitable geometry to display the computed values of the statistic. Table **??** summarizes the statistical operations (or statistical transformations) performed on the data and the corresponding geometries used to visualize the computed statistics.

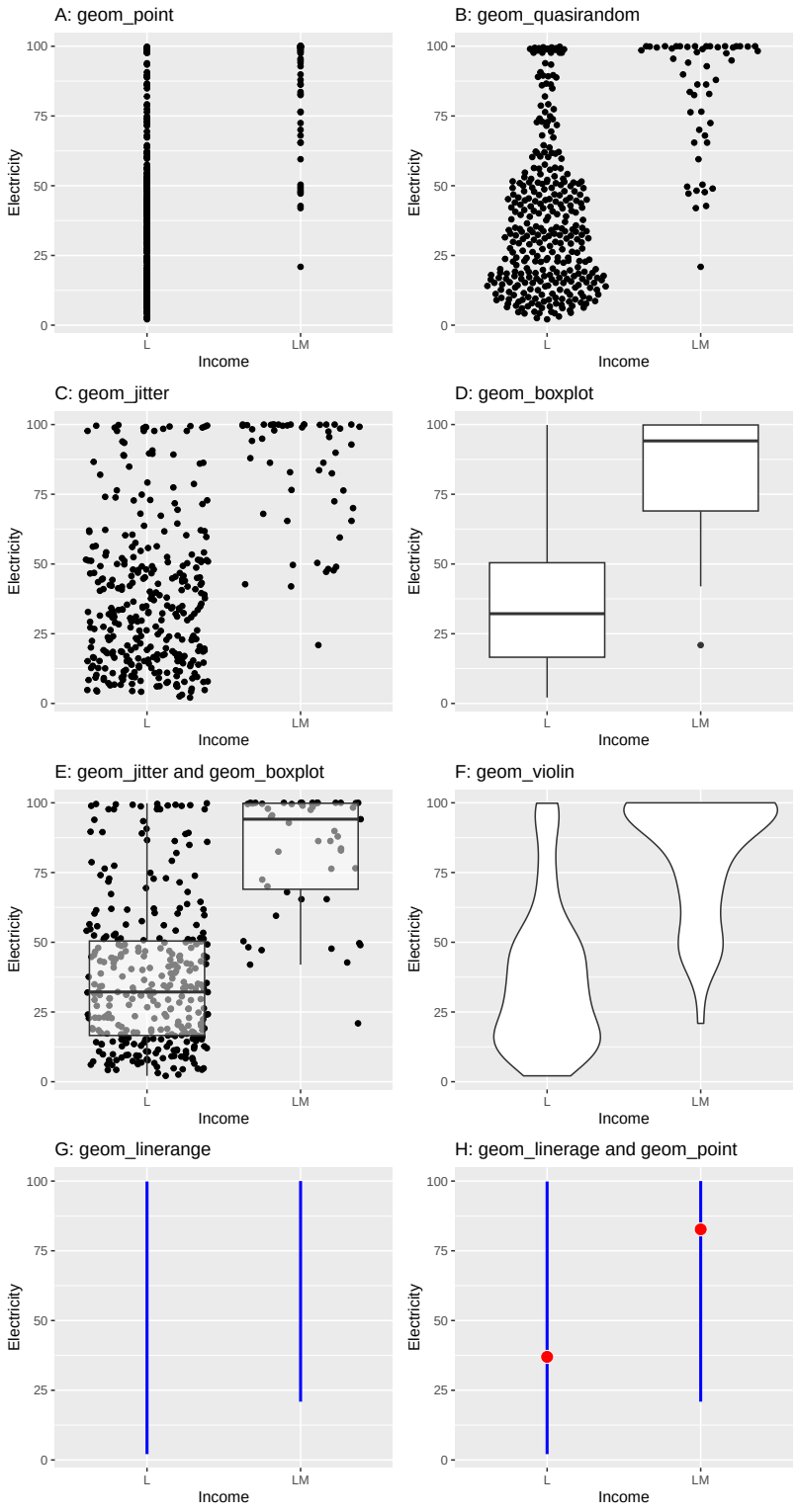


Figure 2: Six plot types used to visualize the relationship between income level (low income L, lower-middle income [LM]) and the percentage of people with access to electricity in 2021.

Table 2: Statistical transformations and geom/stat layers are used in the plots.

Plot	Statistical Operation	Visualisation Metho
A	Use individual data points	Point geometry
B	Use individual data points	Quasirandom geome
C	Use individual data points	Jitter geometry
D	Compute quartiles, Q1 $-$ 3 $\cdot$ IQR, and Q3 + 3 $\cdot$ IQR	Box-and-whisker plo
E	Use individual data points; compute quartiles, Q1 $-$ 3 $\cdot$ IQR, and Q3 + 3 $\cdot$ IQR	Jitter geometry and
F	Compute kernel density	Violin geometry
G	Compute minimum and maximum of data	Line geometry
H	Compute minimum, mean, and maximum	Line geometry (rang

This encyclopedia is a collection of geoms, in other words plot types that you can create using the ggplot2 or extensions of ggplot2 under the grammar of graphics framework. In other words Encyclopedia of Plots.

Motivation to write this book

The motivation behind writing this encyclopedia is, there is no centralized resource where all geoms can be viewed in one place. Additionally, no comprehensive book exists that catalogs the different types of plots available for data visualization. Having them in one place help data visualizers to craft more effective analyses and create new geoms. Further, this also helps to avoid duplicate efforts.

What you will learn?

In this geom encyclopedia you will learn different types of geoms and their applications. Furthermore, each geom has a set of aesthetics that it understands. These aesthetics can be divided into two parts: i) required aesthetics and ii) optional aesthetics. Further, every geom has a default stat; and every stat has a default geom.

To give you an idea about the associated aesthetics and stat, let's look at the plot shown in Figure **??**. In this case `x`, `y` are required aesthetics and `color size` and `alpha` is optional aesthetics. The statistics layer is identity since the data points are plotted as it is.

The aesthetic mappings, defined with `aes()`, describe how variables in the dataset are mapped to aesthetics (or visual properties of the plot). This `aes(x=Income, y=Electricity, color=Income)` is called mapping variables to the visual properties of the chart. This `alpha=0.5, size=2` is called setting values to visual properties.

```r
ggplot(worldbankdata2021, aes(x=Income, y=Electricity, color=Income)) +
  geom_point(alpha=0.5, size=2) +
  scale_color_brewer(palette = "Dark2")
```

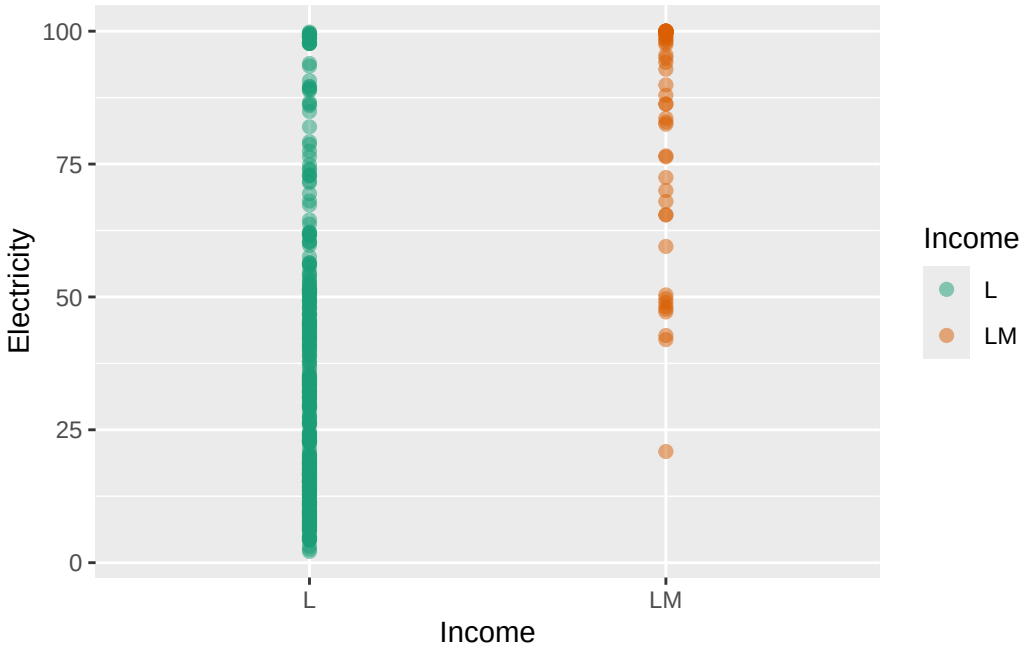


Figure 3: Distribution of Electricity by Income Category

In this geom encyclopedia, you will also learn about the required and optional aesthetics that each geom supports. Additionally, an example is provided demonstrating the application of each geom with reproducible R code.

In summary you will learn:

i. What each geom does The purpose of different geoms (e.g., points, lines, bars, boxes, tiles) and the types of data and questions they are best suited for.

ii. Required and optional aesthetics Which aesthetics (such as x, y, colour, fill, size, group) are required for each geom and how optional aesthetics change the appearance and meaning of a plot.

iii. Typical use cases When to use each geom—for example, comparing distributions, showing relationships, visualizing trends over time, or displaying uncertainty.

iv. Common variations and set parameters Key arguments (such as stat, position, alpha, width, binwidth) and how they affect the visual output.

v. Strategies to improve clarity of plots Strategies to overcome the frequent challenges such as overplapping, missing values, highlight trend, etc.

vi. Connections between geoms and statistics How geoms interact with statistical transformations (e.g., geom_histogram() with binning, geom_smooth() with model fitting).

```
vii. Connections between geom
Similar geoms and different combinations of geoms that can use to visualise data.
```

What you won't learn?

This book is not focused on teaching R programming fundamentals or providing a comprehensive guide to data visualization principles. It assumes you already have a basic understanding of R and ggplot2, and it will not cover how to start from scratch in these areas.

How this geom encyclopedia is organized?

As this is an encyclopedia, the chapters are organised according to the alphabetical order. However, within a chapter geoms are not organized according to the alphabetical order. At the beginning of each chapter, I have tabulated the geoms listed under that letter.

Audience

The Figure **??** shows my target audience for the book. In general for all data enthusiasts, this can be considered as a Encyclopedia of Plots. For R, tidyverse, gggplot2 users this can be considered as a Data Visualisation Geometries Encyclopedia.

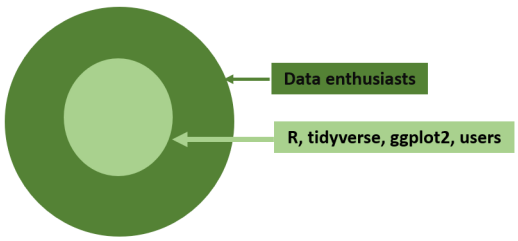


Figure 4: Target Audience

What prior knowledge is needed to use this geom encyclopedia?

1. **For a complete newcomer to get an idea about the possible charts that can be drawn for a data set.**

   None

2. **To get hands-on-experience with the examples provided you need to have following knowledge.**

   i. R and RStudio Integrated Development Environment (IDE)

   ii. Basics of R programming

   iii. tidyverse (Wickham et al. 2019) style of coding

   iv. Data wrangling with tidyr(Wickham, Vaughan, and Girlich 2024) and dplyr(Wickham et al. 2023)

   v. Knowledge in using the pipe operator: |>

   vi. Knowledge in creating data visualisations using the grammar of graphics frame work implemented in ggplot2(Wickham 2016)

If you want to learn more of them, following are the books recommended:

- R for Data Science (2e) written by Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemund.
- ggplot2: Elegant Graphics for Data Analysis (3e) written by Hadley Wickham, Danielle Navarro, and Thomas Lin Pedersen.

Acknowledgments

I would like to thank all the package authors and contributors who have developed and shared geoms as R packages.

I would like to thank all the students who took my ASP 460 2.0/STA 492 2.0 Data Visualization course in 2024 for their valuable contributions in exploring geoms with me.

This book was built with Quarto.

Declaration of generative AI in scientific writing

During the preparation of this work the author used ChatGPT in order to improve readability and language. After using this tool/service, the author reviewed and edited the content as needed and take full responsibility for the content of the published article.

Cite

Please use the below text and bibtex for citing the book.

```
T. S. Talagala, Data Visualisation Geometries Encyclopedia: Geoms in the
Grammar of Graphics - All Types of Plots. [Online]. Available: https://thiyangt.github.io/ge
[Accessed: Jan. 18, 2025
```

```
@online{talagala2024geoms,   author       = {Thiyanga S. Talagala},   title
= {Data Visualisation Geometries Encyclopedia: Geoms in the Grammar of Graphics
- All Types of Plots},   year         = {2024},   url          = {https://thiyangt.github.io/
note         = {Accessed: 2025-01-18} }
```

Please use the below text and bibtex for citing the associated R package drone.

```
Talagala T (2025). _drone: Data for   Data Visualisation Geometries   Encyclopedia_.
R package version   2.0.0,   <https://github.com/thiyangt/geom.encyclopedia>..
```

```
@Manual{drone,     title = {drone: Data for Data Visualisation Geometries
Encyclopedia},     author = {Thiyanga S. Talagala},     year = {2025},     note
= {R package version 1.0.0, commit 86d4fc19bb3a03da3eeb8f6748cb0bfc21dfdf72},
url = {https://github.com/thiyangt/drone},   }
```

Colophon

The field of data visualization is dynamic, and new techniques and visualizations may emerge over time. Hence, I will be regularly updating this encyclopedia to ensure it remains a relevant and comprehensive resource for users.

Data and setting up your workflow

The goal of this chapter is to provide readers with an overview of the datasets used in the book's examples. Having an initial understanding of the data helps readers easily navigate between the examples.

Installation of associated packages

To run the examples in the book you need to install the following packages. In addition, to this package list, the associated package corresponds to the geom should be installed. The `drone` package provides the datasets associated with this geom encyclopedia.

```
#install.packages(drone)
#install.packages("devtools")
devtools::install_github("thiyangt/drone")
install.packages(tidyverse)
install.packages(patchwork)
```

Data set use in the geom Encyclopedia

The datasets used in the book, main are from country-wise statistics obtained from public reliable websites. Only two datasets are used to explain all geoms so that readers can build intuition by seeing the same data represented in multiple ways. The reasons for using these datasets are:

- The dataset context is familiar and easily understood by individuals from diverse backgrounds.
- Ability to create cross-sectional, time-series, spatial, and spatio-temporal visualizations.
- Ability to visualize relationships between qualitative–qualitative, qualitative–quantitative, and quantitative–quantitative variables.
- Experience addressing common data challenges such as missing values, overplotting, and large-scale datasets.

WorldHappinessScore dataset

The dataset is obtained from World Population Review report (World Population Review (n.d.)). The World Happiness Score is reported on a 0–10 scale, with higher values indicating greater happiness. This dataset containes the world happiness score for 145 countries. The variable description is given in Table **??**.

The first few rows of the dataset is shown below

Table 3: Variable Description for WorldHappinessScore dataset

Variable	Description
flagCode	Country flag code
country	Country name
WorldHappinessScore_2024	World happiness score for 2024
WorldHappinessScore_2023	World happiness score for 2023
WorldHappinessScore_2022	World happiness score for 2022

```
library(drone)
head(WorldHappinessScore)
```

```
  flagCode     country WorldHappinessScore_2024 WorldHappinessScore_2023
1       FI     Finland                     7.74                     7.804
2       DK     Denmark                     7.58                     7.586
3       IS     Iceland                     7.53                     7.530
4       SE      Sweden                     7.34                     7.395
5       IL      Israel                     7.34                     7.473
6       NL Netherlands                     7.32                     7.403
  WorldHappinessScore_2022
1                    7.821
2                    7.636
3                    7.557
4                    7.384
5                    7.364
6                    7.415
```

`WorldHappinessScore`: Data Profiling

Table **??** provides a compact summary of the dataset.There are 5 variables. Among the 2 are character variables and 3 are numeric variables. Summary of both character variables and numeric variables are given in the output. For character variables,

`n_missing` tells how many values are missing for each variable.

`complete_rate` is the proportion of non-missing values.

`n_unique` is the number of unique values in the variable.

For numeric variables Table **??** shows mean, sd (standard deviation), p0 (min), percentiles: p25, p50 (median), p75, p100 (max), and a small histogram for each numeric variable.

Table 4: Summary description of the WorldHappinessScore dataset

Table 4: Data summary

	-
Name	WorldHappinessScore
Number of rows	145
Number of columns	5
_____	
Column type frequency:	
character	2