Variables & Visualization

What Is The Story You Are Trying To Tell?

Andy Grogan-Kaylor

2021-04-02

Contents

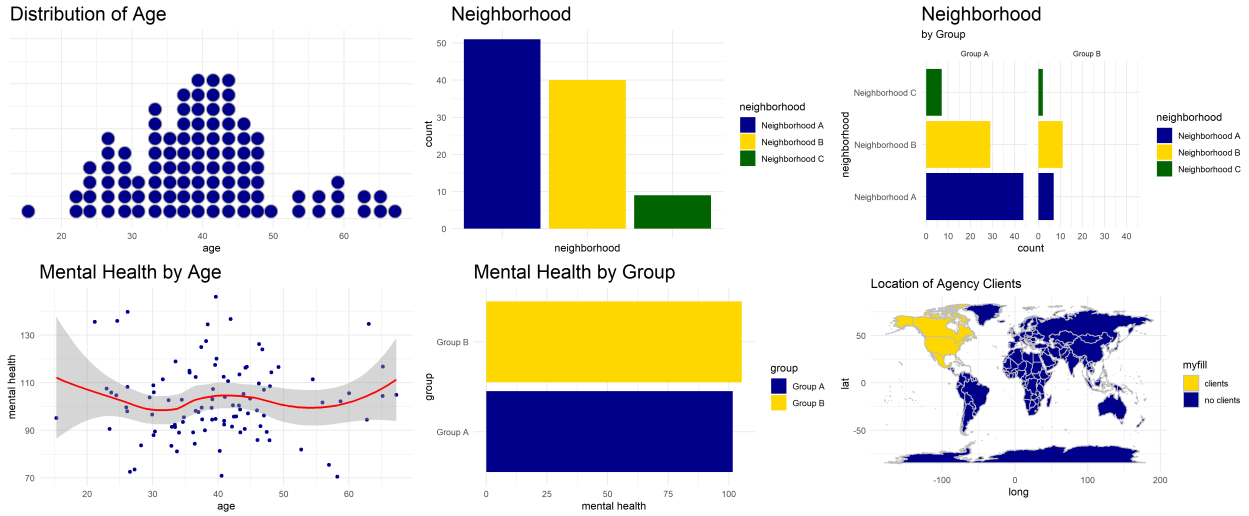


Figure 1: possible visualizations

1 Possibilities

2 Background

- Deciding upon the right data visualization to represent your data can be a daunting process.
- I believe that a *starting point* for this thinking is some basic statistical thinking about the *type* of variables that you have.
- At the broadest level, variables may be conceptualized as *categorical* variables, or *continuous* variables.

3 Data Often Come From A Survey Questionnaire.

4 What is Data?

A data set is nothing more than a series of rows and columns that contain answers to responses to a survey.

- Rows are usually used for individuals, while columns indicate the questionnaire answers, or measures, from those people.
- Answers to questions are often given numerical responses (e.g. “no” is frequently coded as “0” and “yes” is frequently coded as “1”)

Table 1: Hypothetical Data

person	Q1	Q2	Q3
1	1	0	100
2	2	0	200
3	1	1	-9

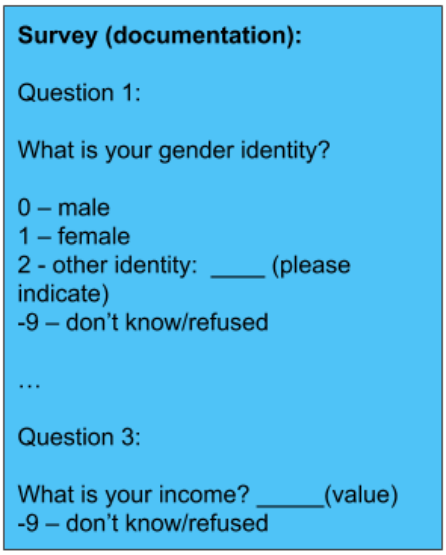


Figure 2: hypothetical questionnaire

4.1 Some Notes on Data

- In working through our research questions, we'll constantly be going back and forth between the actual data (to see the pattern of responses) and the documentation, to figure out the actual question asked as well as how the different responses are coded.
- Often in a spreadsheet, you'll see the full text of a question written out (e.g. "What is your gender identity"?)
- Most programs that work with data are going to want abbreviations (e.g. "Q1" or "gender"") for the questions. These abbreviations should usually have no spaces and be 8 characters or less.

4.2 Missing Data

- One cell of the sample data set has a negative number.
- Frequently negative numbers are used to indicate what are called "missing values". A missing value is a response like "don't know"" or "refused to answer" or "did not answer".
- Before we start doing calculations with our data, we'll want to change negative numbers to true missing values (usually symbolized by a "." or "NA", so that they don't goof up our calculations.)

5 What are Variables?

- By variables, I simply mean the columns of data that you have.
- For our purposes, you may think of variables as synonymous with questionnaire items, or columns of data.

5.1 Variable Types

- *categorical variables* represent unordered categories like *neighborhood*, or *religious affiliation*, or *place of residence*.

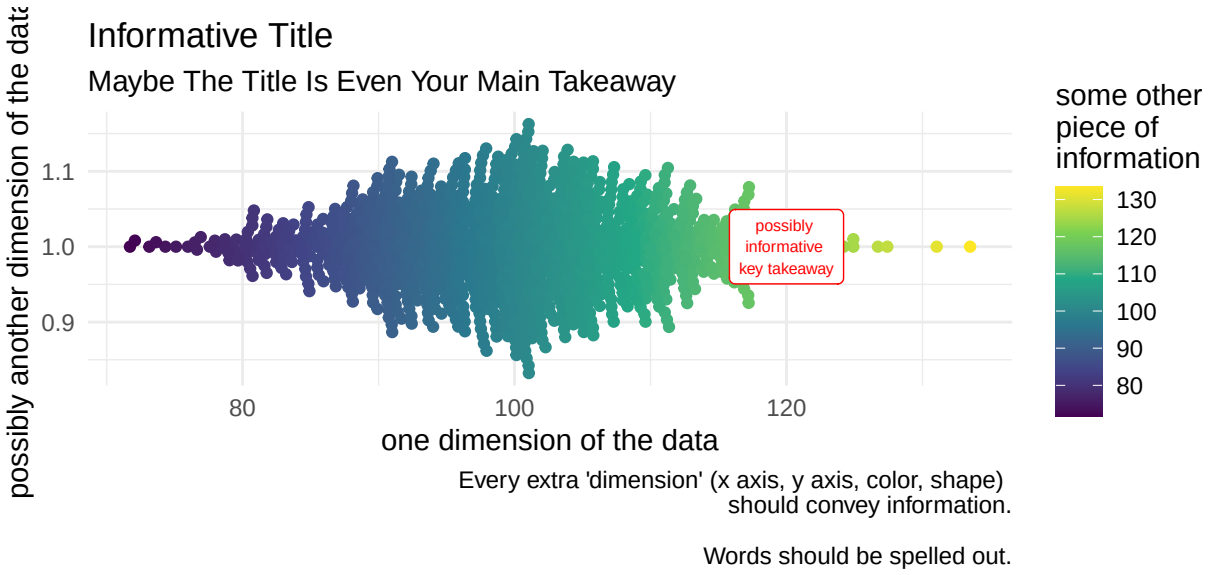
- *continuous variables* represent a continuous scale like a *mental health scale*, or a *measure of life expectancy*.

6 A Data Visualization Strategy

Once we have discerned the type of variable that have, there are two followup questions we may ask before deciding upon a chart strategy:

- Is our graph about **one thing at a time**?
  - How much of *x* is there?
  - What is the distribution of *x*?
- Is our graph about **two things at a time**?
  - What is the relationship of *x* and *y*?
  - How are *x* and *y* associated?

6.1 More On Strategy



7 Simulated Data

This example uses simulated data on social work clients, of the kind that a social service agency might collect.

Table 2: Simulated Data

age	mental_health	group	neighborhood
41.27	91.5	Group A	Neighborhood A
24.05	109	Group A	Neighborhood A
28.28	106.1	Group A	Neighborhood B
42.4	116.1	Group B	Neighborhood B
56.02	88.83	Group A	Neighborhood B

8 Show One Thing At A Time

We start by visualizing one indicator at a time.

8.1 Continuous Variable

Sometimes the most interesting visualizations, are visualizations that give us a sense of the maximum, minimum, and average values. For example, the *histogram* and *dotplot* display information on *age*.

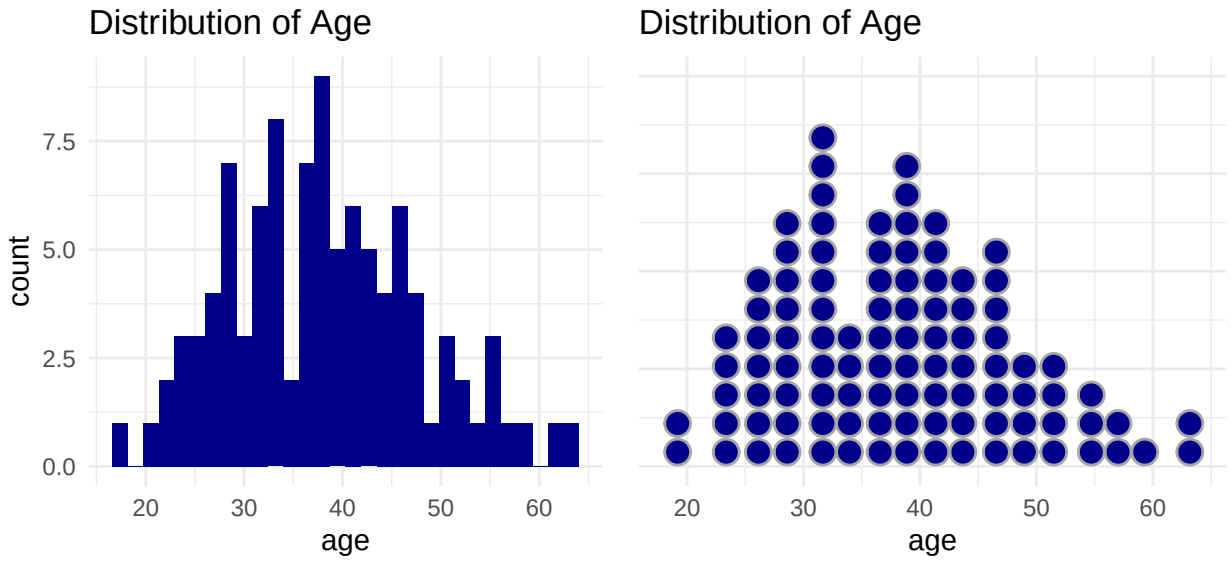


Figure 3: Histogram and Dotplot

8.2 Categorical Variable

We would use a slightly different visualization, for example, a barchart, when our data are grouped into categories.

9 Show The Relationship Of Two Things

Our task becomes somewhat more complicated when we want to understand the relationship of *one thing* to *another thing.*

9.1 Categorical by Categorical

Here, for example, we visualize two *categorical* variables, *neighborhood*, by *group*. In this graph, the height of the bars represents the *count* of observations.

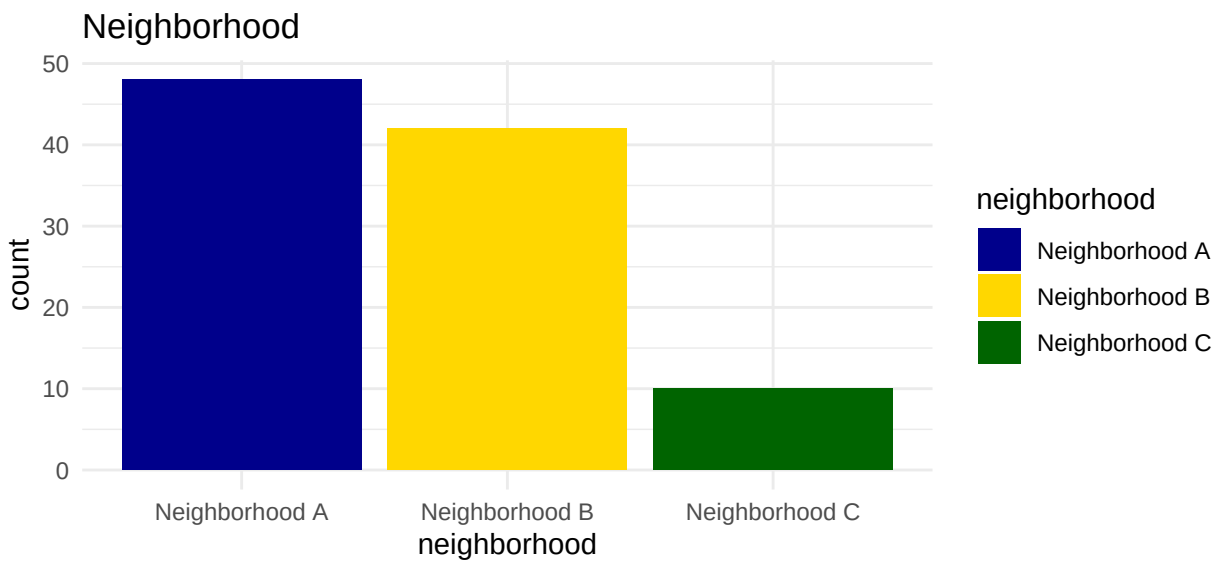


Figure 4: Barchart

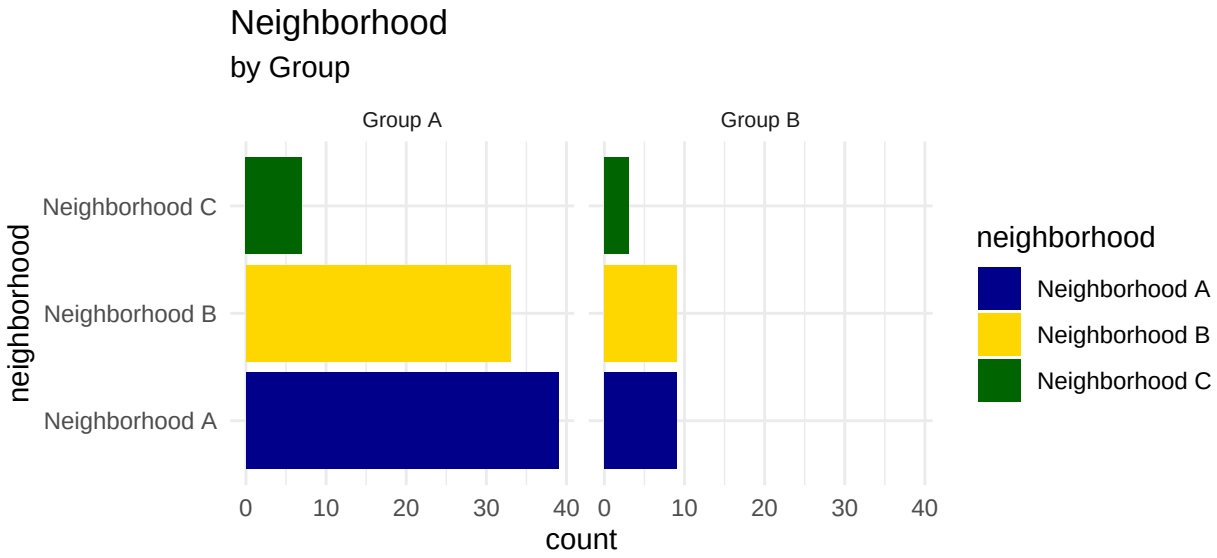


Figure 5: Barchart

9.2 Continuous by Continuous

Here, we visualize two *continuous* variables, *mental health*, by *age*.

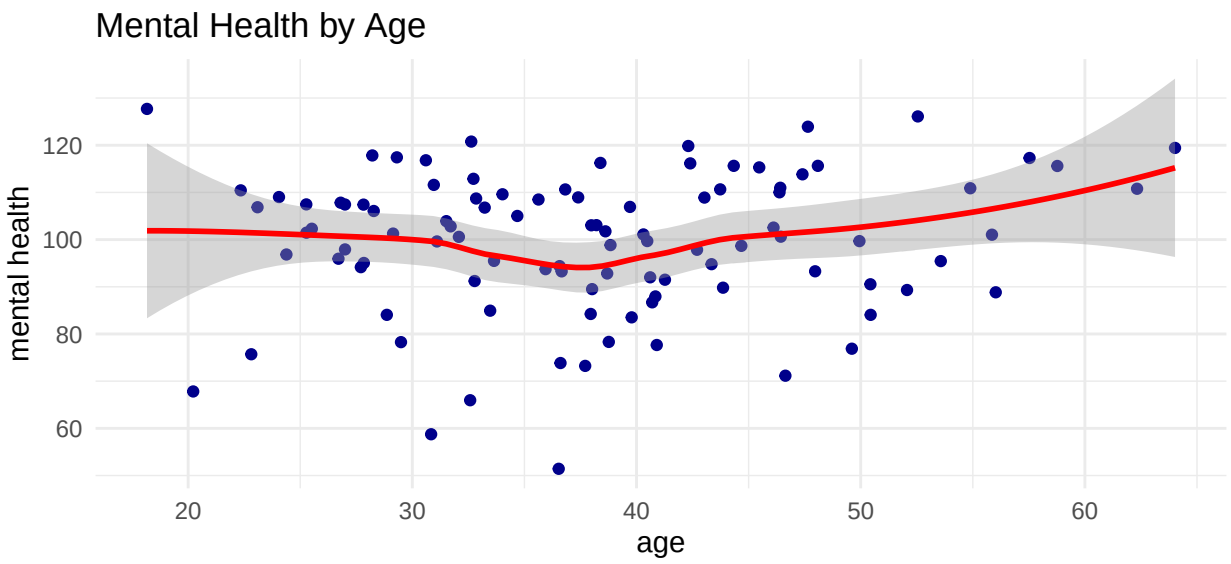


Figure 6: Scatterplot

9.3 Continuous by Categorical

Last, we visualize a *continuous* variable by a categorical variable, *mental health*, by *group*. In this graph, the height of the bars represents the *mean score*.

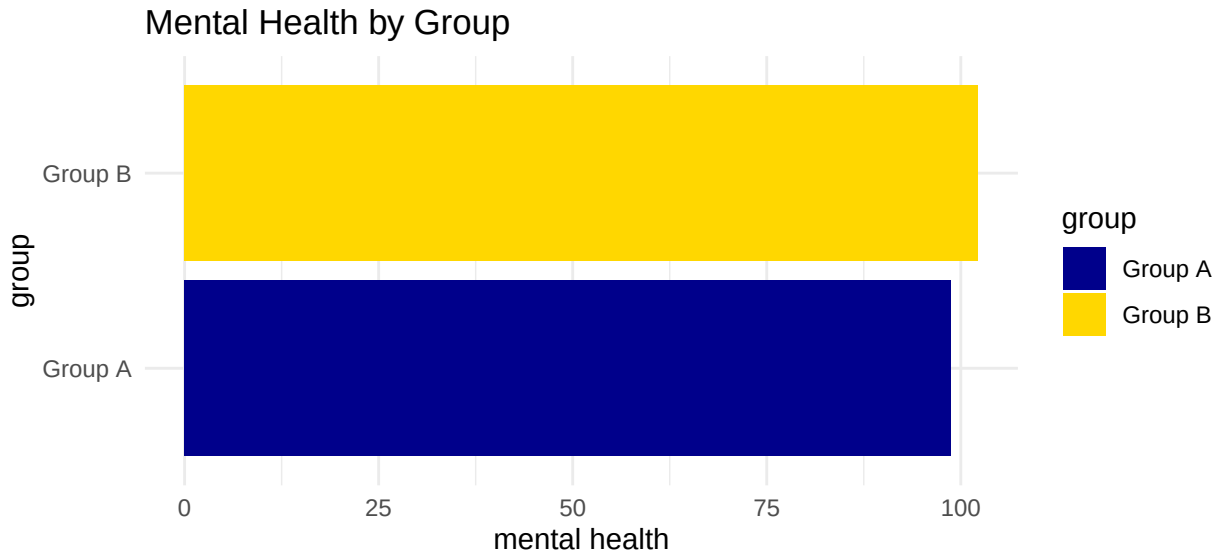


Figure 7: BarChart

10 Show Where Something Is

Sometimes our task is different. We want to visualize information, but add information on spatial location, using a map.

10.1 Map

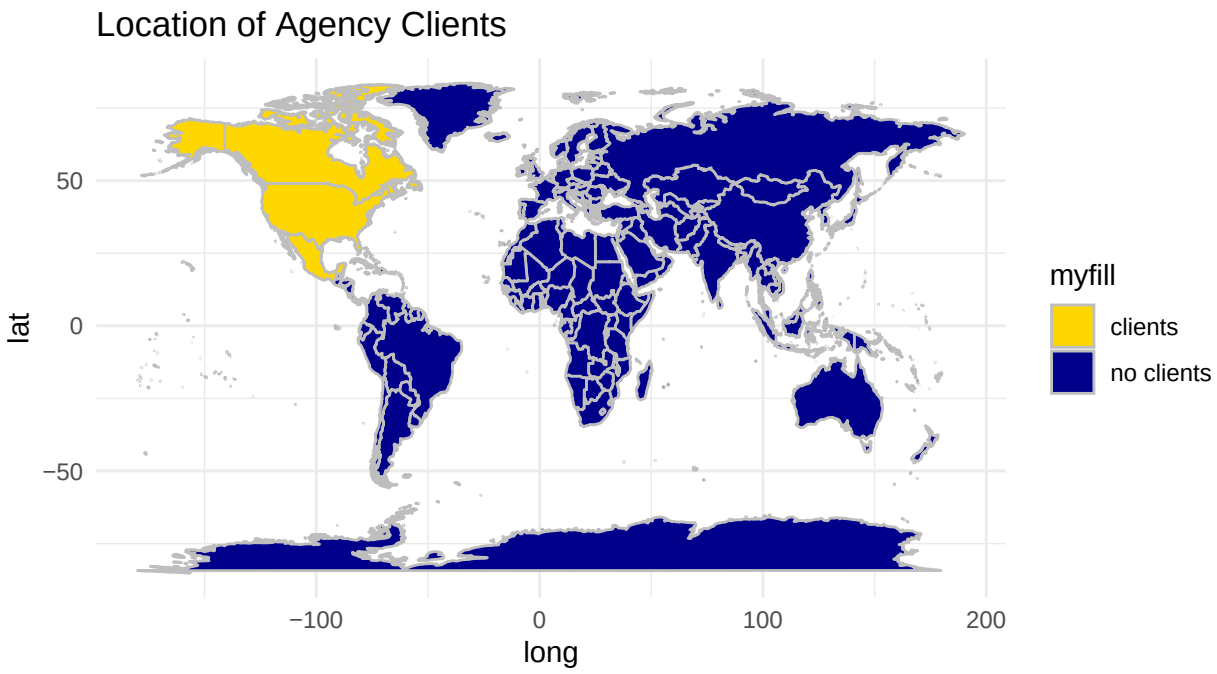


Figure 8: Map

11 Credits

Graphics made with the ggplot2 graphing library created by Hadley Wickham.

12 More Information

- A more exhaustive set of graphical possibilities, along with the code to generate them in R and ggplot2, can be found at my tutorial on How To Choose A Chart.
- I have also written an Introduction to R