6
Frames, Glyphs, and other Components of Graphics

Data graphics are built from parts. Chapter 5 showed the parts assembled together. This chapter looks at the parts individually.

Of course, a data table provides the basis for drawing a data graphic. The relationship between a data table and a graphic is simple: Each case in the data table becomes a mark in the graph. The designer of the graphic — you — chooses which variables the graphic will display and how each variable is to be represented graphically: position, size, color, and so on. The marks themselves are called *glyphs*. A data graphic has one glyph for each case in the data table.

Key Graphics Vocabulary

FRAME: The relationship between position and the data being plotted.

GLYPH: The basic graphical "unit" that represents one case. Other terms used include "mark" and "symbol." Variables set graphical attributes of the shape: size, color, shape, and so on. The location of the glyph — location is an important graphical attribute! — is set by the two variables defining the frame.

AESTHETIC: Any graphical attribute of a glyph: size, location, shape, color, etc.

SCALE: The relationship between the value of a variable and the graphical attribute to be displayed for that value.

GUIDE: An indication for the human viewer of the scale, that is, graphics how a variable encodes into its graphical attribute. Common guides are x- and y-axis tick marks and color keys.

6.1 *The Frame*

The frame of a graphic provides the space for drawing glyphs. But there is more to a frame than a blank canvas or piece of paper. The frame defines what position means. Most often, the frame is a rectangular region and position is described in terms of the familiar (x, y) Cartesian coordinate system. In creating a frame, you must decide which variable in your data will correspond to the x coordinate, and which to the y coordinate.

For instance, consider a dataset relevant to economic productivity. Table 6.1 gives per capita GDP for each country as well as some of the explanatory candidates: average educational level in the population, length of roadways per unit area, Internet use as a fraction of the population.

country	gdp	educ	roadways	net_users
Australia	44353.87	5.60	0.11	>60%
Ghana	3509.96	8.10	0.46	>0%
Mozambique	1140.04	5.00	0.04	>0%
	... and so on for 256 rows			

Table 6.1: Data relevant to economic performance. The complete table is available at `http://tiny.cc/dcf/table-6-2.csv`

You define a frame by selecting two variables from the glyph-

ready data table. For instance, Figure 6.1 shows a frame based on GDP and length of roadways. The frame provides the meaning to location in space.

6.2 *Glyphs*

The frame itself doesn't display any of the cases. Instead, the glyphs positioned in the frame represent the cases. There will be one glyph for each case in the data table.

The basic shape used in scatter plots is a simple glyph: a dot, a square, a triangle, an x, and so on. The following graph uses small dots. Since each case is a country, each dot represents one country.

In Figure 6.2 the glyphs are simple. Only position in the frame distinguishes one glyphs from another. The shape, size, etc. of all of the glyphs are identical. There's nothing about the glyph itself which identifies the country. It's possible to use a glyph with several attributes. Figure 6.3 location and label, mapping country name to the label.

But glyphs can have several properties. The aspects of each glyph that we can perceive are called *aesthetics*, or equivalently *graphical attributes*. The word *aesthetics* applied in the context of glyphs is not used in the modern sense. Nowadays, most people associate aesthetics with notions of beauty and artistic taste. The earlier meaning of the word, *properties relating to perception by the senses*, is the one intended when it comes to glyphs.

Location in the frame are the (x, y) aesthetics for a glyph, but other aesthetics can display variables in the data table. For instance, color could be used to show Internet use (as a fraction of the population), as in Figure 6.4. Another aesthetic is size. The size is fixed in 6.4; the same for every country. Figure 6.5 maps the average years of eduction onto the size aesthetic.

6.3 *Scales and Guides*

There are four aesthetics in Figure 6.5. Each of the four aethetics is set in correspondence with a variable; we say the variable is *mapped* to the aesthetic. Length of roadways is being mapped to horizontal position, GDP to vertical position, Internet connectivity to color, and educational attainment to size.

A *scale* is the relationship between a variable and the aesthetic to which it is mapped. For `roadways`, the scale says what value of the variable will correspond to position at the bottom of the frame, what value will correspond to the top of the frame, and where things fall inbetween.

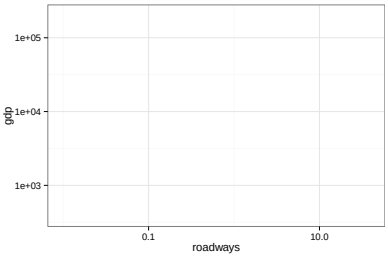


Figure 6.1: A graphics frame set by the `GDP` and `roadway` variables. No glyphs have been set in this frame.

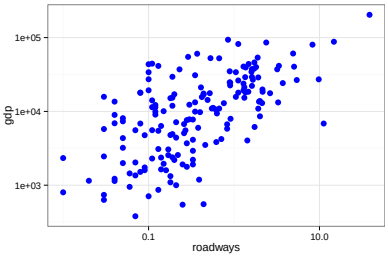


Figure 6.2: Using only position as the aesthetic for glyphs.

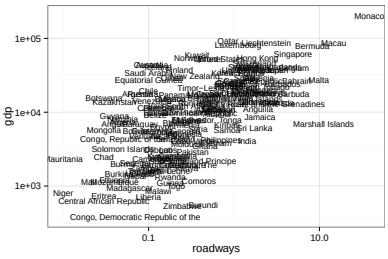


Figure 6.3: Using both location and label as aethetics.

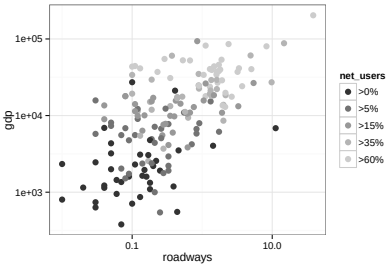


Figure 6.4: `net_users` mapped to color.

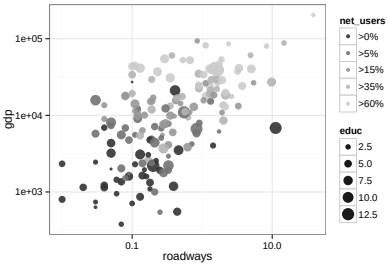


Figure 6.5: `net_users` mapped to color, `educ` mapped to size. Compare this graphic to Figure effig:facet-internet, which shows the same data using facets.

Not all scales are about position. For instance, in Figure 6.5, `net_users` is translated to color. Similarly, average educational attainment (in years) is translated to size: the middle-sized dot corresponds 7½ years of education.

Scales translate values into aesthetic properties. *Guide*s help to human reader to do the back translation. For position aesthetics, the most common sort of guide is the familiar axis with its tick marks and labels. But notice also the guide that tells how dot color corresponds to Internet connectivity. There's still another guide telling how dot size corresponds to education.

GUIDE: a display of a scale

6.4 *Facets*

Using multiple aesthetics such as shape, color, and size to display multiple variables can produce a confusing, hard-to-read graph. *Facets* provides a simple and effective alternative. Figure 6.6 uses facets to show different levels of Internet connectivity, providing a better view than Figure 6.5.

FACETS: Multiple side-by-side graphs used to display levels of a categorical variable.

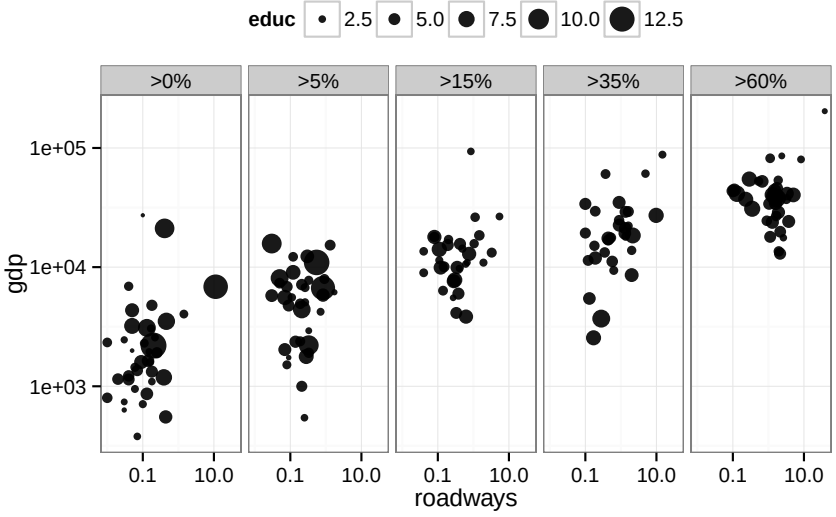


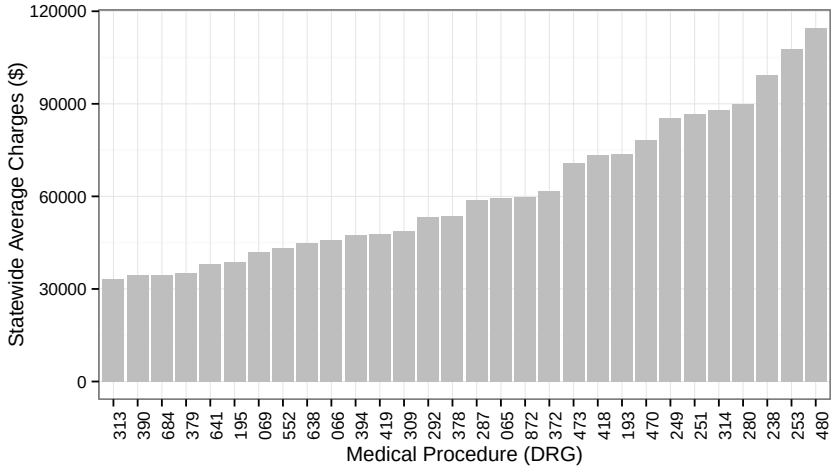
Figure 6.6: Using facets for different ranges of Internet connectivity.

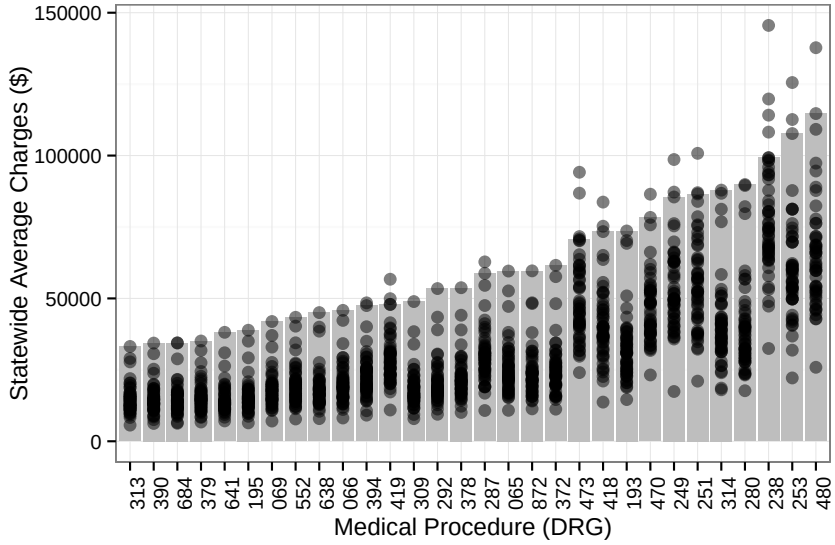
6.5 *Layers*

On occasion, data from more than one data table are graphed together. For instance, suppose you want a display of one state's hospital providers' charges for different medical procedures. The glyph-ready data table for New Jersey looks like Table 6.2. The glyph-ready table can be translated to a chart (Figure 6.7 (top)) using bars to give

a fair impression of the range in charges for different medical procedures in New Jersey.

How do the New Jersey charges compare to those in other states? Tables 6.2 and 6.3 provide relevant data. The two data tables, one for New Jersey and one for the whole country, can be plotted with different types of glyph: bars for New Jersey and dots for the whole country as in Figure 6.8.





With the context provided by the individual states, it's easy to see the charges in New Jersey are among the highest in the country for each medical procedure. (A description of each medical procedure number is given in the data table `DirectRecoveryGroups` in the `DataComputing` package.)

drg	stateProvider	mean_charge
313	NJ	33183.08
390	NJ	34361.27
684	NJ	34396.18
379	NJ	35091.47
641	NJ	38134.32
	... and so on for 100 rows	

Table 6.2: Glyph-ready data for the barplot layer in Figure 6.7

Figure 6.7: Average charges for medical procedures in New Jersey.

drg	stateProvider	mean_charge
039	ME	14480.66
039	KS	25859.60
039	FL	42636.68
039	DC	47369.88
039	MO	26128.30
039	TN	26417.38
057	GA	18508.69
057	SC	21741.92
057	AK	31858.55
057	UT	20247.73
	... and so on for 5,025 rows	

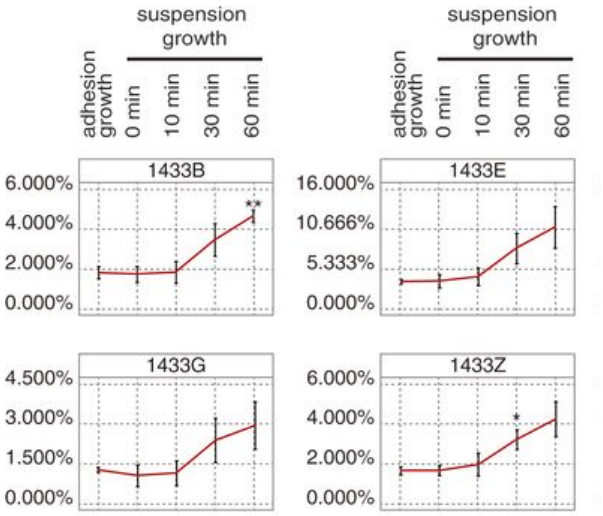
Table 6.3: Glyph-ready data table for the scatter-plot layer in Figure 6.7

Figure 6.8: Adding a second layer to provide a comparison of New Jersey to other states. Average charges for medical procedures in New Jersey.

6.6 *Exercises*

Problem 6.1

The following chart contains four facets. Each shows the amount of a substance in different conditions:

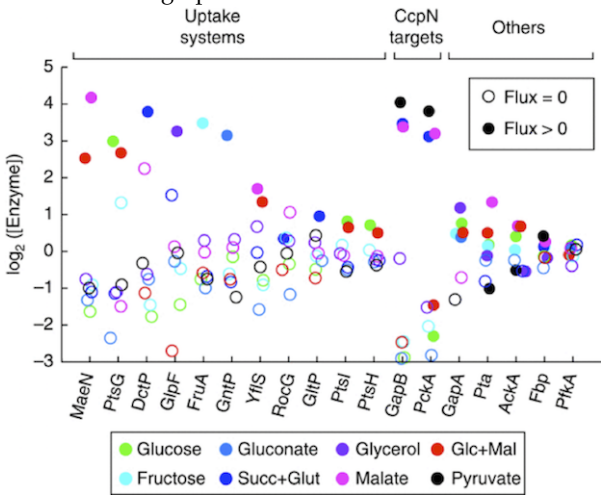


- when the cells are adhering to a surface
- when the cells are growing in suspension for different amounts of time

1. What are the labels/identifiers for the facets?
2. Are the frames the same in each facet?
3. There are three different glyphs shown in the frames. Describe each type in terms of its graphical properties.

Problem 6.2

Consider this graph



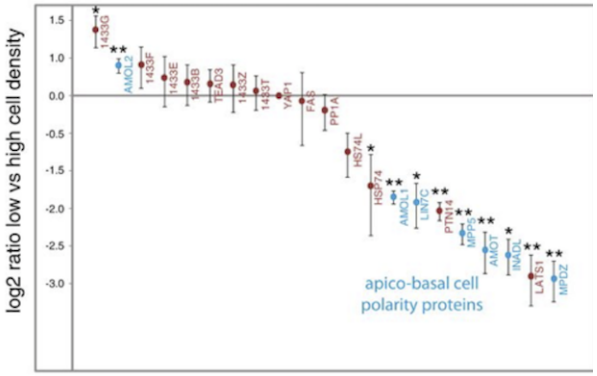
Here are some of the variables and their levels:

- Log enyzme `concentration`: numerical −3 to 5
- `target`: CcpN, Uptake, Other
- `flux`: zero or positive
- `gene`: MaeN, PtsG, DctP, ...
- `molecule`: Glocose, Fructose, Gluconate, ...

1. List all of the **guides** in the graph. For each one, say which variable is being mapped to which graphical attribute.
2. The basic glyph is a dot. Say what are the graphical attributes of the dot (e.g. color, size, ...). For each graphical attribute found in the graph, say which variable is mapped to that attribute.
3. Which two variables set the frame?
4. The scaling of the horizontal variable (e.g. the translation of position to variable levels) is set by a combination of two variables. Which two?

Problem 6.3

Consider this graphic:



Suppose the glyph-ready data underlying the graphic were structured as follows:

protein	center	low	high	polarity	signif
1433G	1.35	1.18	1.54	plus	1
AMOL2	0.78	0.63	1.01	minus	2
1433F	0.79	0.18	1.19	plus	0
1433E	0.42	-0.15	1.01	plus	0
⋮	⋮	⋮	⋮	⋮	⋮

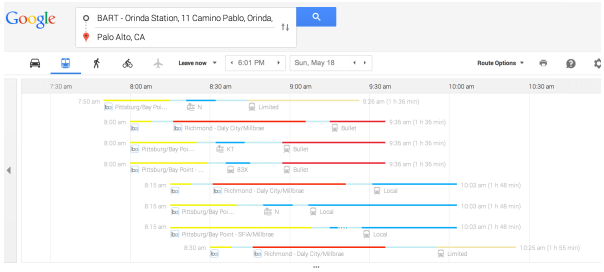
Consider these two kinds of glyph present in the graph:

and **

1. For each of the two glyphs, list the set of graphical attributes both geometrically (e.g. "dot") and in terms of the variable from the table that is mapped to that attribute (e.g., `polarity`).
2. Which variables define the frame? Give variables for both the horizontal and vertical coordinates.
3. Is color an attribute of the ** glyph?
4. What guides (if any) are displayed?

Problem 6.4

The graph, from Google Maps, shows mass transit options on a Monday morning for getting from Orinda, CA (in the East Bay), to Palo Alto, CA (in the West Bay).



(For a larger version, see `Data-Computing.org/images#C133`.)

1. Considering only that part of the graphic below the blue underlined bus and other modes of transportation, what is the frame?
2. Describe the different types of glyphs used.
3. For each different type of glyph
   - What information is encoded in the shape/style of the glyphs?
   - What information is encoded in the position of the glyph?
4. What guides are there?

Figure 6.9 presents forecasts for the US Senate elections in Nov. 2014. The numbers or words give the forecast probability of one party's candidate — Democrat or Republican — winning. The forecasts are made based on polls up through the end of August 2014. Individual results from several different polling organization are shown. The graphic is an excerpt from the full graphic at `http://www.nytimes.com/newsgraphics/2014/senate-model/comparisons.html`, which shows predictions for all 36 senate seats up for election in 2014. Source: New York Times

Competitive States	NYT Aug 31	538 Aug 4	Cook Aug 22	Roth. Aug 29	Sabato Aug 27	WaPo Aug 29
New Hampshire	84% Dem.	90% Dem.	Leaning	Likely	Likely	>99% Dem.
Michigan	74% Dem.	65% Dem.	Tossup	Leaning	Likely	99% Dem.
Colorado	57% Dem.	60% Dem.	Tossup	Tossup	Leaning	65% Dem.
Iowa	53% Dem.	55% Dem.	Tossup	Tossup	Tossup	63% Rep.
Alaska	52% Dem.	Even	Tossup	Tossup	Tossup	66% Dem.
North Carolina	51% Rep.	Even	Tossup	Tossup	Tossup	91% Dem.
Louisiana	60% Rep.	55% Rep.	Tossup	Tossup	Tossup	51% Dem.
Arkansas	66% Rep.	60% Rep.	Tossup	Tossup	Tossup	65% Rep.
Georgia	82% Rep.	75% Rep.	Tossup	Likely	Leaning	83% Rep.
Kentucky	86% Rep.	80% Rep.	Tossup	Leaning	Likely	94% Rep.

* Rothenberg ratings are converted from a nine-category scale to a seven-category scale to make comparisons easier.

Solid Dem. · Likely Dem. · Leaning Dem. · Tossup · Leaning Rep. · Likely Rep. · Solid Rep.

Figure 6.9: Forecasts made before the Nov. 2014 US Senate elections.

Problem 6.5

In Figure 6.9, what variables define the frame in this graphic?

a. Probability and State.

b. State and Polling Organization.

c. Democrats and Republicans.

d. Just State

e. Just Probability

Problem 6.6

In Figure 6.9, what is the glyph and its graphical attributes?

a. Glyph: names of the states. Graphical attribute: font.

b. Glyph: names of the polling organization. Graphical attribute: the organization's logo.

c. Glyph: Rectangle. Graphical attribute: color.

d. Glyph: Rectangle. Graphical attribute: color and text.

Problem 6.7

In Figure 6.9, what sets the order of the categorical variable in the scale for the vertical variable?

a) State

b) Poll

c) Roth poll probability for the Democratic candidate.

d) NYT poll probability for the Democratic candidate.

e) Date of the poll.

Problem 6.8

The `NCHS` data (in the `DataComputing` package) has 31126 rows. To speed things up, work with a small subset of `NCHS`:

```
Small <-
  NCHS %>%
  sample_n(size=5000)
```

Using the data in `Small`, make this plot with `scatterGraphHelper()` (in the `DataComputing` package). Then, write down the mapping between variables and graphical attributes.