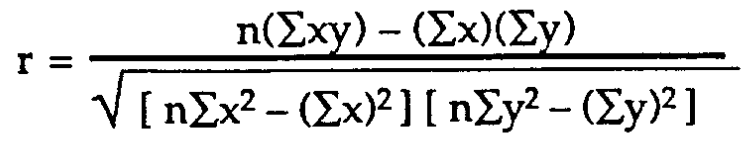
a **dummy variable** is one that takes the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome.[[3]](https://en.wikipedia.org/wiki/Dummy_variable_(statistics)#cite_note-3)[[4]](https://en.wikipedia.org/wiki/Dummy_variable_(statistics)#cite_note-Interpreting_Coefficients-4) Dummy variables are used as devices to sort data into [mutually exclusive](https://en.wikipedia.org/wiki/Mutually_exclusive_events) categories (such as smoker/non-smoker, etc.).[[2]](https://en.wikipedia.org/wiki/Dummy_variable_(statistics)#cite_note-Gujarati-2) For example, in [econometric](https://en.wikipedia.org/wiki/Econometrics) [time series analysis](https://en.wikipedia.org/wiki/Time_series_analysis), dummy variables may be used to indicate the occurrence of wars or major [strikes](https://en.wikipedia.org/wiki/Strike_action). In the [panel data](https://en.wikipedia.org/wiki/Panel_data), [fixed effects estimator](https://en.wikipedia.org/wiki/Fixed_effects_estimator) dummies are created for each of the units in [cross-sectional data](https://en.wikipedia.org/wiki/Cross-sectional_data) (e.g. firms or countries) or periods in a pooled time-series

The [Pearson product-moment correlation coefficient](https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient), also known as *r*, *R*, or Pearson's *r*, is a measure of the strength and direction of the linear relationship between two variables that is defined as the [covariance](https://en.wikipedia.org/wiki/Covariance) of the variables divided by the product of their standard deviations.



**geocoding:** Services that take addresses and zip code and turn them into latitude and longitude.

**ggplot2** *Grammar of Graphics*—a general scheme for data visualization which breaks up graphs into semantic components such as scales and layers. ggplot2 can serve as a replacement for the base graphics in R and contains several defaults for web and print display of common scales.

**Stacked chart**s are commonly used to compare proportions within a whole. communicate totals, trends and proportions in a single illustration. Hard to interpret. clustered column chart makes it easy to compare the values of any of the regions for any of the quarters. The trends in the data are much more evident than in the stacked chart. You use stacked bar charts when there are subcategories, and the sum of these subcategories is meaningful.

**Points** Sometimes it makes more sense to use points instead of bars. They use less space and because there are no bins, points can provide a better feeling of flow from one point to the next – scatter plot - It’s often used to show the relationship between two variables

The **treemap** is an area-based visualization where the size of each rectangle represents a metric. Outer rectangles represent parent categories, and rectangles within the parent are like subcategories. You can use a treemap to visualize straight-up proportions, but to fully put the technique to use, it’s best served with hierarchical, or rather, tree-structured data. Proportions/change over Time **Stacked Continuous** Imagine you have several time series charts. Now stack each line on top of the other. Fill the empty space. What you have is a stacked area chart, where the horizontal axis is time, and the vertical axis is a range from 0 to 100 percent. **Conclusion:** Only have a few values? The pie chart might be your best bet. Use donut charts with care. If you have several values and several categories, consider the stacked bar chart instead of multiple pie charts. If you’re looking for patterns over time, look to your friend the stacked area chart or go for the classic time series. With these steady foundations, your proportions will be good to go.

**Visualizing Relationships** Statistics is about finding relationships in data. What are the similarities between groups? Within groups? Within subgroups? The relationship that most people are familiar with for statistics is correlation. For example, as average height goes up in a population, most likely average weight will go up, too. This is a simple positive correlation.

**Correlation**

Correlation is probably the first thing you think of when you hear about relationships in data. The second thing is probably causation. Now maybe you’re thinking about the mantra that correlation doesn’t equal causation. The first, correlation, means one thing tends to change a certain way as another thing changes. For example, the price of milk per gallon and the price of gasoline per gallon are positively correlated. Both have been increasing over the years.

Now here’s the difference between correlation and causation. If you increase the price of gas, will the price of milk go up by default? More important, if the price of milk did go up, was it because of the increase in the gas price or was it an outside factor, such as a dairy strike?

It’s difficult to account for every outside, or confounding factor, which makes it difficult to prove causation. Researchers spend years figuring stuff like that out. You can, however, easily find and see correlation, which can still be useful, as you see in the following sections.

Correlation can help you predict one metric by knowing another. To see this relationship, return to scatterplot and multiple scatterplots.

A **Density Plot** visualises the distribution of data over a continuous interval or time period. This chart is a variation of a [Histogram](http://datavizcatalogue.com/methods/histogram.html) that uses [kernel smoothing](https://en.wikipedia.org/wiki/Kernel_smoother) to plot values, allowing for smoother distributions by smoothing out the noise. The peaks of a Density Plot help display where values are concentrated over the interval.

An advantage Density Plots have over Histograms is that they're better at determining the [distribution shape](https://en.wikipedia.org/wiki/Shape_of_the_distribution)because they're not affected by the number of bins used (each bar used in a typical histogram). A Histogram comprising of only 4 bins wouldn't produce a distinguishable enough shape of distribution as a 20-bin Histogram would. However, with Density Plots, this isn't an issue.

Historams are constructed by binning the data and counting the number of observations in each bin.The objective is usually to visualize the shape of the distribution.

The number of bins needs to be

large enough to reveal interesting features;

small enough not to be too noisy.

A very small bin width can be used to look for rounding or heaping.

Common choices for the vertical scale are-bin counts, or frequencies-counts per unit, or densities

The count scale is more intepretable for lay viewers.

The density scale is more suited for comparison to mathematical density models.

Constructing histograms with unequal bin widths is possible but rarely a good idea.

 LOESS fit is a locally moving weighted regression based on the original data points. A simple linear model fits a straight line through a set of points. The line is the best possible straight line (at least, for one sensible definition of best) A loess model fits a complicated curve through a set of points. In some ways, it can be thought of as a complicated moving average.

**Prepare yourself**: When you explore your own data, you don’t need to do much in terms of storytelling. You are, after all, the storyteller. However, the moment you use your graphic to present information—whether it’s to one person, several thousand, or millions—a standalone chart is no longer good enough. Sure, you want others to interpret results and perhaps form their own stories, but it’s hard for readers to know what questions to ask when they don’t know anything about the data in front of them. It’s your job and responsibility to set the stage. How you design your graphics affects how readers interpret the underlying data. You need to know your source material to tell good stories with data. Know what each metric represents. Learn all you can about the data, and the visual storytelling will come natural.

**Prepare your reader:** The lesson: Don’t assume your readers know everything or that they can spot features in your graphic. Your job as a data designer is to communicate what you know to your audience.

**good design** not only lends to aesthetics, but also makes graphics easier to read and can change how readers feel about the data or the story you tell.

**color** choice can play a major role in data graphics. It can evoke emotions (or not) and help provide context. It’s your responsibility to choose colors that represent an accurate message. Your colors should match the story you are trying to tell. a simple color change can change the meaning of your data completely**.**

**Visual Cues:** In Chapter 1, “Telling Stories with Data,” you saw how encodings work. Basically, you have data, and that data is encoded by geometry, color, or animation. Readers then decode those shapes, shades, and movement, mapping them back to numbers. This is the foundation of visualization. Encoding is a visual translation. Decoding helps you see data from a different angle and find patterns that you otherwise would not have seen if you looked only at the data in a table or a spreadsheet.Through all the examples in previous chapters, you’ve seen how good design not only lends to aesthetics, but also makes graphics easier to read and can change how readers actually feel about the data or the story you tell. Graphics with default settings from R or Excel feel raw and mechanical. This isn’t necessarily a bad thing. Maybe that’s all you want to show for an academic report. If, however, you do want to display your graphic prominently, a quick color change can make all the difference. A darker color scheme might be used for a somber topic, whereas a brighter color scheme can feel more happy-go-lucky. **Visualization**: displays the data of interest fairly and presents the data in an engaging way.

**Good Visualization**: Statisticians and analysts, for example, generally think of visualization as traditional statistical graphics that they can use in their analyses. If a graphic or interactive doesn’t help in analysis, then it’s not useful. It’s a failure. On the other hand, if you talk to graphic designers about the same graphic, they might think the work is a success because it displays the data of interest fairly and presents the data in an engaging way. The analytically minded can learn a lot from designers about making data more relatable and understandable, whereas design types can learn to dig deeper into data from their analytic counterparts. Ultimately, it’s all about your goals for the graphic, what story you want to tell, and who you tell it to. Take all the above into account—and you’re golden.

**Conclusion:** A lot of data people see design as just a way to make your graphics look pretty. That’s certainly part of it, but design is also about making your graphics readable, understandable, and usable. You can help people understand your data better than if they were to look at a default graph. You can clear clutter, highlight important points in your data, or even evoke an emotional response.

When you have a big dataset, and you don’t know where to begin, the best place to start is with a question. What do you want to know? Are you looking for seasonal patterns? Relationships between multiple variables? Outliers? Spatial relationships? Then look back to your data to see if you can answer your question. If you don’t have the data you need, then look for more.

**Cultural sensitivity** is being aware that cultural differences and similarities between people exist without assigning them a value – positive or negative, better or worse, right or wrong.

Comprehend information quickly

By using graphical representations of business information, businesses are able to see large amounts of data in clear, cohesive ways – and draw conclusions from that information. And since it’s significantly faster to analyze information in graphical format (as opposed to analyzing information in spreadsheets), businesses can address problems or answer questions in a more timely manner.

**how visualization help make sense of data**

**Identify relationships and patterns**

Even extensive amounts of complicated data start to make sense when presented graphically; businesses can recognize parameters that are highly correlated. Some of the correlations will be obvious, but others won’t. Identifying those relationships helps organizations focus on areas most likely to influence their most important goals.

**Pinpoint emerging trends**

Using data visualization to discover trends – both in the business and in the market – can give businesses an edge over the competition, and ultimately affect the bottom line. It’s easy to spot outliers that affect product quality or customer churn, and address issues before they become bigger problems.

**Communicate the story to others**

Once a business has uncovered new insights from visual analytics, the next step is to communicate those insights to others. Using charts, graphs or other visually impactful representations of data is important in this step because it’s engaging and gets the message across quickly.

**Comprehend information quickly**

By using graphical representations of business information, businesses are able to see large amounts of data in clear, cohesive ways – and draw conclusions from that information. And since it’s significantly faster to analyze information in graphical format (as opposed to analyzing information in spreadsheets), businesses can address problems or answer questions in a more timely manner.

**What to Look for over Time**

The most common thing you look for in time series, or temporal, data is trends. Is something increasing or decreasing? Are there seasonal cycles? Are there outliers? Are there any periods of time that look out of place? Are there spikes or dips? If so, what happened during that time? Often, these irregularities are where you want to focus. Other times the outliers can end up being a mi stake in data entry.