In 1977, statistician John Tukey published his book Exploratory Data Analysis, which detailed how and encouraged data professionals to analyze data through visualization. This was during a time when most analysis was performed in the context of hypothesis tests and statistical models, one computer filled a room, and graphs were typically drawn by hand. For example, in his book, Tukey provides a tip on how to draw darker symbols with a pen instead of a pencil.

Nevertheless, although the technology was bigger and slower back then, the driving principle is the same. You can see a lot in a picture, and what you see can lead to answers or generate more questions you otherwise never would have thought of.

> *"The greatest value of a picture is when it forces us to notice what we never expected to see."*
>
> *—John W. Tukey, Exploratory Data Analysis (1977)*

The public-facing side of visualization—the polished graphics that you see in the news, on websites, and in books—are fine examples of data graphics at their best, but what is the process to get to that final picture? There is an exploration phase that most people never see, but it can lead to visualization that is a level above the work of those who do not look closely at their data. The better that you understand what your data is about, the better you can communicate your findings.

Note: *The New York Times* and *The Washington Post* discuss the process behind their graphics at http://chartsnthings.tumblr.com/ and http://postgraphics.tumblr.com/, respectively. Work often starts with rough sketches on paper or dry erase board and then moves to exploration and production.

Even if you don't plan to show your results to a wide audience, visualization as an analysis tool enables you to explore data and find stories that you might not find with formal statistical tests. You just need to know what to look for and what questions to ask based on the data that you have available.

The great thing is that tools and access to data are less of a limiting factor than they were in Tukey's time, so you aren't stuck with just pencils, paper, and a ruler to draw thousands of dots and lines.

PROCESS

The specific steps you take in any analysis varies by dataset and project, but generally speaking, you should consider these four things when you explore your data visually.

- What data do you have?
- What do you want to know about your data?

- What visualization methods should you use?
- What do you see and does it makes sense?

The answer to each question depends on the answers that come before, and it's common to jump back and forth between questions. As shown in Figure 4-1, it's an iterative process. For example, if your dataset is only a handful of observations, this limits what you can find in your data and what visualization methods are useful, and you won't see much.

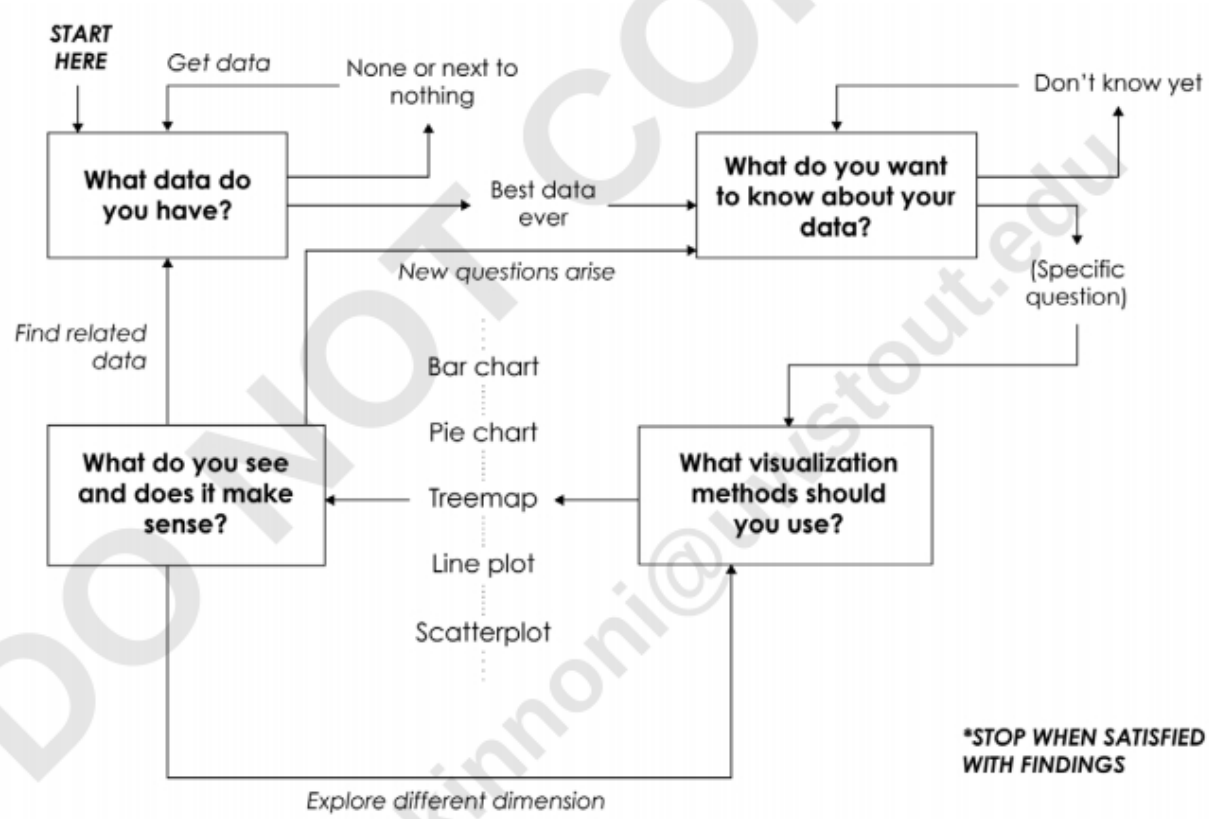


FIGURE 4-1 *The iterative data exploration process*

On the other hand, if you have a lot of data, what you see when you visualize one aspect of it can lead to a curiosity about other dimensions, which in turn leads to different graphics. This is the fun part.

WHAT DATA DO YOU HAVE?

People often form a picture in their head of what a visualization should look like or find an example that they want to mimic. The excitement is great, but

when it's time to visualize, they realize they either need more data or their data doesn't work with the chart they want to make.

The common mistake is to form a visual first and get the data later. It should be the other way around—data first and visualization follows.

Often, getting the data that you need is the hardest and most time-consuming part of the visualization process. In school, data is handed to you formatted the way you need so that you can easily load it into the software of choice, but this is hardly ever the case in practice. You might need to scrape data from a website, access an API, or derive values from existing data.

Note: I sometimes spend most of the time gathering data and little time visualizing it. Don't be surprised if you have to do the same. Totally normal.

For example, you might have a list of addresses, but to map them, you need latitude and longitude coordinates. Or you have observations for individuals of a population, but you might be more interested in subpopulations.

Programming can be helpful in this case to automate parts of the process, but there are a growing number of click-and-play applications to manage data, too.

Note: See Chapter 7, "Where to Go from Here," for tools and resources to work with and visualize data.

When you have data you want to explore, pause for a second to consider what values represent, where the data is from, and how variables were measured. Essentially, apply everything you learned in Chapter 1, "Understanding Data."

WHAT DO YOU WANT TO KNOW ABOUT YOUR DATA?

Imagine you have some data to explore. Where do you begin? The answer is easy if you have only one data point. You can just read the value, and most of your findings, if any, will come from outside information and additional data. On the other hand, when you have a dataset with thousands or millions of observations—think spreadsheet with a lot of rows—it can be challenging and often intimidating to figure out what to look at first.

This is where the phrase "drowning in data" comes from. You stare at a bunch of numbers on your computer screen, and values start to blur together the longer you stare. Soon all you see is a blob of data that feels suffocating, but wait; there's hope. Take a step back. Breathe.

To avoid drowning in data, you learn to swim. When you learn to swim, you start at the shallow end and work your way toward the deep end. If you're

more adventurous, you snorkel or go deep-sea diving. Even then you don't swim the entire ocean. You explore a little bit at a time, and what you learn during one dive carries over to the next. People drown in the ocean, but when you drown in data, you still get more chances to learn and try again.

To start, ask yourself what you want to know about the data. Your answer doesn't need to be complex or profound. Just make it less vague than, "I want to know what the data looks like." The more specific you are the more direction you get. Maybe you want to know the best or worst thing (such as a country, sports team, or school) in your data, so you explore rankings, and if you have multiple variables, you decide what makes something good and something bad. If you have time series data, you might want to know if something has improved or gotten worse over the past decade.

Note: Early exploration of a dataset can be overwhelming, because you don't know where to start. Ask questions about the data and let your curiosities guide you.

For example, journalist Tim De Chant explores world population densities, as shown in Figure 4-2, and was curious how large a city might be if everyone who lived in the world had the same amount of space. A straightforward method could be to directly map population density around the world, but De Chant put it into a more relatable perspective.

When you ask questions about your data, you give yourself a place to start, and if you're lucky, as you investigate, you'll come up with more questions, and then you dig into those. Coming up with and answering potential questions a user might have while you explore also provides focus and purpose, and helps farther along in the design process when you make graphics for a wider audience.

WHAT VISUALIZATION METHODS SHOULD YOU USE?

As you saw in the previous chapter, there are many chart options and combinations of visual cues to choose from. It's easy to obsess over picking the right chart for your data, but during the early stages of exploration, it's more important to see your data from different angles and to drill down to what matters for your project.

Make multiple charts, compare all your variables, and see if there are interesting bits that are worth a closer look. Look at your data as a whole and then zoom in on categories and individual data points.

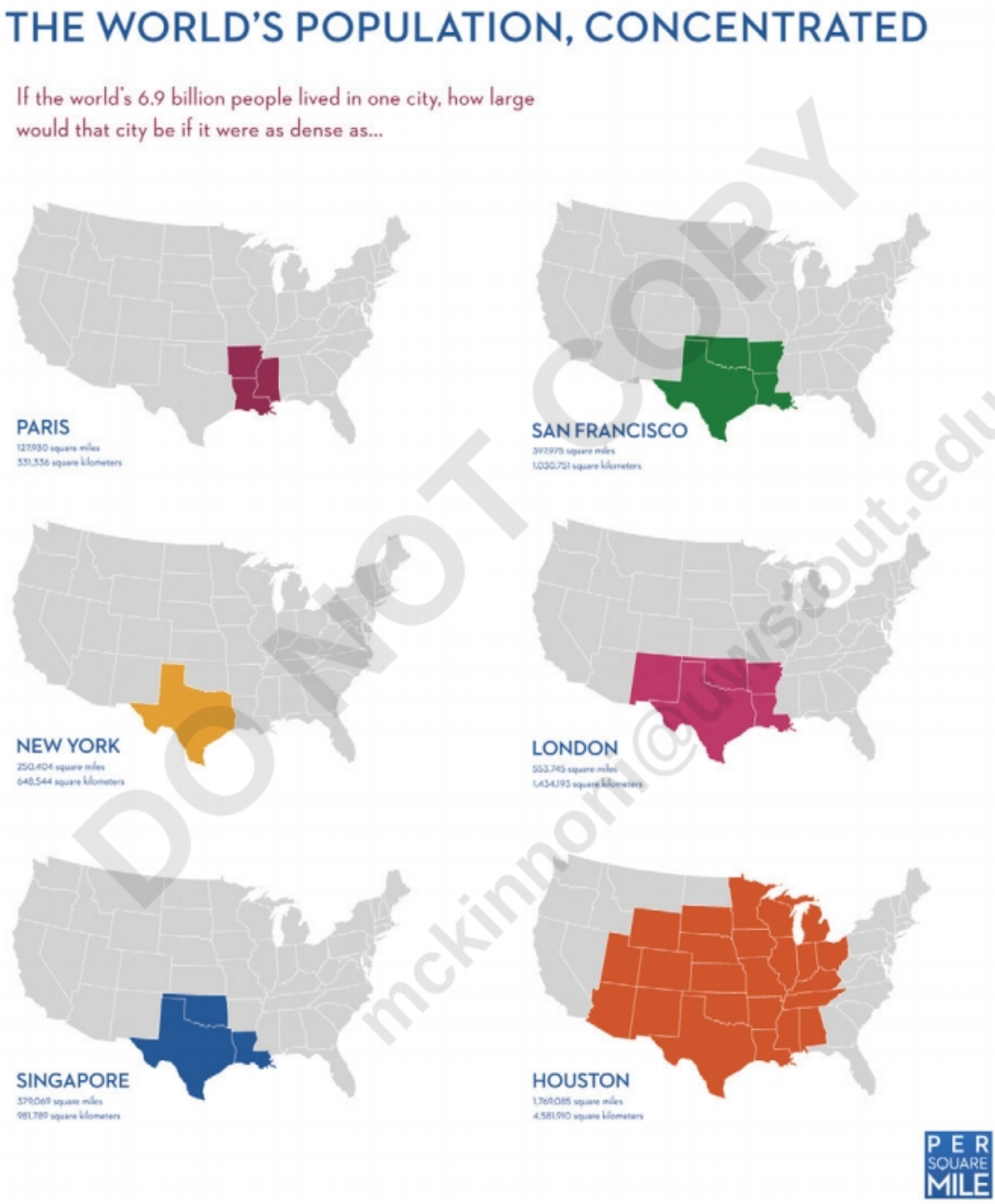


FIGURE 4-2 The World's Population, *Concentrated (2011) by Tim De Chant, http://persquaremile.com/*

This is also a great time to experiment with visual forms. Try different scales, colors, shapes, sizes, and geometries, and you might stumble upon a graphic worth pursuing further. You don't always need to stick to the visual cues that are the "best" at showing data most accurately and are easiest to read. When exploration is your goal, don't let a list of best practices stop you from trying something different because complex data often requires complex visualization.

Note: Ben Shneiderman, a professor of computer science and inventor of the treemap, is often quoted for, "Overview first, zoom and filter, then details-on-demand" in his paper "The Eyes Have It."

For example, Figure 4-3 shows an interactive exploration of article deletions on Wikipedia by Mortiz Stefaner, Dario Taraborelli, and Giovanni Luca Ciampaglia. Wikipedia is a large resource of data with small and large data tables within articles, article edits over time, and user interaction with articles and between each other. The data can be explored on many dimensions, but the topic focus of Notabilia shows a clearer picture.

Note: A common misconception is that you must understand a graphic in under 10 seconds. Relationships and patterns aren't always straightforward, so just because a visualization takes a few minutes to understand doesn't make it a failed attempt.

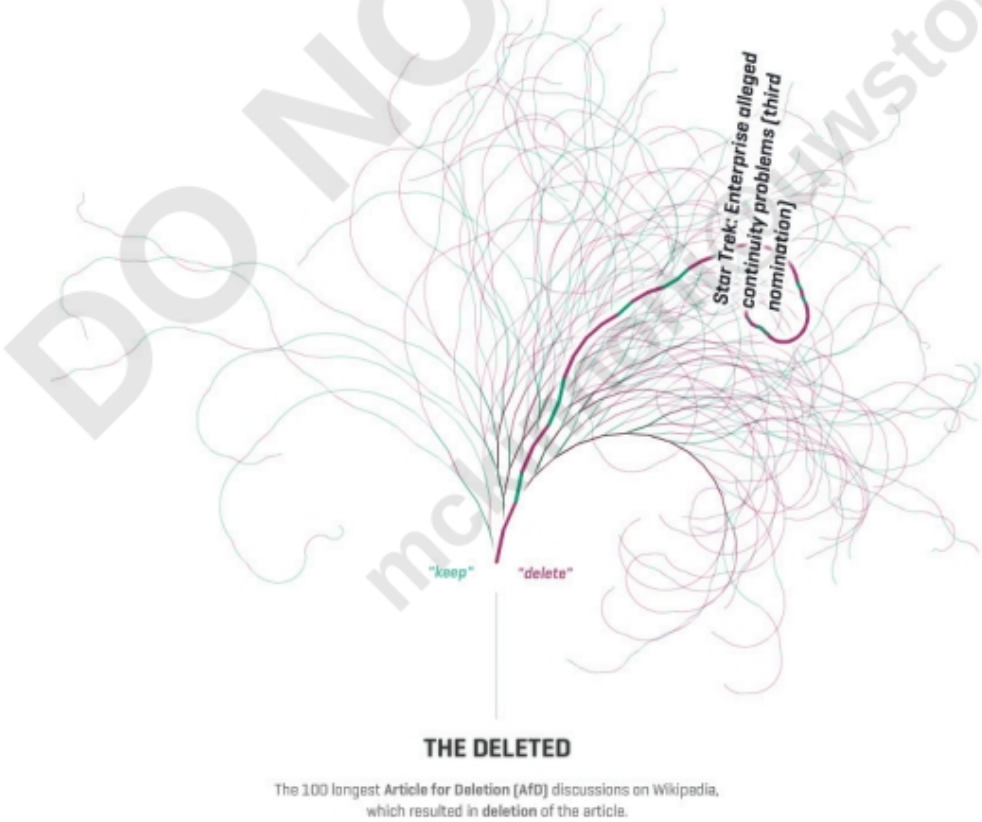


FIGURE 4-3 Notabilia *(2011) by Mortiz Stefaner, Dario Taraborelli, and Giovanni Luca Ciampaglia, http://notabilia.net/*

Each branch represents a user discussion about whether an article should be deleted, and those that curl to the right are discussions that lean strongly for deletion. A curl to the left is a discussion leaning toward keeping an article. The more prominent the curl is, the stronger the agreement between users. Although the visualization isn't traditional, you get still get something out of it.

That said, traditional visualization, such as bar graphs and line charts, can be made easily and read quickly, which makes them great tools to explore data.

As your goals shift, so do your choices of visualization. If you were to design a dashboard that provides the status of a system at a glance, you must visualize the data in a way that is straightforward to digest. On the other hand, if the goal is to encourage reflection or to evoke emotions, efficiency might not be your main concern.

WHAT DO YOU SEE AND DOES IT MAKE SENSE?

After you visualize your data, there are certain things to look for, as shown in Figure 4-4: increasing, decreasing, outliers, or some mix, and of course, be sure you're not mixing up noise for patterns.

Also note how much of a change there is and how prominent the patterns are. How does the difference compare to the randomness in the data? Observations can stand out because of human or mechanical error, because of the uncertainty of estimated values, or because there was a person or thing that stood out from the rest. You should know which it is.

When you find something interesting, ask yourself: Does it make sense? Why does it make sense? This is massively important.

The tendency is to automatically think of data as fact because numbers can't possibly waver. But again, there's uncertainty attached to the data because each data point is a snapshot of what happened during a moment in time. You infer everything else.

Note: Inference and uncertainty: This is what statistics is all about. If you can, take a statistics course. Although you can learn a lot from visual exploration alone, traditional statistics can help you examine data in greater detail.

In the rest of this chapter, you look at specific data types more closely. Keep the process in mind as you make your way through.

VISUALIZING CATEGORICAL DATA

You might like to group people, places, and things. The classifications and categorizations lend structure to what otherwise would be an amorphous blob of stuff. Figure 4-5 shows some of your options to visualize such categories.

The bar graph, of course, is one of the most common ways to show categorical data. Each rectangle represents a category, and the longer the rectangle is, the greater the value that it represents. Whether a higher value means better or worse can, of course, vary by dataset and point of view.

For example, in February 2012, the Pew Internet and American Life Project surveyed approximately 2,200 people about how they use the Internet, social networking sites such as Facebook and Twitter, and whether politics was a regular occurrence on those sites. Figure 4-6 shows the results for four of the fifty questions.

As you might expect, Google was the most common chosen search engine; Facebook was far ahead of Twitter and career-based social network LinkedIn. The responses to the other questions probably aren't that surprising to you either.

Note: The Pew Internet & American Life Project makes its survey data freely available at http://www.pewinternet.org/.

In this example, the bar graph is the visual equivalent of a list. Each bar represents a value, and you use separate bars and charts for separation. Length is your visual cue, with rectangles placed on a linear scale. You could however use a different scale and shapes to represent the same data. Figure 4-7 shows the same poll results with squares sized by area.

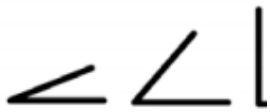
Notice that the differences among categories doesn't look as dramatic in the symbols plots as they do in the bar graphs. For example, the bar for Google looks a lot longer than the rest in the search engine bar graph, but when you compare the square for Google, it looks bigger, but not quite by the same magnitude relative to the other squares.

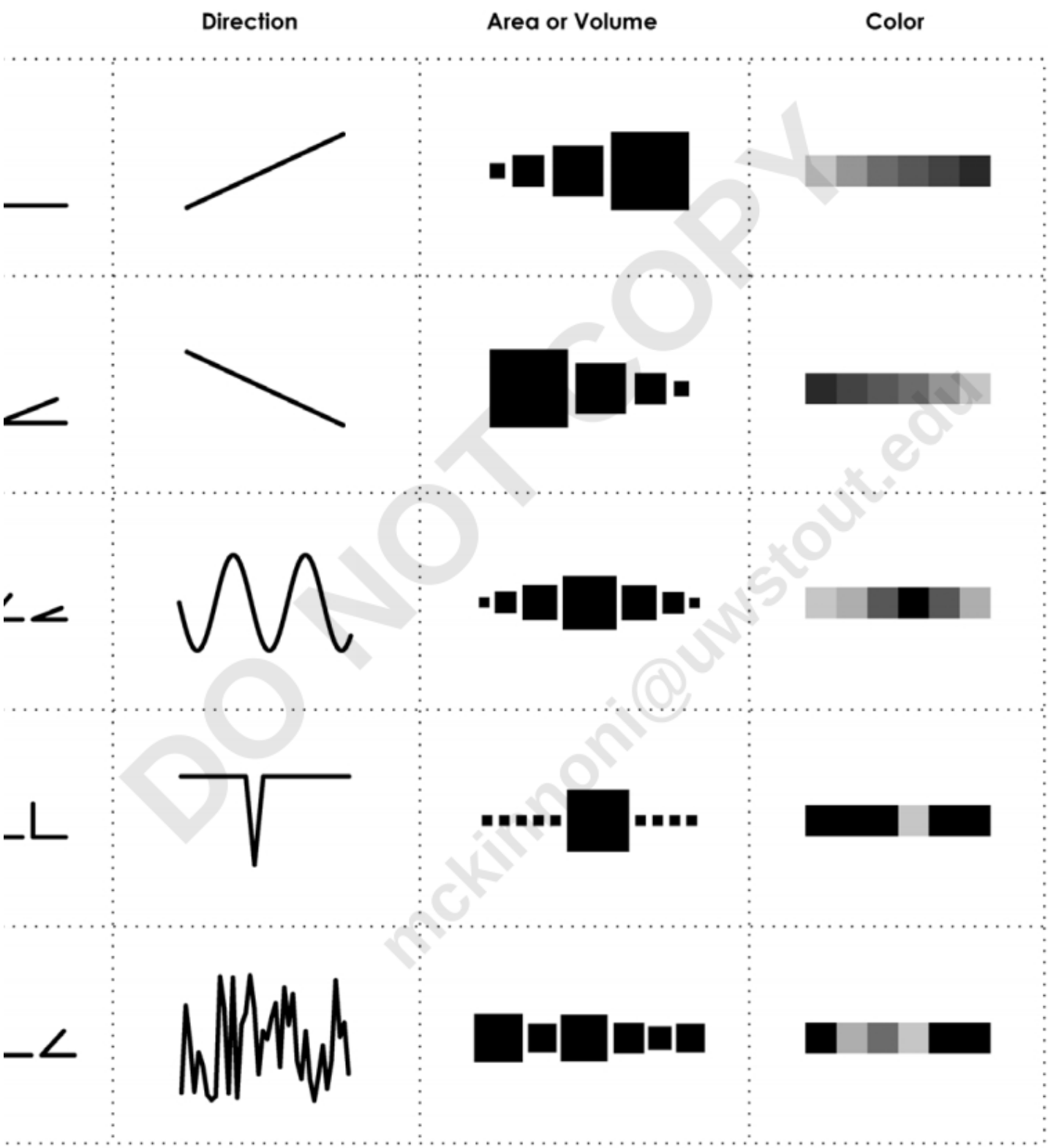
This might be considered a drawback, but it can also be an advantage when you have hundreds of values that vary by orders of magnitude. With symbol plots, you can organize squares and circles in any way you want in two-dimensional space, as shown in Figure 4-8. On the other hand, bar graphs are restricted in that each bar must start at the zero-axis and must extend straight across or upward to the corresponding value.

Note: Because there aren't many categories for each question, the bar chart seems like a better choice in this example, but you don't need to rule out area as a visual cue automatically.

FIGURE 4-4 *(following page) Patterns and visual cues*

Visual Cues

Patterns	Position	Length	Angle
Increase			
Decrease			
Combination			
Outlier			
Noise			

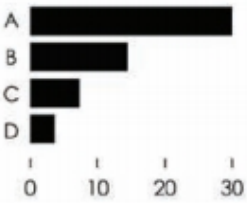


/

Categories

When your data is straightforward, with a value for each category, these are easy to read and create.

Bar graph



With length as visual cue, useful for straightforward comparisons

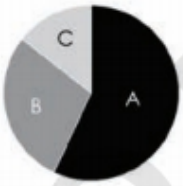
Symbol plot

Can be used in place of bars, but can be hard to see small differences

Parts of a whole

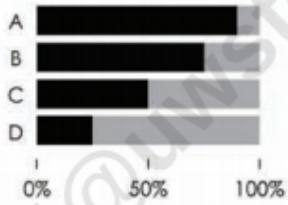
The categorical breakdown within a population can be interesting, and you might want to keep the groups together, although often not essential.

Pie chart



Parts add to 100 percent, typically sorted clockwise for readability

Stacked bar chart

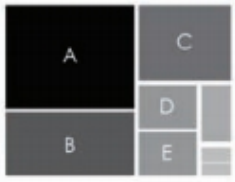


Often used to show poll results and can also be used for raw counts

Subcategories

Data can have a hierarchical structure, which can be important in data interpretation and it often allows for different points of view.

Treemap



Shows hierarchical structure in a compact space, area often combined with color

Mosaic plot

Allows comparisor across multiple categories in one view

FIGURE 4-5 *Visualizing categorical data*

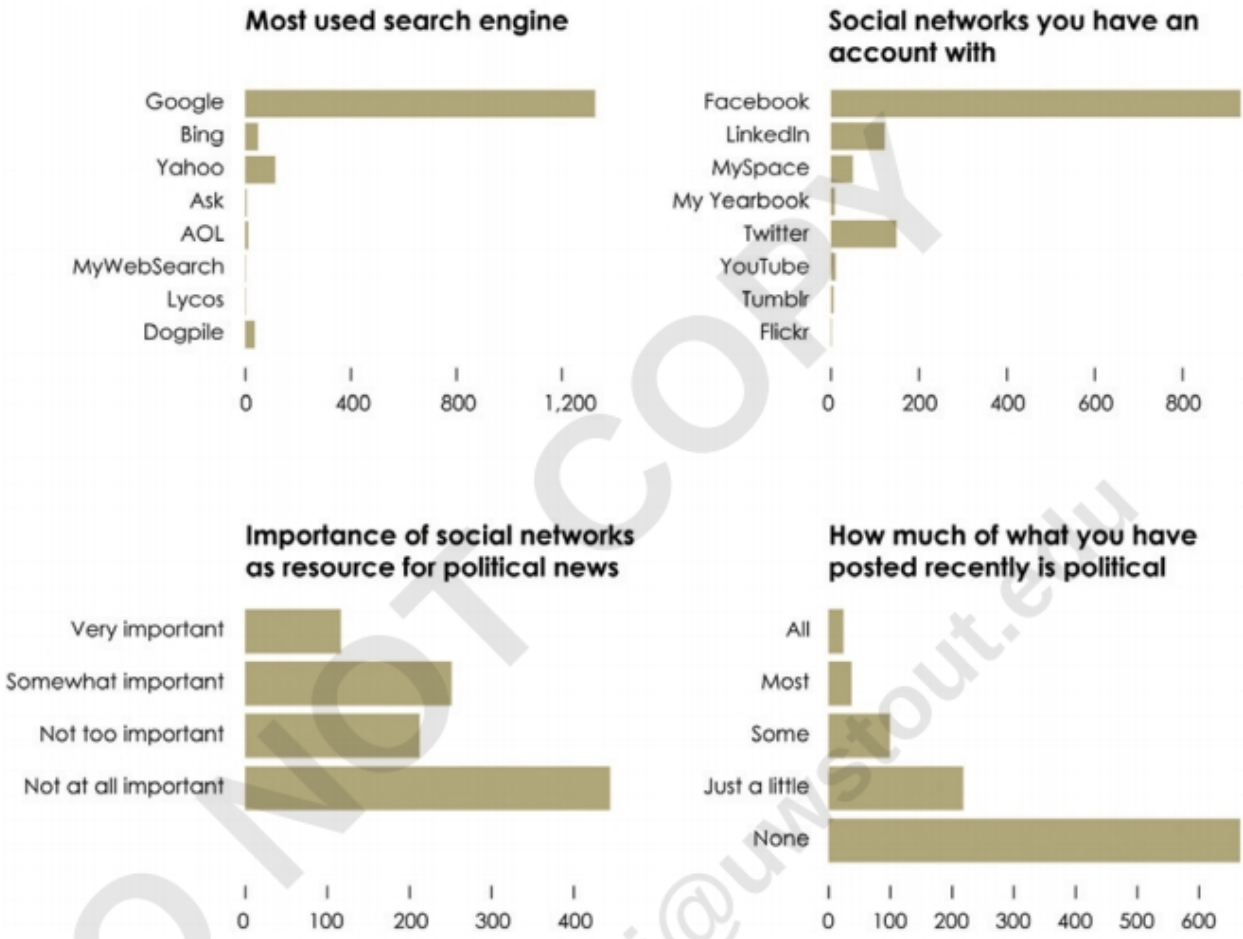


FIGURE 4-6 *Bar graphs for survey results*

PARTS OF A WHOLE

When you put categories together, the sum of the parts can equal a whole. Count everyone in all the states and you have a national aggregate; combine sports divisions and you have a league. Seeing categories as a single unit can be beneficial if you want to see distributions or the spread across a single population.

This is when the pie chart comes into the picture. A full circle represents 100 percent of a whole, and each wedge is a portion of that 100 percent. The sum of all the wedges equals 100 percent. Angle is the visual cue.

Note: You might not be able to get the exact value from a pie chart, but you can still make comparisons when there aren't a lot of categories.

Discussions on whether the pie chart is useful end up running in circles, so you decide if you want to use them. If you do use pie charts, they tend to clutter quickly when you have a lot of categories, simply because there is only so much space in a circle, and small values end up as slivers.

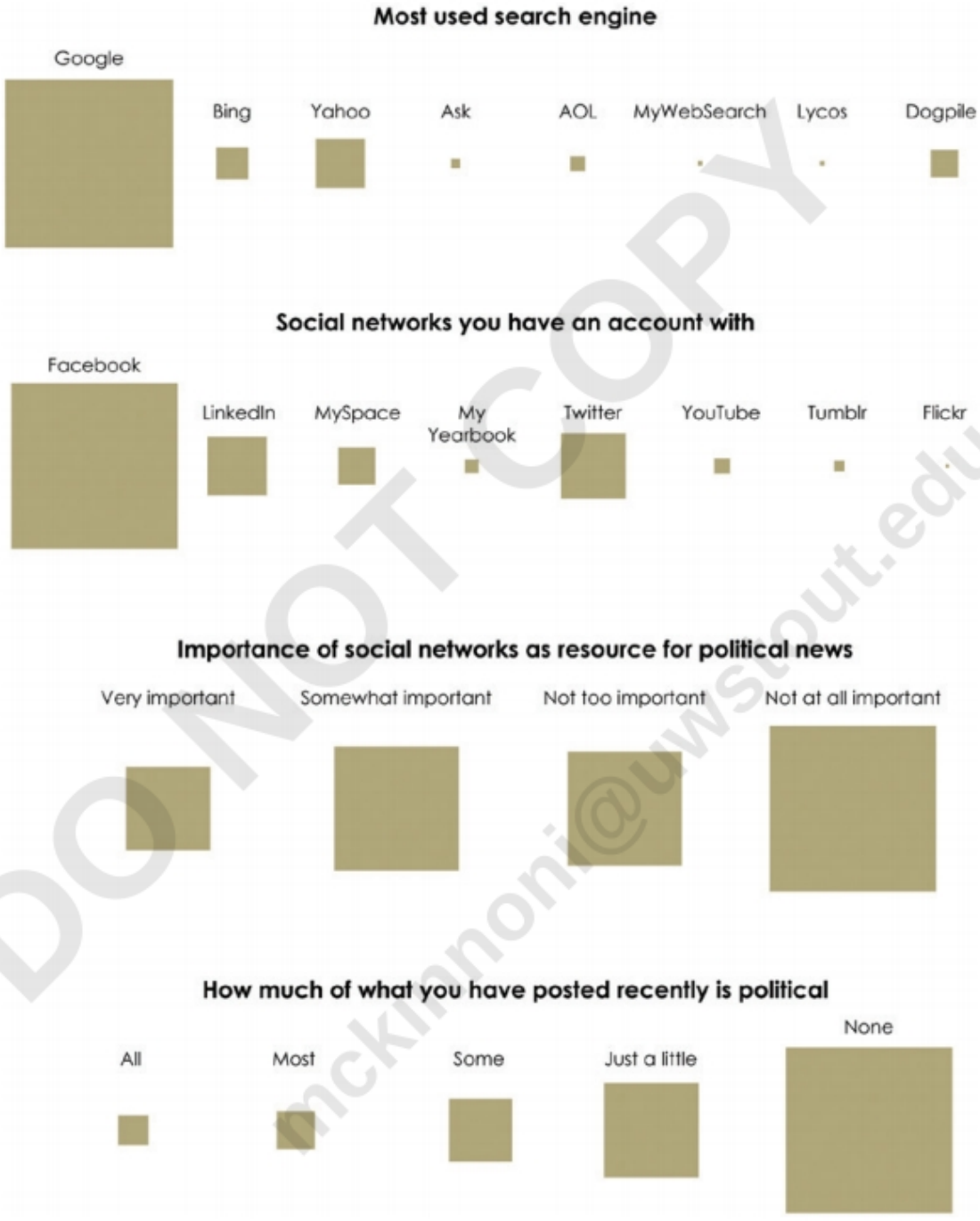


FIGURE 4-7 *Symbols plot for survey results*

Returning to the Pew Research poll on Internet usage, Figure 4-9 shows breakdowns of awareness of targeted advertising online. The first three pie charts show the percentage of respondents who were aware of targeted advertising, those who were okay with it, and those that knew there was a way to limit it, respectively. The next three pie charts show actions that people took, given that they knew they were aware of online tracking.

If you're not fond of pie charts, you can also use stacked bar charts, as shown in Figure 4-10. The full length of the bars represents 100 percent, and each small bar is the equivalent of a wedge in a pie chart.

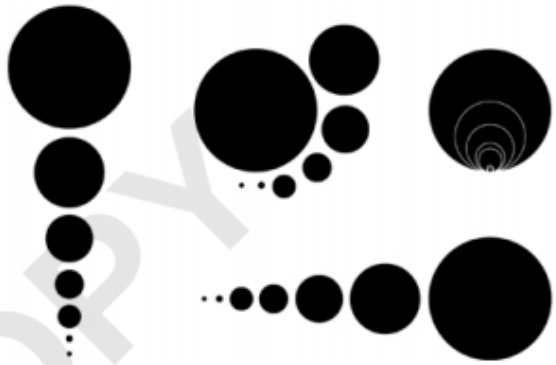


FIGURE 4-8 *Bubble plots organized differently*

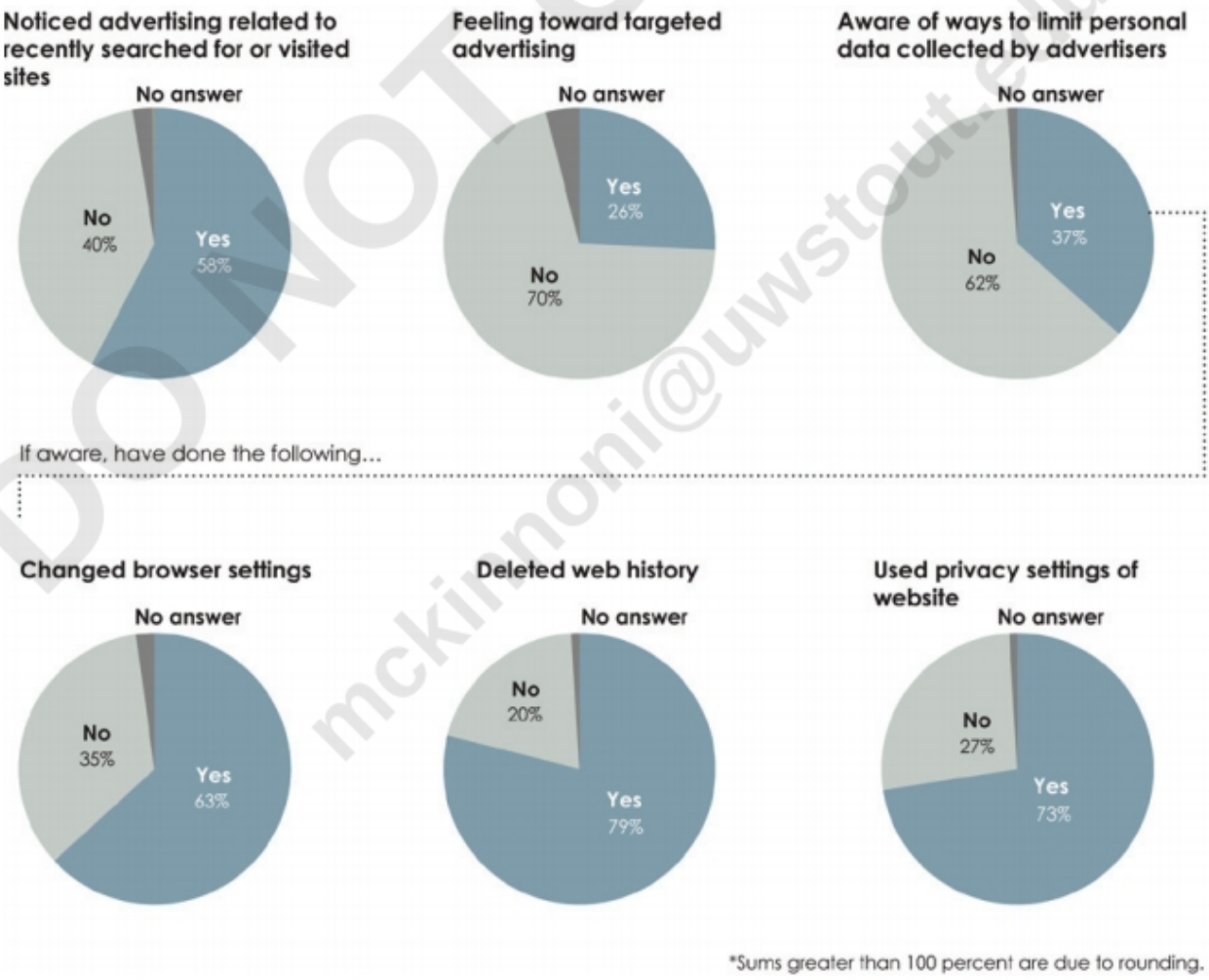


FIGURE 4-9 *Pie charts to show categories*

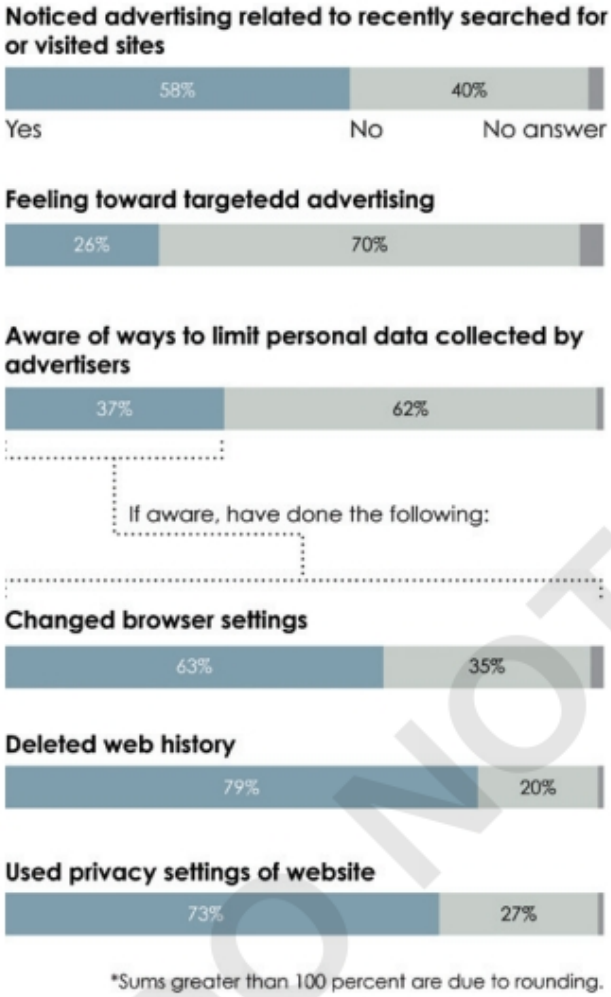


FIGURE 4-10 *Stacked bars to show categories*

Note: "Map of the Market" by SmartMoney is another popular treemap. It shows the status of the United States stock market in real-time. Check it out at: http://www.smartmoney.com/map-of-the-market/

SUBCATEGORIES

Subcategories, the categories within categories (within categories), are often more revealing than the main categories. As you drill down, there can be higher variability and more interesting things to see.

At the least, showing subcategories can make it easier to browse your data, because you can visually jump to the areas that you care most about. For example, you saw categorical hierachy of the news in Marcos Wescamp's newsmap in Chapter 2, "Visualization: The Medium."

As shown in Figure 4-11, you can use a treemap with the Pew Research survey data. It shows those who use the Internet regularly and those who don't. Within the group of people who use the Internet regularly is a grouping of those who used the Internet the day before the survey and those who did not. However, the survey data doesn't work well with a treemap. Whereas newsmap shows a rectangle for each story sized by current popularity, individuals within a survey are equally weighted.

Instead, a mosaic plot, which shows you proportions within categories and category combinations is more fitting. Like the treemap, you can use a mosaic plot with multiple levels of data, but interpretation can get difficult quickly, so start with the minimum and work your way to the more complex.

Figure 4-12 shows the proportion of people in the survey who said they were the parent or guardian of a child younger than 18 living in the household.

The plot looks like one column from a stacked bar graph. The bigger a section, the more people who gave that answer, so from this view, you see most people said no, some said yes, and there were a few who declined to answer.

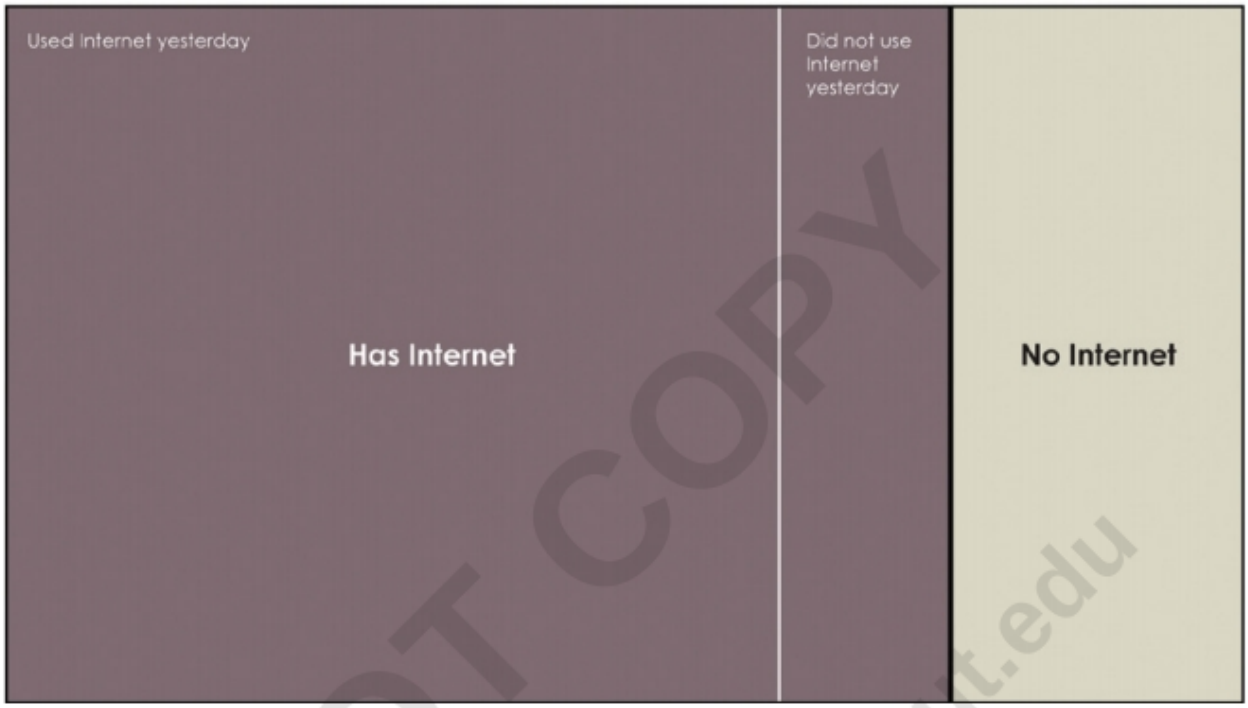


FIGURE 4-11 *Treemap on survey data*

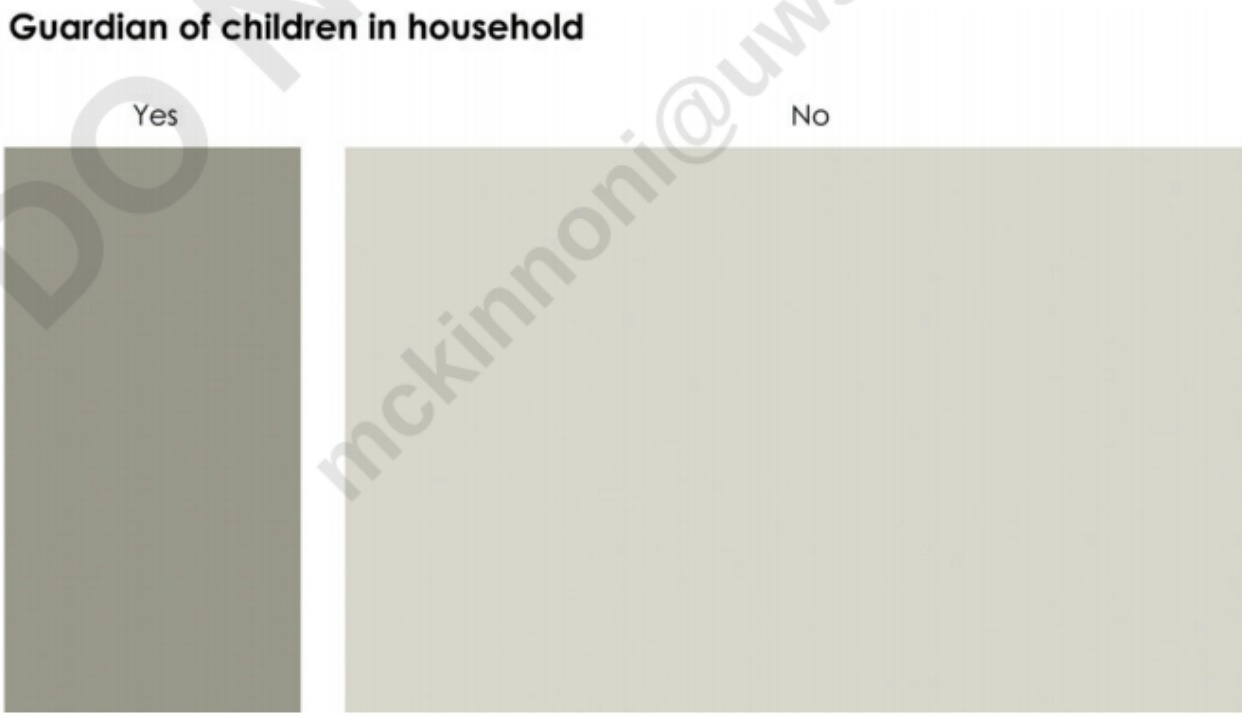


FIGURE 4-12 *Basic mosaic plot with one variable*

What if you want to know the education level of those who are parents or guardians versus those who are not? As shown in Figure 4-13, you can introduce another dimension. It's the same geometry, where more area equals a higher percentage. But now for example, you can see that of those who are parents, a slightly lower percentage were college graduates than those who were not.

You can keep going and bring in a third variable. The orientation of education and parenting are the same, but you can also see e-mail usage. Notice the vertical split on the subsection in Figure 4-14.

You could keep on adding variables, but as you can see, the plot grows more challenging to read, so proceed with caution.

WHAT TO LOOK FOR

With categorical data, you often look for the minimum and maximum right away. This gives you a sense of the range of the dataset, and is easily found with a quick

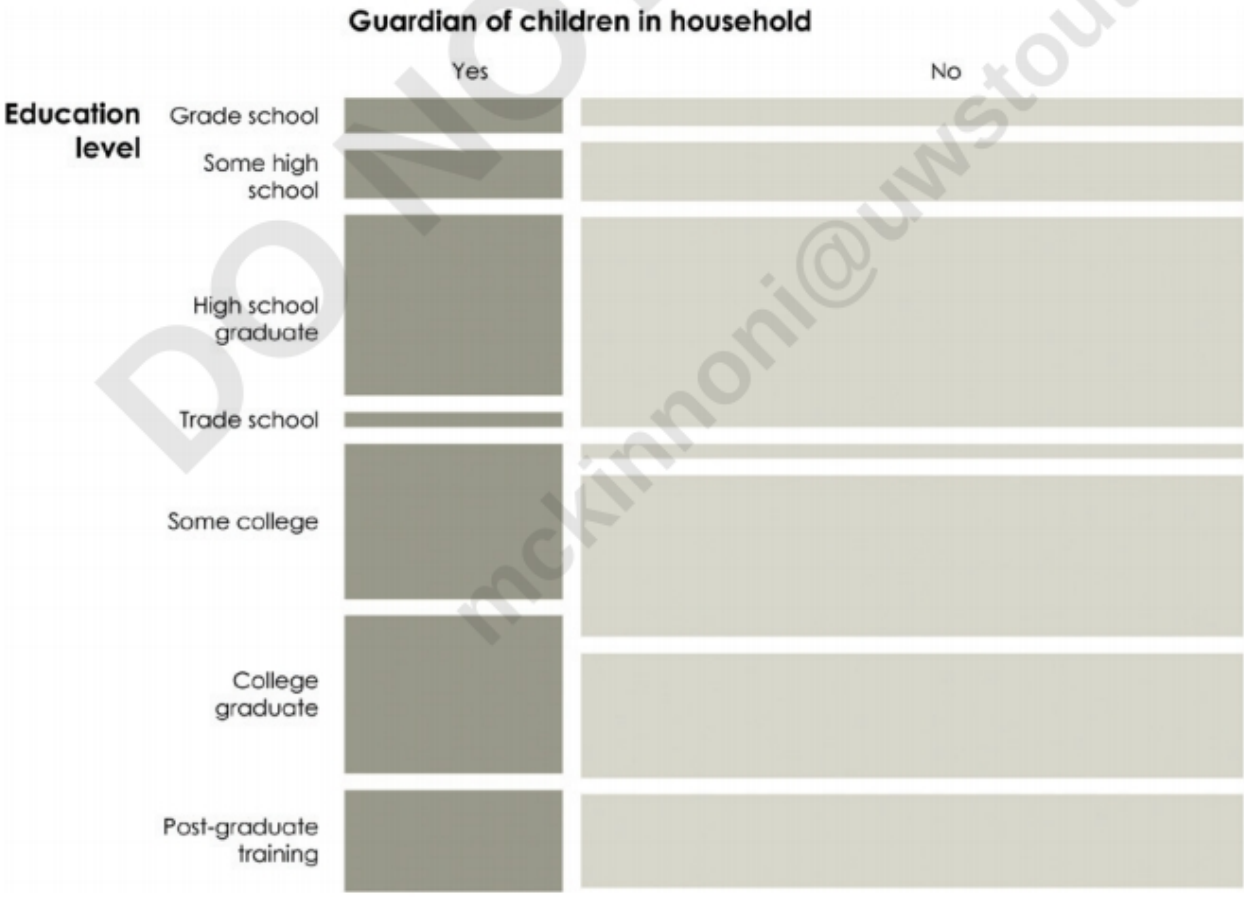


FIGURE 4-13 *Mosaic plot with two variables*

sorting of values. After that, look at the distribution of the parts. Are most values high? Low? Somewhere in between? Finally, look for structure and patterns. If a couple of categories have the same value or high differing ones, it's worth asking why and what makes the categories similar or different, respectively.

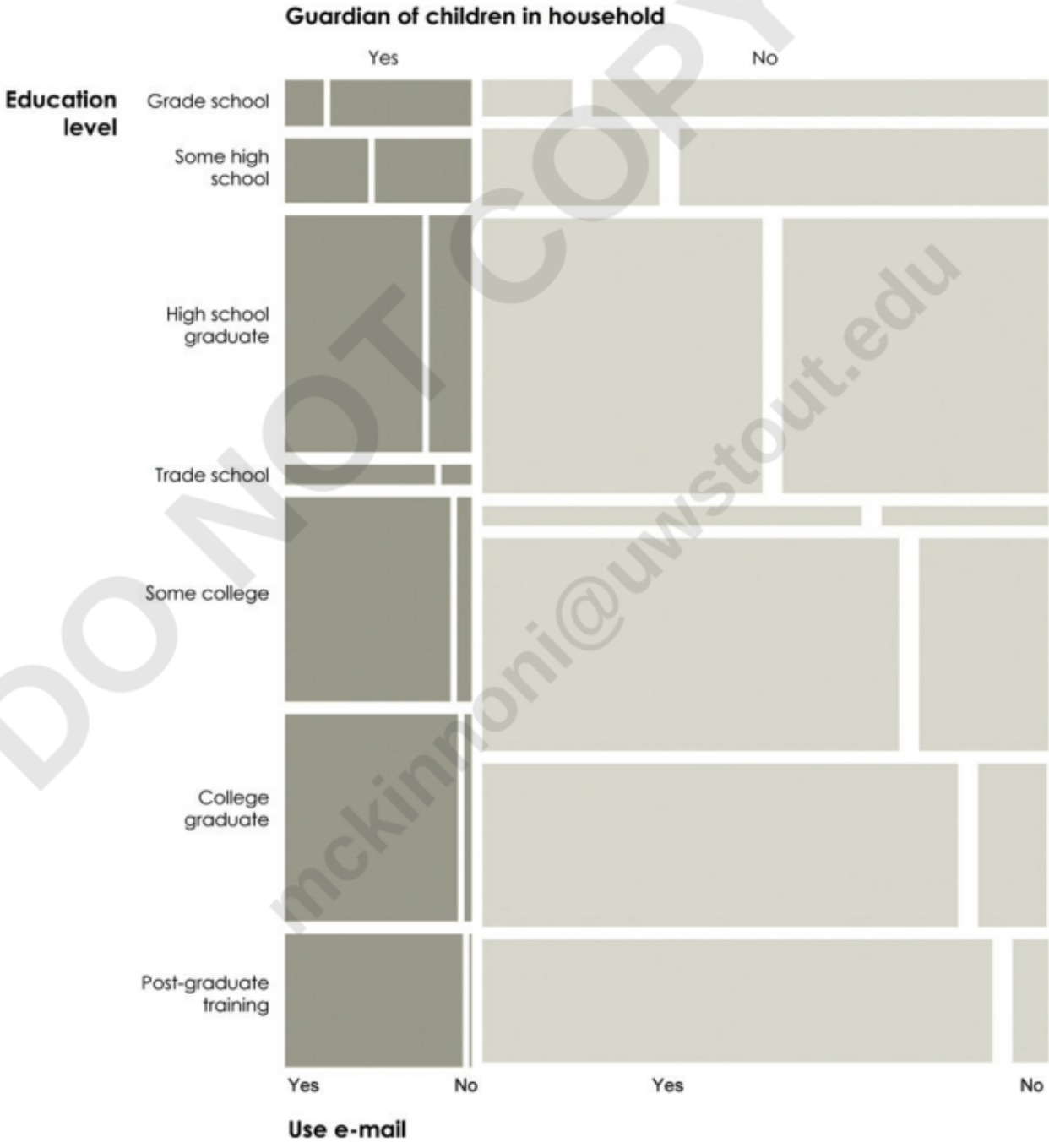


FIGURE 4-14 *Mosaic plot with three variables*

VISUALIZING TIME SERIES DATA

Time passes. Things change, people change, and places change. You can feel time through the sunrise and sunset, your clocks and watches, and the coffee you need to drink when you wake up. When you visualize time series data, as shown in Figure 4-15, your goal is to see what has passed, what is different, and what is the same, and by how much. Compared to last year, is there more or less? What are possible explanations for the increase, decrease, or nonchange? Is there a recurring pattern, and is that good or bad? Expected or unexpected?

As with categorical data, the bar chart is a straightforward way to look at data over time, except instead of categories on one of the axes, you use time.

Figure 4-16 shows the unemployment rate in the United States from 1948 to 2012, according to the Bureau of Labor Statistics. On top is the rate month-to-month, and because there is a high point density, it looks like a continuous area. On the other hand, the graph on the bottom shows only the unemployment rate in January of each year, which allows for space in between bars and makes it easier to distinguish individual points.

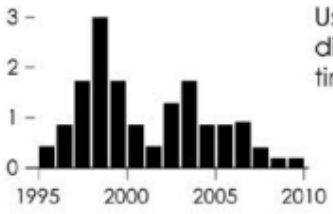
In Chapter 1, you saw how car crashes vary over time, and how you can explore time series data at different granularities. The same applies here. You can look at data hourly, daily, annually, by decade, by century, and so on. Sometimes the data format dictates the level of detail because the metric was measured, say, only every 5 years. However, if for example you had measurements by the hour, a high variability might obscure a trend that's more obvious if you take a step back and look at your data by the day.

Usually the magnitude of change between segments of time is more interesting than the value at each point. Although you can interpret trends from a bar graph, you must visually calculate rates. You look at one bar and compare it to the ones before and after.

Time series

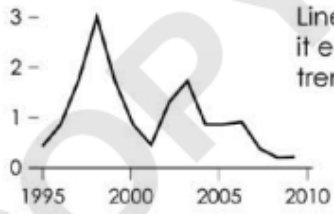
There are a variety of ways to see patterns over time, using cues such as length, direction, and position.

Bar graph



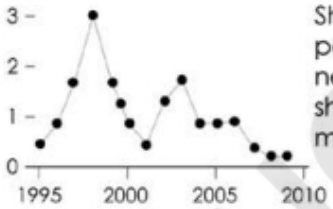
Useful for discrete points in time

Line chart



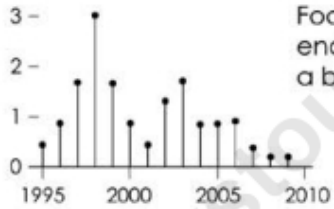
Lines can make it easier to see trends

Dot plot



Shows distinct points but might need line to show trend if not much data

Dot-bar graph

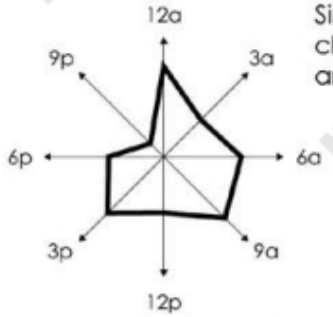


Focuses more on endpoints than a bar graph

Cycles

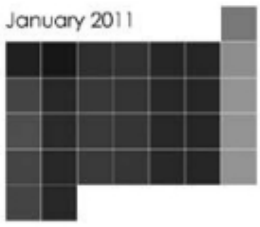
TIme of day, day of the week, and month of the year repeat themselves, so it is often beneficial to align the segments in time.

Radial plot



Similar to line chart but wraps around

Calendar



Patterns for days of week seen more easily than views above

FIGURE 4-15 *Visualizing time series data*

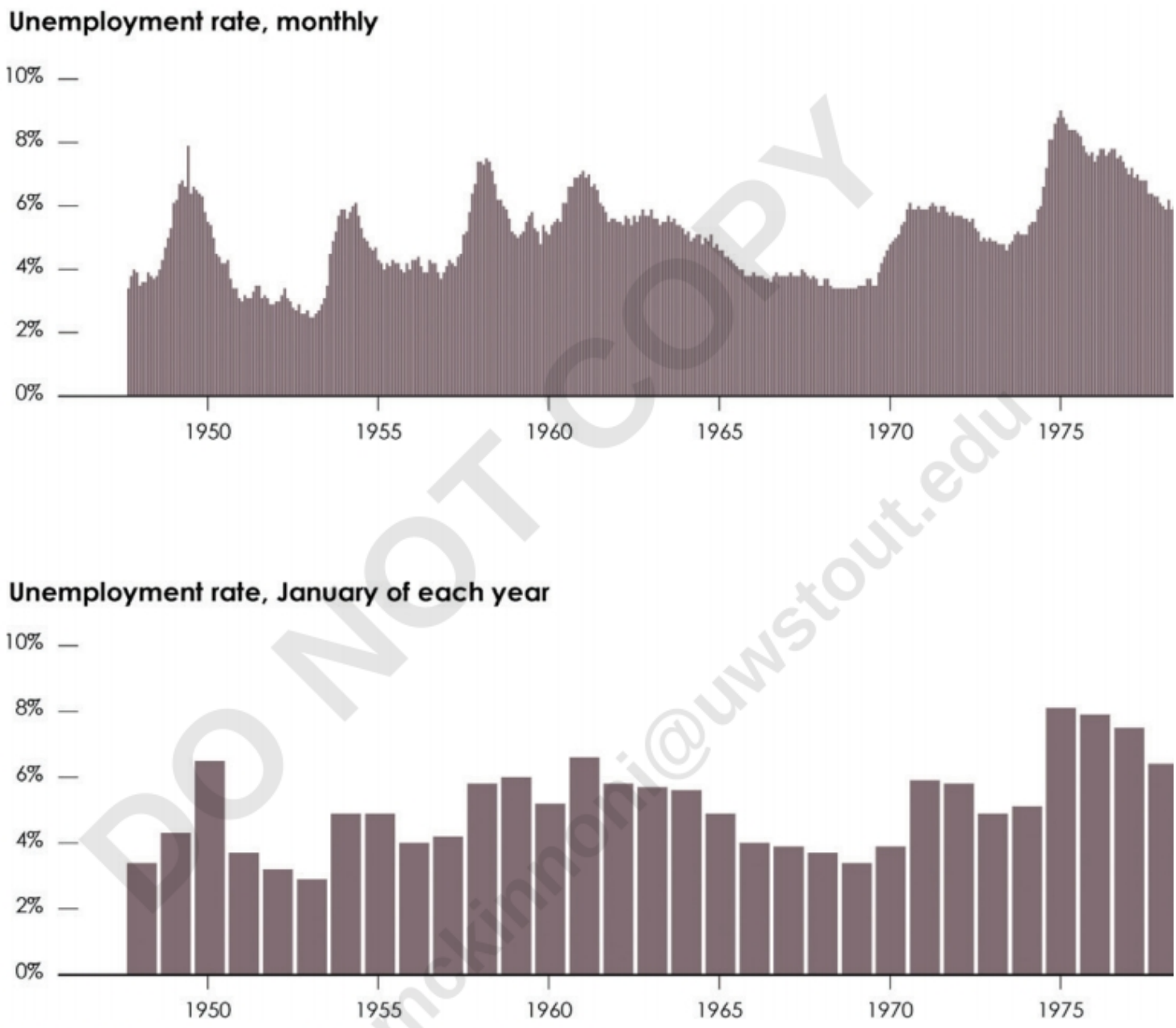
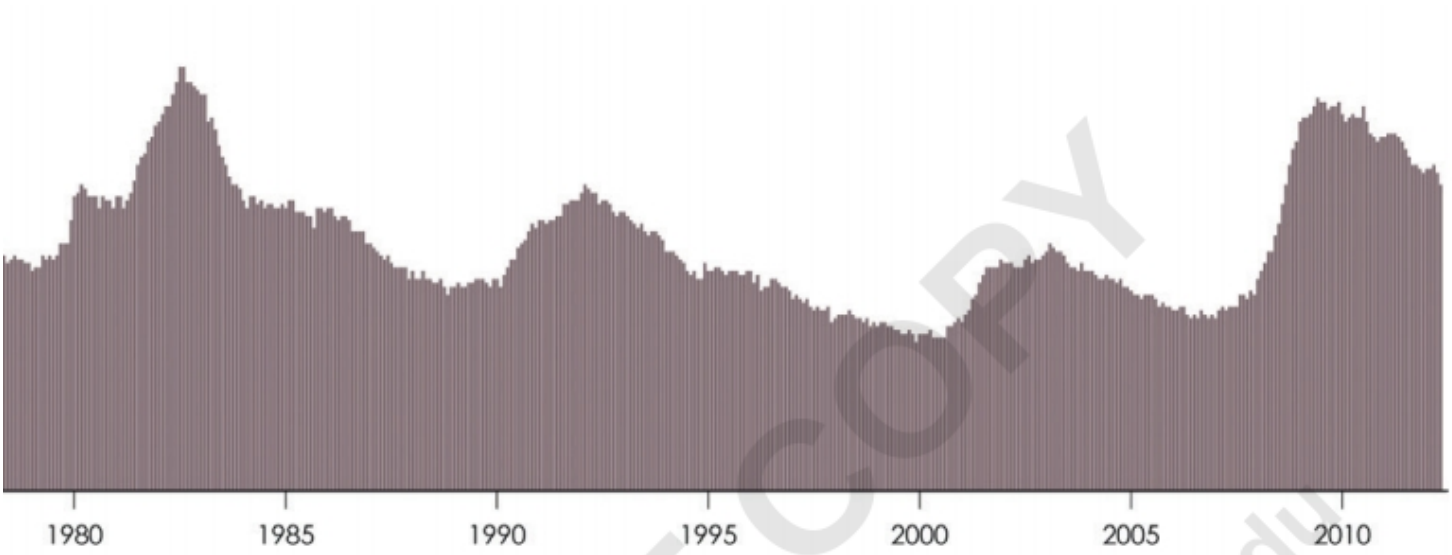
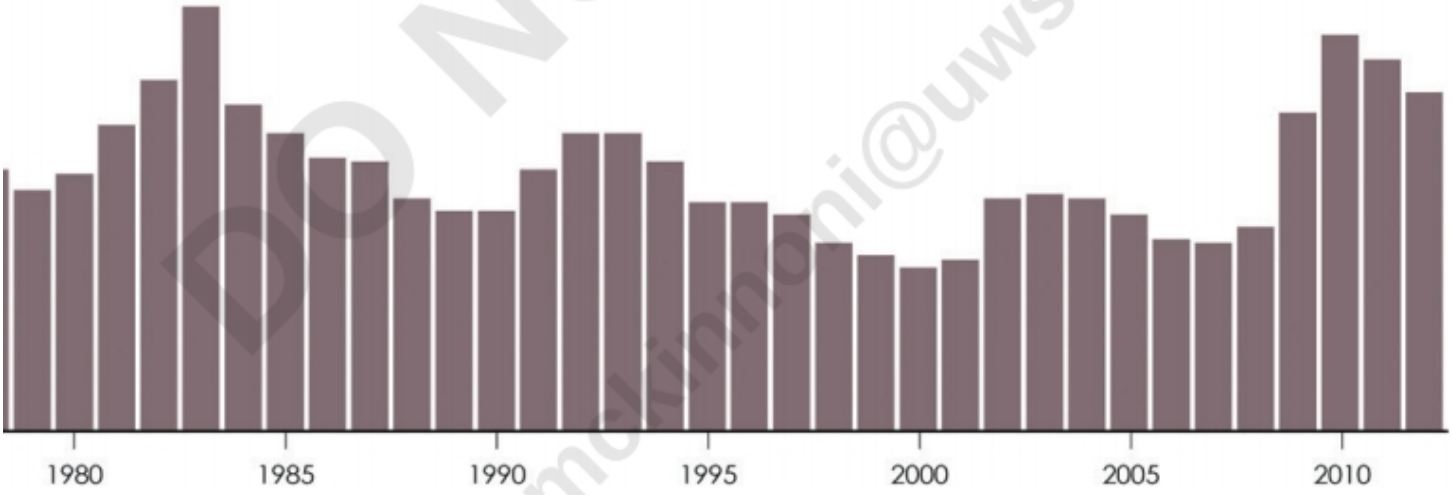


FIGURE 4-16 *Bar graphs for time series data, with monthly on top and annual on the bottom*





However, where the bars look like a continuous unit (refer to Figure 4-16), it's easier to distinguish changes because you can see the slope, or the rate of change in between points. It's even easier to see the slope when you use a continuous line, as shown in Figure 4-17. The line chart shows the same data as the bar graph on the same scale, but change is directly displayed via direction as a visual cue.

A dot plot can be used in the same way, as shown in Figure 4-18. Again, the data and axes are the same and the visual cue is different.

Like bar charts, dots put focus on each value, and trends can be harder to see. Although in this example, there are enough data points, so you don't need to mentally fill in the gaps. If the data were more sparse, such as in Figure 4-19, changes are less obvious.

When you connect the sparse dots with a line, as shown in Figure 4-20, the focus of the plot shifts again.

If you care more about an overall trend than you do about the more specific monthly variability, you can fit a LOESS curve to the dots, as shown in Figure 4-21, instead of connecting every dot. The closer you fit the curve to the dots, the more it resembles Figure 4-17.

Note: *LOESS* (or *LOWESS*) stands for *locally weighted scatterplot smoothing*. It's a statistical technique created by William Cleveland, which fits a polynomial function to a subset of the data at different points. When combined, they form a continuous line.

Of course, the chart style you choose depends on your data, and although it might seem like a grab bag of options at first, you get a feel for what type of chart to use with practice. It's not an exact science (or computers could do all the work) and options can vary a lot even if you have datasets that look similar.

For example, the previous charts on unemployment rate provide a historical view of the past few decades. You can see peaks and valleys, periods of recession such as in 2001 and from 2007 to 2009, and an overall picture of changing rates. If you were only interested in the five highest peaks and what happened immediately after them, the exploration would take a different route.

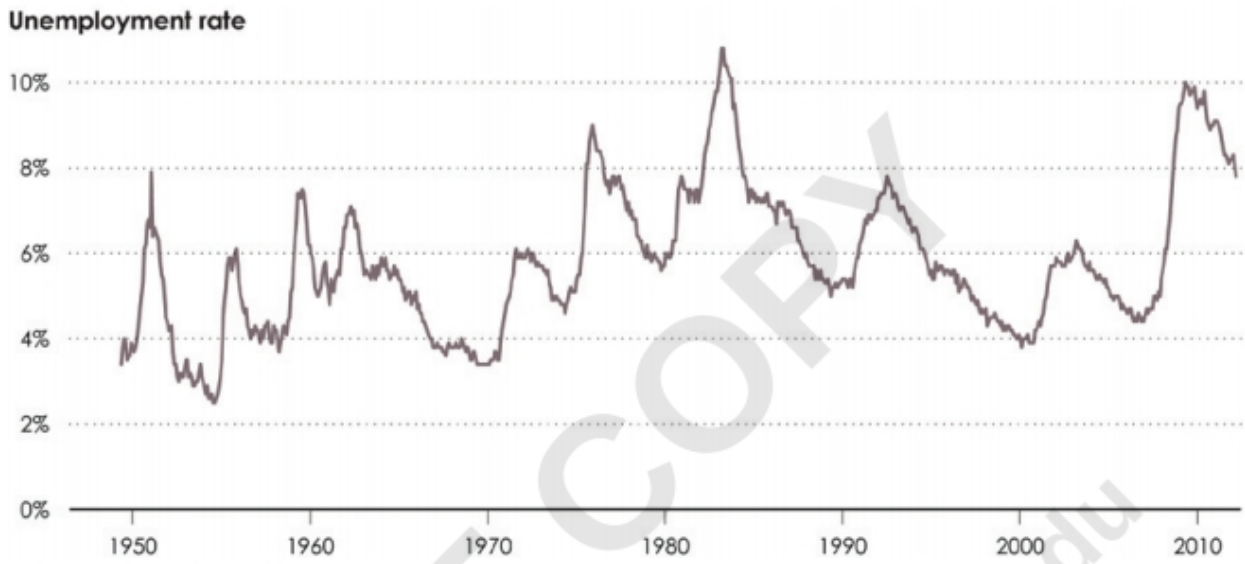


FIGURE 4-17 *Line chart to show time series*

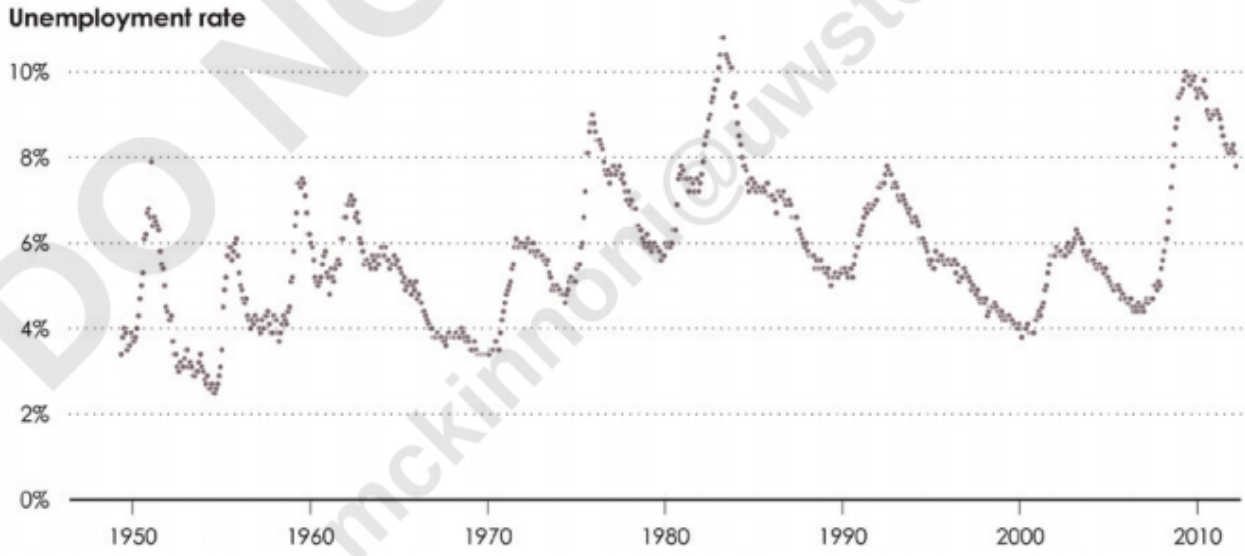


FIGURE 4-18 *Dot plot to show time series*

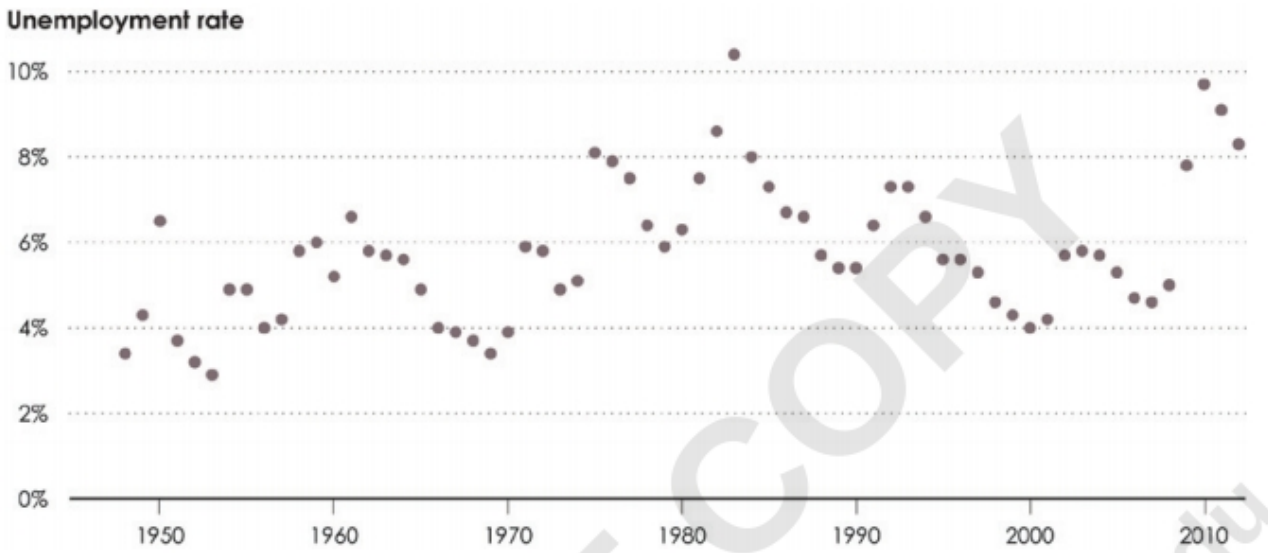


FIGURE 4-19 *Sparse dot plot*

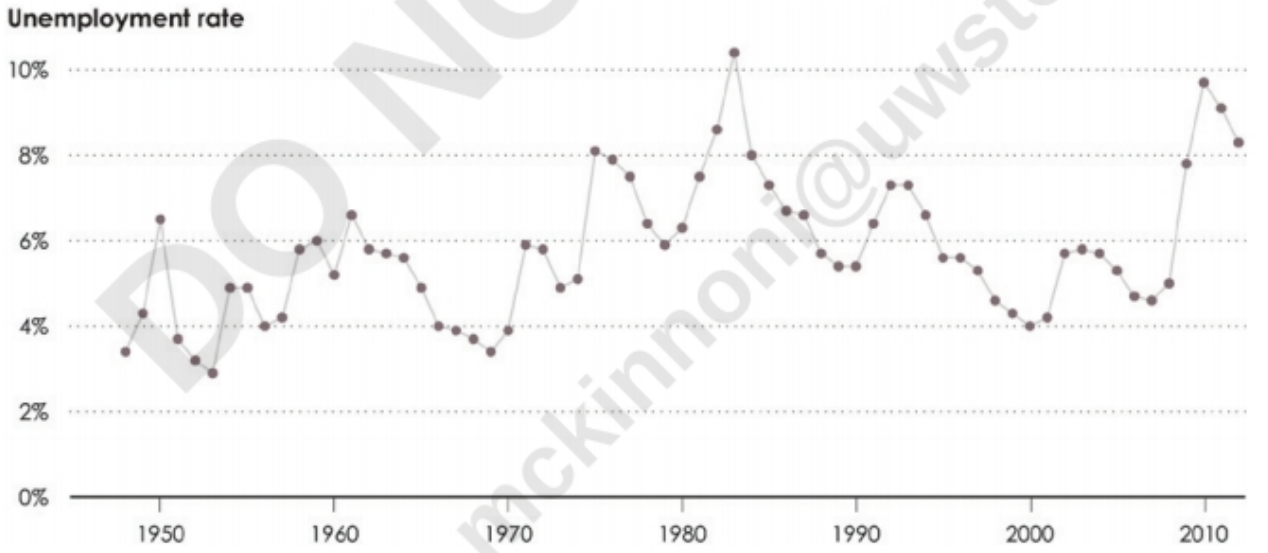


FIGURE 4-20 *Sparse dot plot with connecting line*

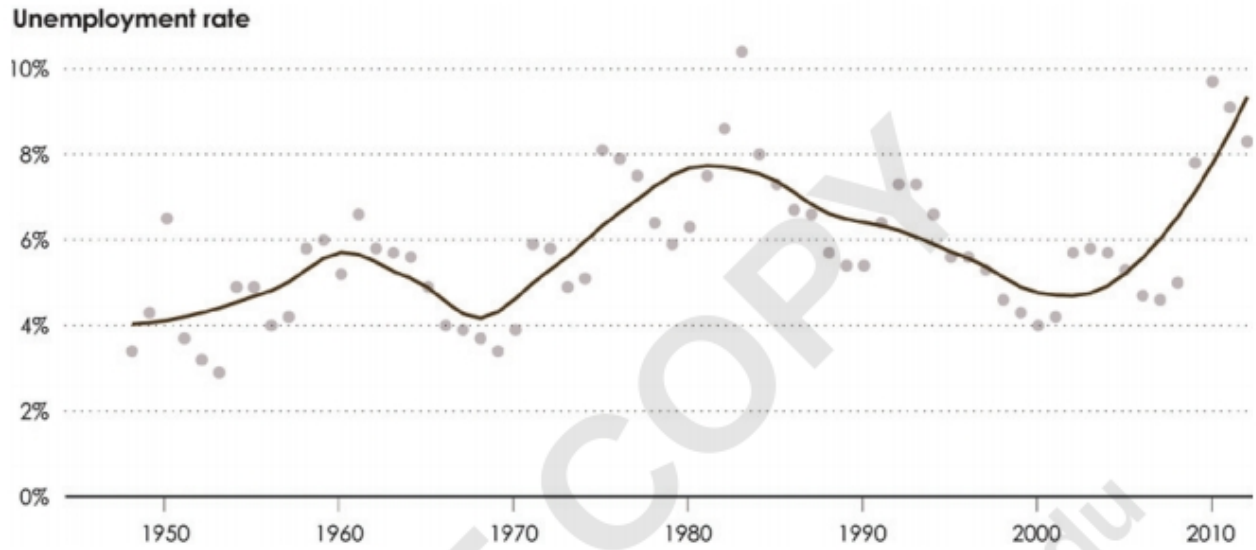


FIGURE 4-21 *Fitted LOESS curve*

CYCLES

A number of factors feed into the economy and affect the unemployment rate, so there aren't regular intervals in between significant increases. For example, the data doesn't suggest that unemployment goes up to 10 percent every 10 years. However, there are a lot of things that repeat themselves on regular intervals. Students get summer breaks and people often take summer vacations, and lunch is typically around noontime, so the restaurant around the corner that makes burritos bigger than your face usually has a longer line during that hour.

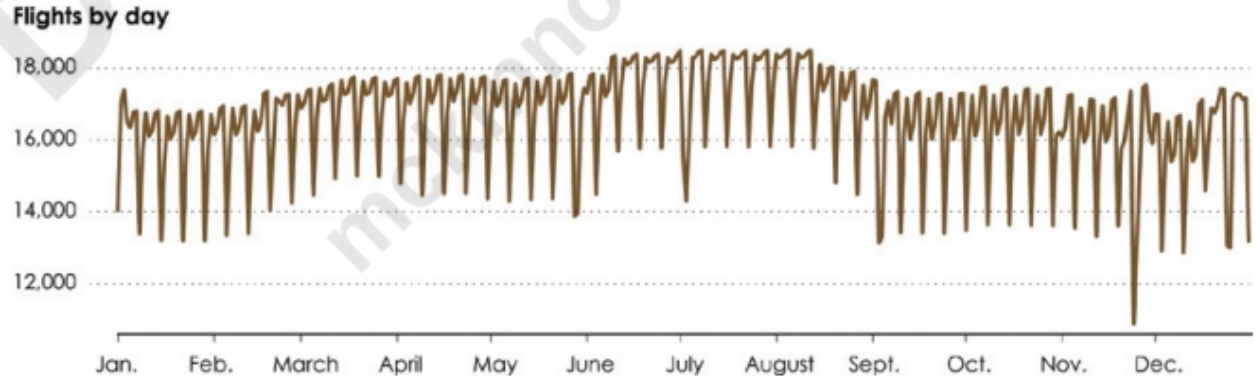


FIGURE 4-22 *Weekly cycle*

You saw repetition in the crashes data: More people travel during the summer months; more people leave work around 5 in the afternoon and head home; and more accidents occur on Saturday than any other day of the week. This information can be used to make sure there are enough people staffed during each day of the week and when to allot vacation times.

Flight data from the Bureau of Transportation Statistics shows a similar cycle, as shown in Figure 4-23. The chart shows a weekly cycle, with the fewest flights on Saturdays and typically the most flights on Fridays (a contrast to car crashes).

You can see the same pattern if you switch to a polar axis, as shown in the star plot in Figure 4-23. The data starts at the top, and you read the chart clockwise. The closer to the center a point is, the lower the value, and greater values move further away.

Note: The *star plot* is also commonly referred to as a *radar chart*, *radial plot*, and *spider chart*.

Because the data repeats itself, it makes sense to compare like days of the week to each other. For example, compare all Mondays. It's hard to do this when time is visualized as a continuous line or loop, but you can split the days

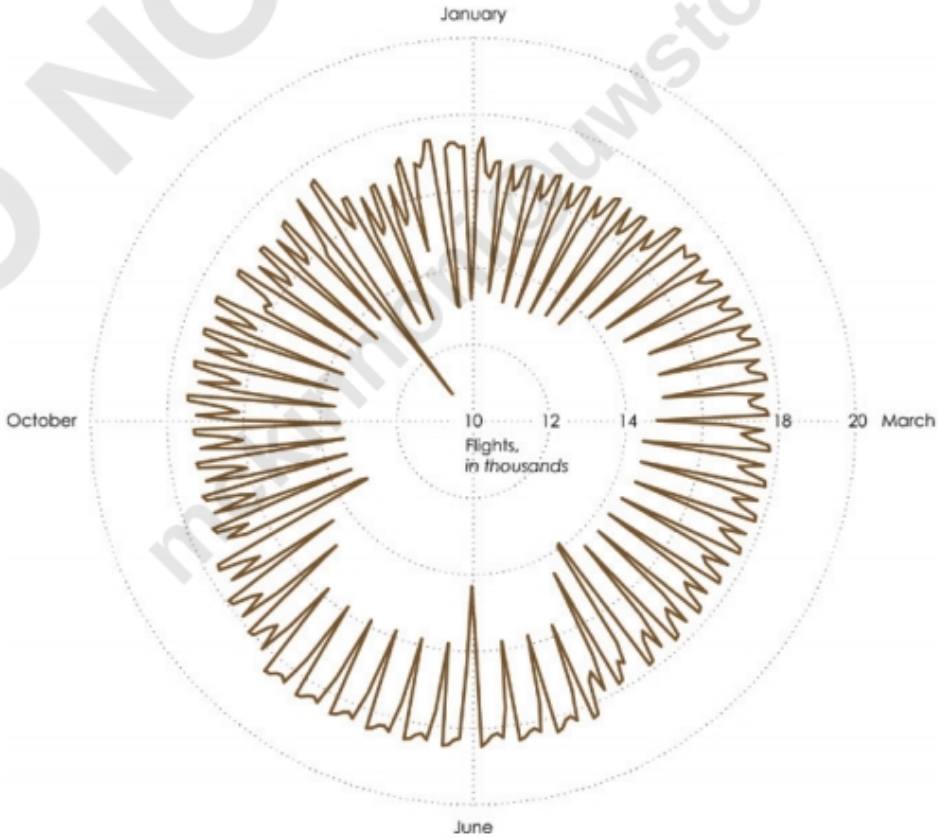


FIGURE 4-23 *Star plot to show time series data*

into weekly segments so that you can directly compare cycles, as shown in Figure 4-24, with both the line chart and star plot.

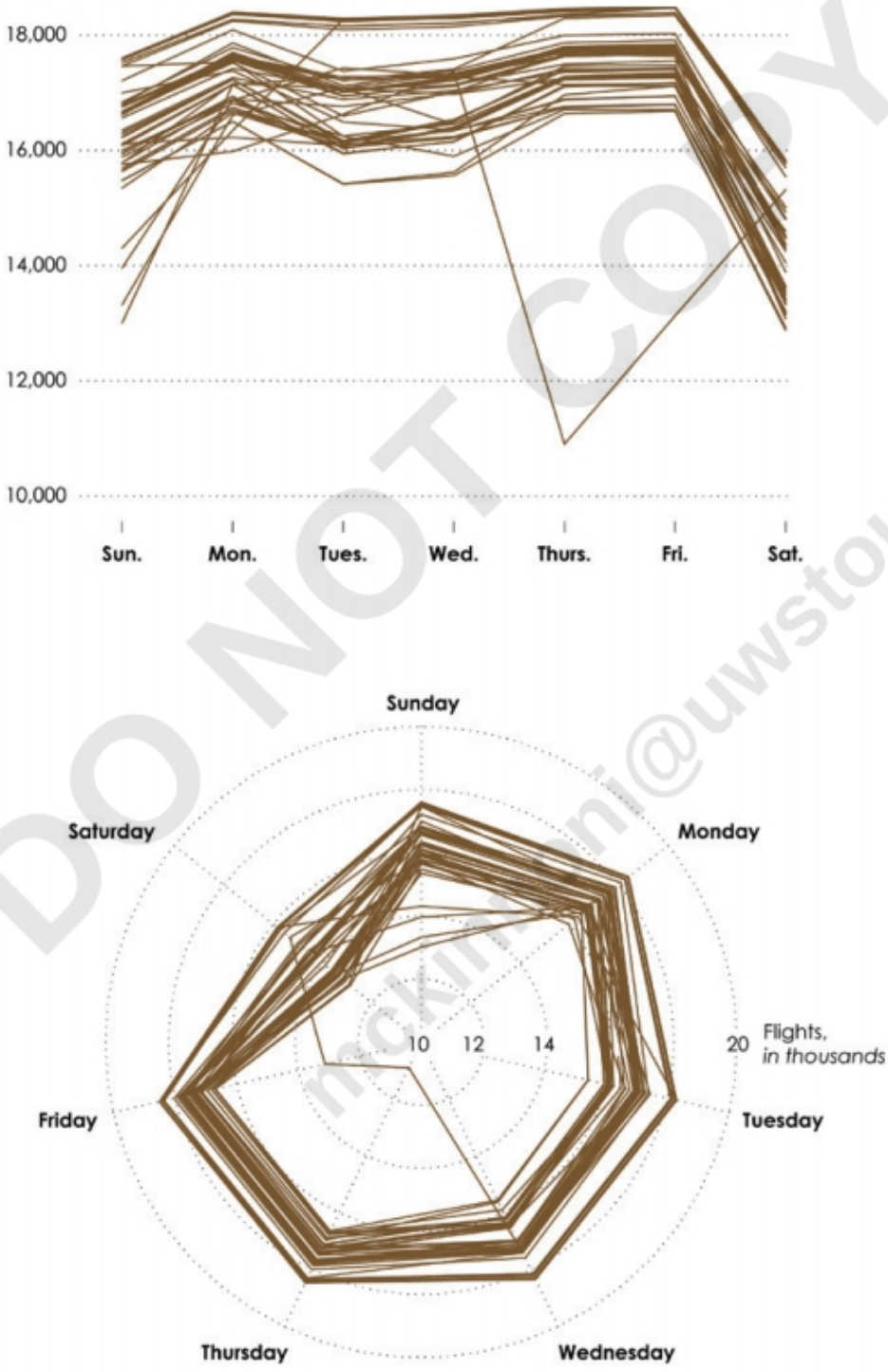


FIGURE 4-24 *Line chart and star plot to show overlapping cycles*

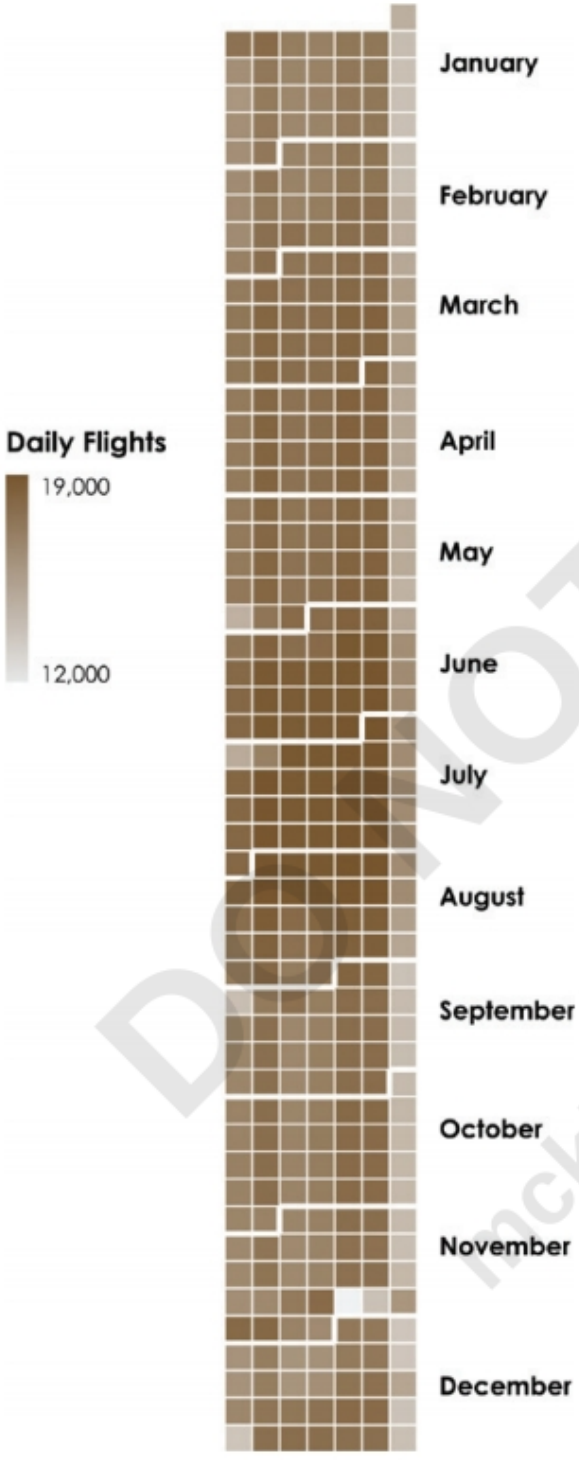


FIGURE 4-25 *Calendar heat map*

There's clearly a regular day-to-day pattern. The thin lines that overlap give the appearance of thicker bands.

There are also some obvious outliers. Thursdays and Fridays are much lower than the rest of the year, and four Sundays also appear lower. Take a moment to think about why travel might be lower than usual on those days.

The obvious way to figure out the dates of the outliers is to go back to the data and look at the minimums for each day. That's always there for your reference. You can also refer to the linear view in Figure 4-17 for a rough idea, but is there a way to directly get the dates and see the data? Figure 4-25 shows the data in a familiar calendar format. The first column is Sunday, the second is Monday, and so on, to Saturday at the end.

The advantage of the calendar heat map over the line chart is that, along with seeing cycles as you scan top to bottom, it's easy to see specific days in rows and columns, so it's easier to reference what day of the year each value is for. The first three low-volume Sundays precede holidays that fall on a Monday in the United States: Memorial Day on May 30, Independence Day on July 4, and Labor Day on September 5. The last Sunday was Christmas. As for the low-volume Thursday and Friday, that was Thanksgiving weekend in November.

A disadvantage of the calendar is that color is the visual cue, and it can be hard to see small differences. It's easier to compare positions on the line chart. So there are trade-offs between different views, but then again, there's no harm in looking at your data from all angles.

Note: The calendar heat map is an intuitive layout that you see often in your everyday life, but is underused as an exploratory tool. It can come in handy.

WHAT TO LOOK FOR

Generally speaking, look for changes over time. More specifically, note the nature of the changes. Are the changes relatively a lot or are they small? If they're small, is the change still significant? Think of possible reasons for what you see over time or sudden blips and if they make sense. The change itself is interesting, but more importantly, you want to know the significance of a change.

VISUALIZING SPATIAL DATA

Spatial data is easy to relate to because at any given moment—as you read this sentence—you have a sense of where you are. You know where you live, where you've been, and where you want to go.

There is a natural hierarchy to spatial data that allows, and often requires, you to explore at different granularities. Far out into space, Earth looks like a small, blue dot with little to see, but as you zoom in, you see land and large bodies of water. There are continents and oceans. Zoom in again, and you get countries and seas, then provinces and states, counties, districts, cities, towns, neighborhoods, all the way down to an individual household.

Global data is often categorized by country and national data by states, provinces, or territories. However, if you have questions about variation across blocks or neighborhoods, such high-level aggregates won't do you much good. So again, the exploration route you choose depends on the data you have or the data you can get.

The most obvious way to explore spatial data is with maps, which place values within a geographic coordinate system. Figure 4-26 shows some of your options, of which there are many.

If you care only about individual locations, you can place dots on a map, as shown in Figure 4-27. The map simply shows the 30 busiest airports in the United States, based on the number of outgoing flights in 2011. As you might expect, the busy airports are in or near major cities such as Los Angeles, Washington, DC, New York, and Atlanta.

Note: A map isn't always the most informative way to visualize spatial data. Often, you can treat regions as categories, and a bar graph might be more useful than seeing a location.

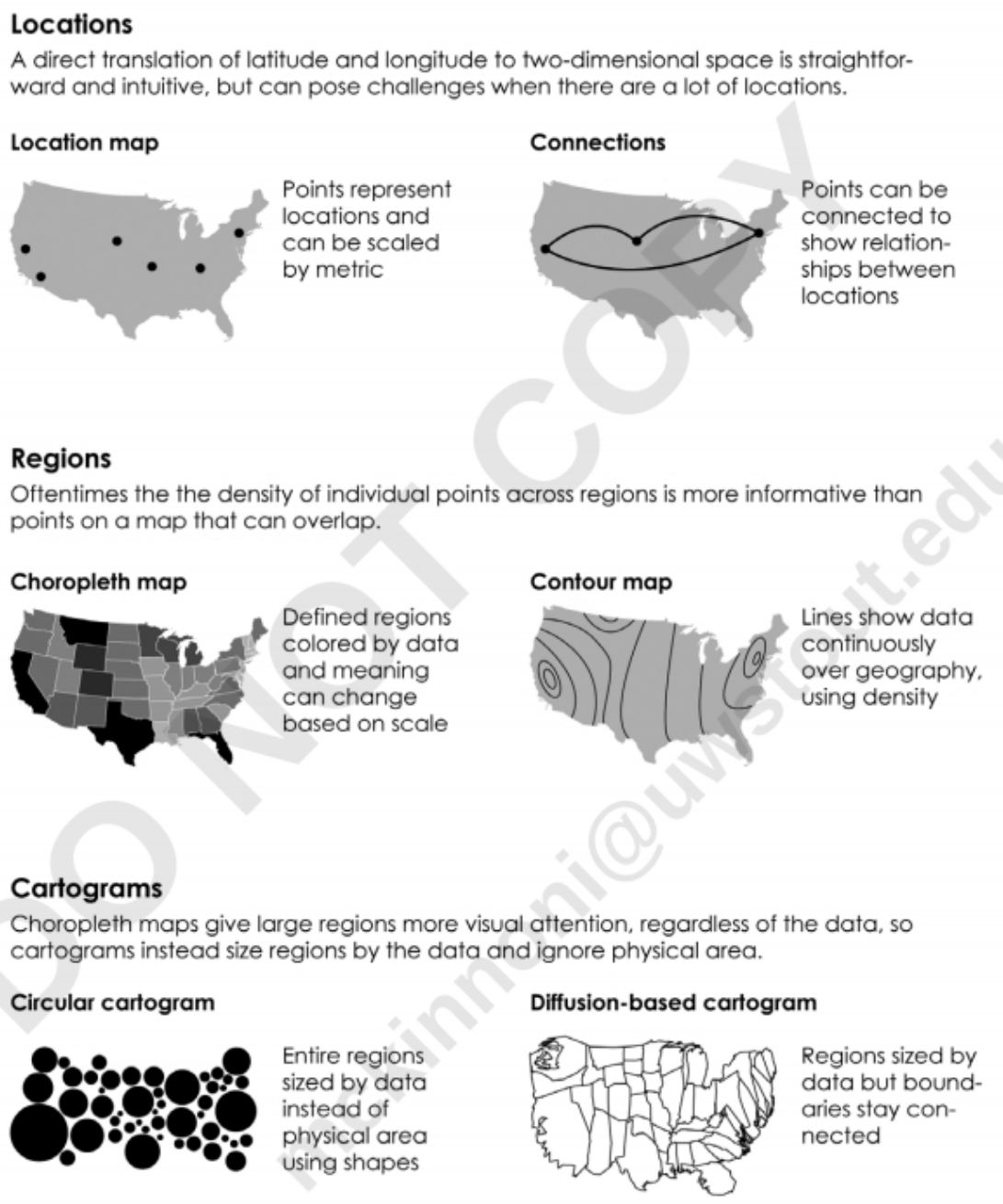


FIGURE 4-26 *Visualizing spatial data*

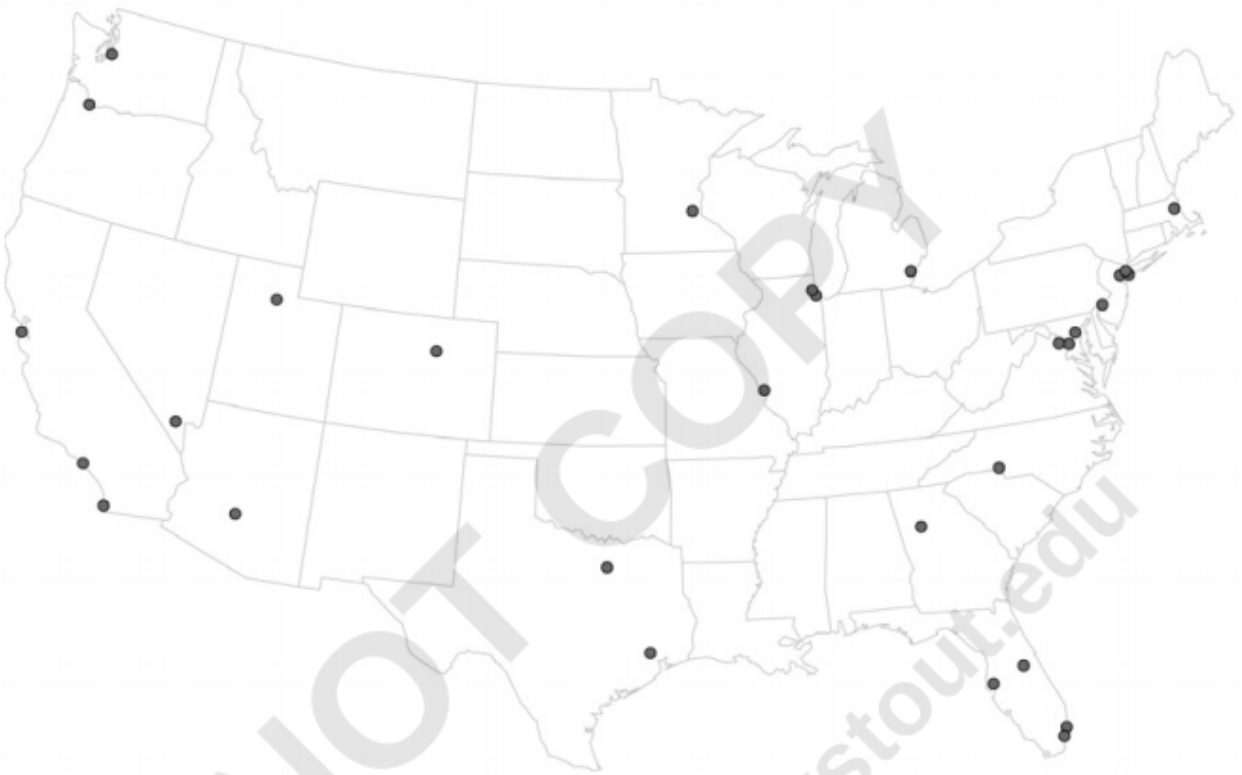


FIGURE 4-27 *Dots in a geographic coordinate system*

Figure 4-28 uses bubbles for the airports, sized by the number of outgoing flights. So, with the addition of an area as visual cue, you don't just see where the busiest airports are, but also how busy they are relative to each other. Atlanta International served the most outgoing flights in 2011, followed by Chicago O'Hare, Dallas-Fort Worth, Denver, and Los Angeles.

Rather than separate locations, you might want to explore connections between locations. For example, in recent years, people have visualized global friendships on social network sites such as Facebook and Twitter. It's one thing to see where people like to use the sites, but it's another to see how they interact.

With the flight data, you already saw counts for outgoing flights via bubbles on a map, but where did those flights go to? Each flight has an origin and a destination. Figure 4-29 shows these connections. The brighter a line is, the

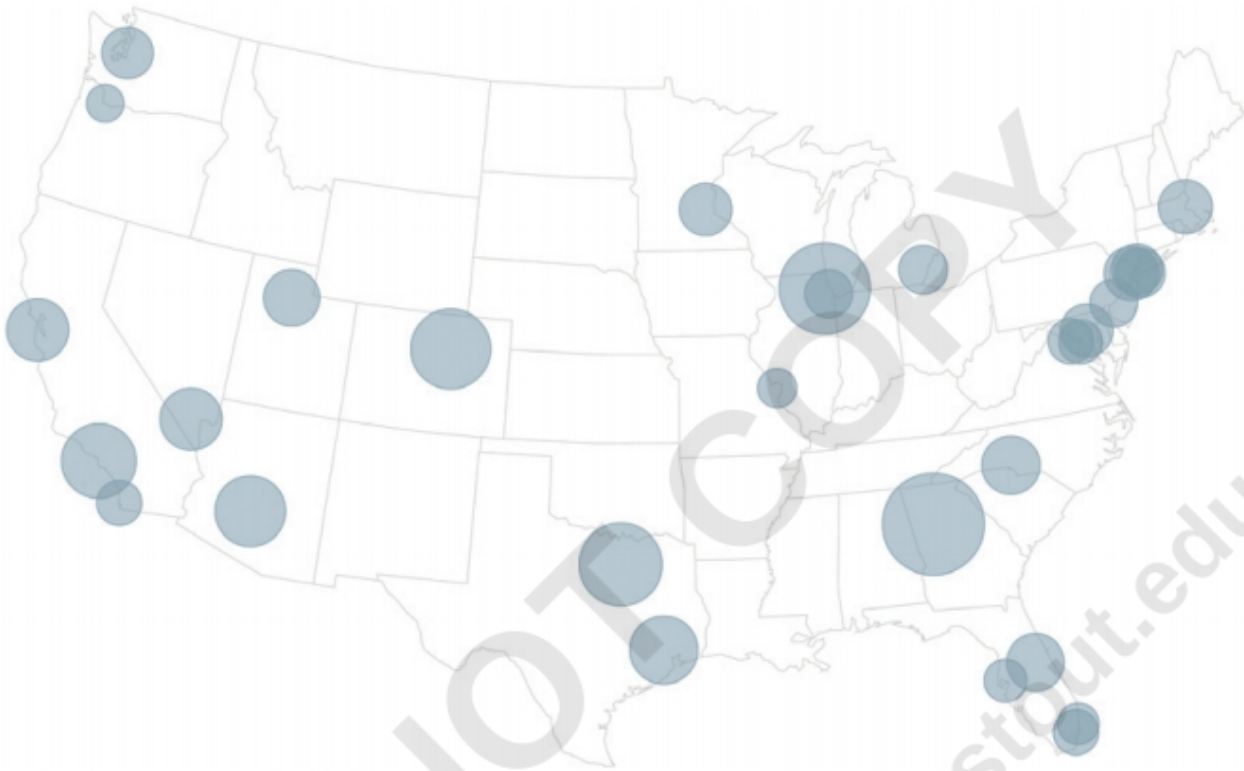


FIGURE 4-28 *Bubbles to show an additional metric on a map*

more flights that went to and from those two airports. Busier airports also appear where there is a higher density of flights.

It's fun to see patterns emerge when you plot a lot of data at once. The map represents more than 6 million domestic flights in 2011, and you gain a rough idea of where people flew to and from. But there's more you can take away from this data by splitting it into categories. For example, map flights by airline, as shown in Figure 4-30, and you see the data with a new dimension.

Note: When you have a lot of data, it is often to your benefit to split it into groups so that you can see details more clearly.

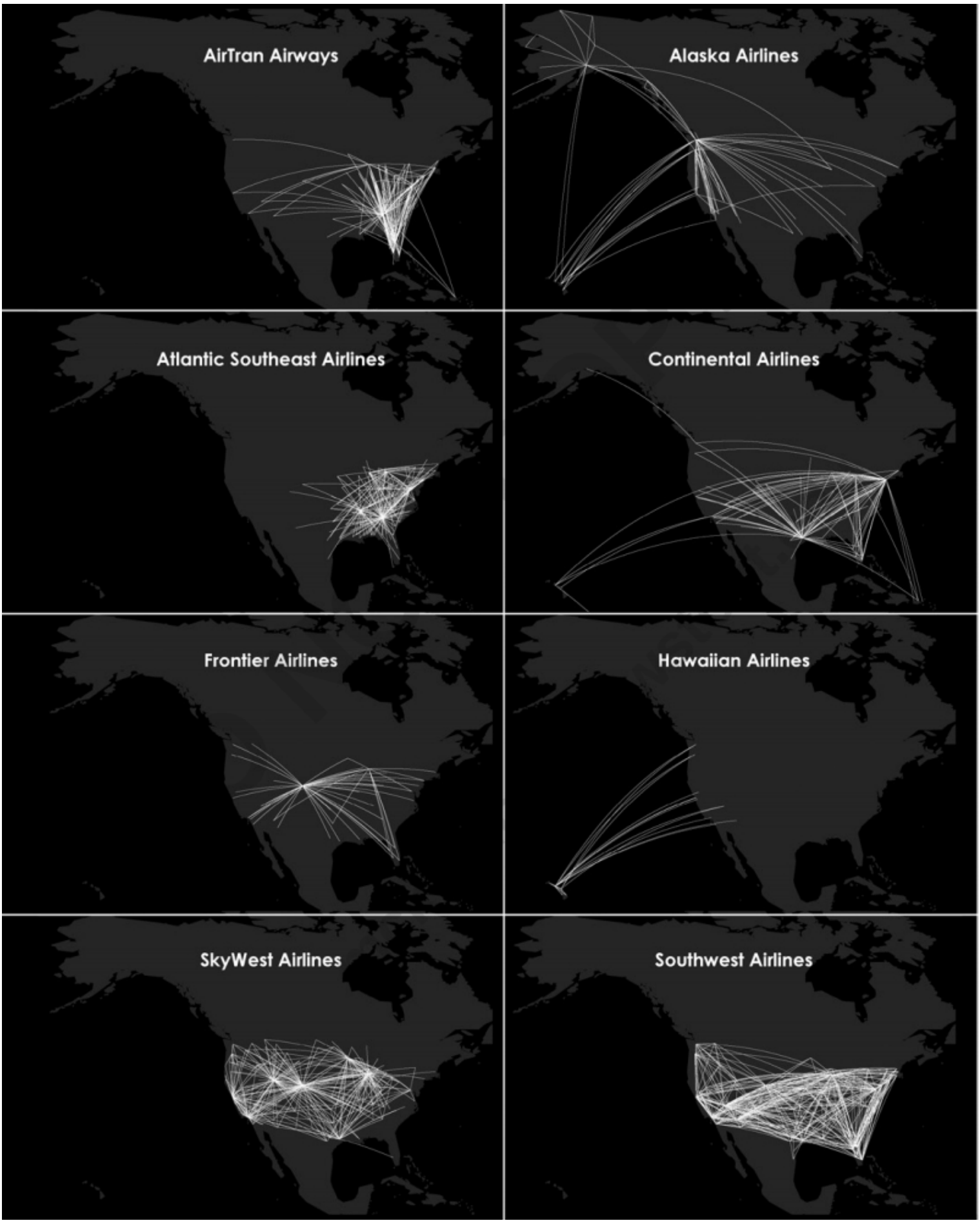
Hawaiian Airlines flies only from the west coast to the islands; Atlantic Southeast Airlines is true to its name; Southwest stays within the contiguous United States; and Delta flies to a number of places, but you can see their major hubs in Atlanta, New York, Detroit, and Salt Lake City.

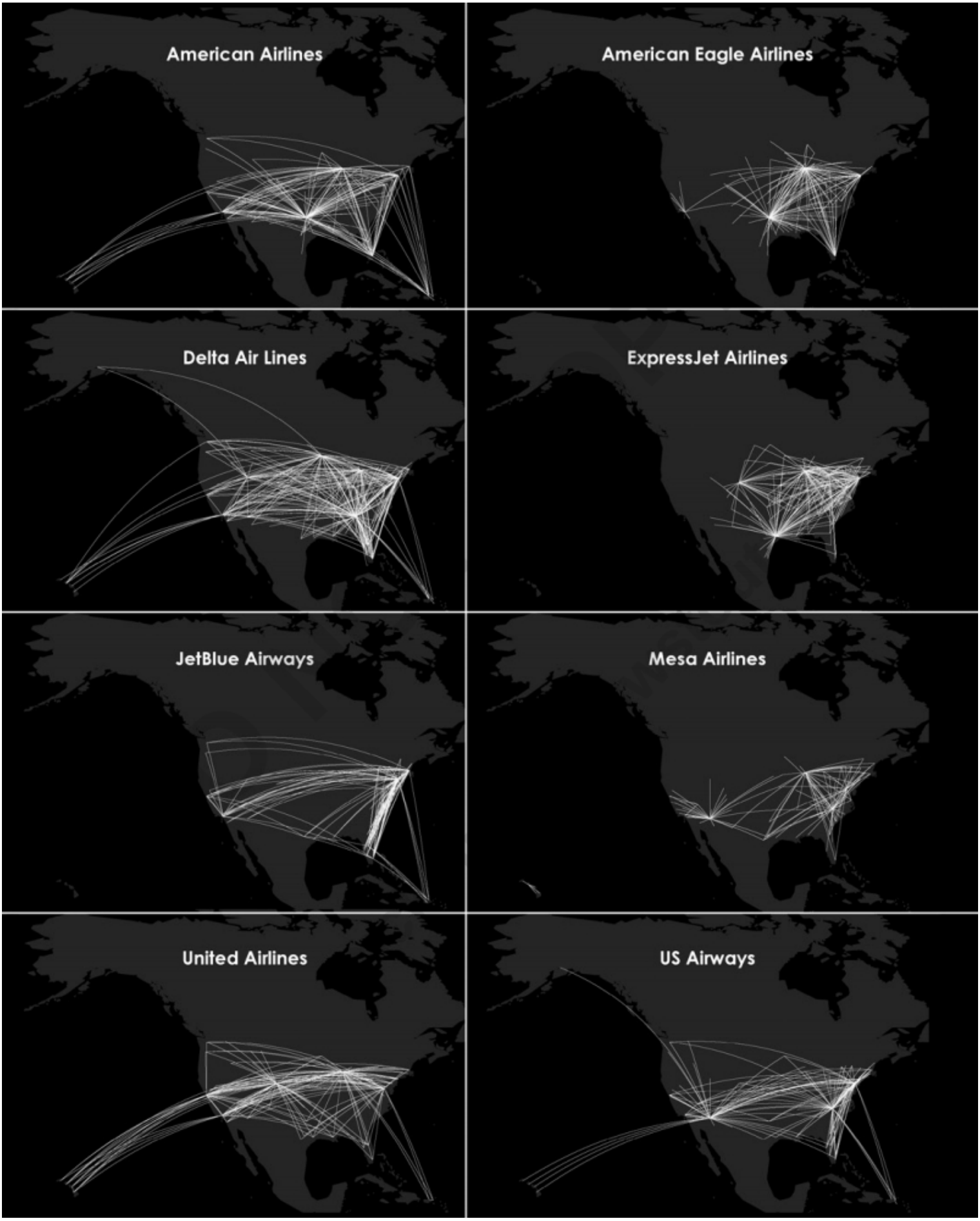
FIGURE 4-29 *Connections between locations*

REGIONS

To maintain the privacy of individuals and to keep personal addresses hidden, it's common to aggregate spatial data before releasing it. Sometimes it's not possible to make estimates at a higher granularity because it would be too big of an undertaking. For example, it's rare to see global data more than country-specific because it's difficult to get a big enough sample in every country for such high detail.

FIGURE 4-30 *(following page) Categorizing data for more specific views*





/

Why not combine studies if they estimate the same thing? If methodology is different, it can be hard to make a case that the results are comparable.

Other times it just makes sense to provide data in aggregate because people want to compare regions. For example, if you work with open data, you often see estimates by country, state, or county. Although less specific, you can still extract information from aggregated data.

Choropleth maps are the most common way to visualize regional data in a spatial context. The method uses color as its visual cue, and regions are filled based on the data. Higher values are typically represented with higher saturation and lower values with lower saturation, such as the map in Figure 4-31.

The map shows estimated national gas prices around the world. The darker the shade of brown, the higher the price per gallon. Gray indicates that there was no data available for that country. Prices are relatively high in Europe and Africa, compared to that of the United States.

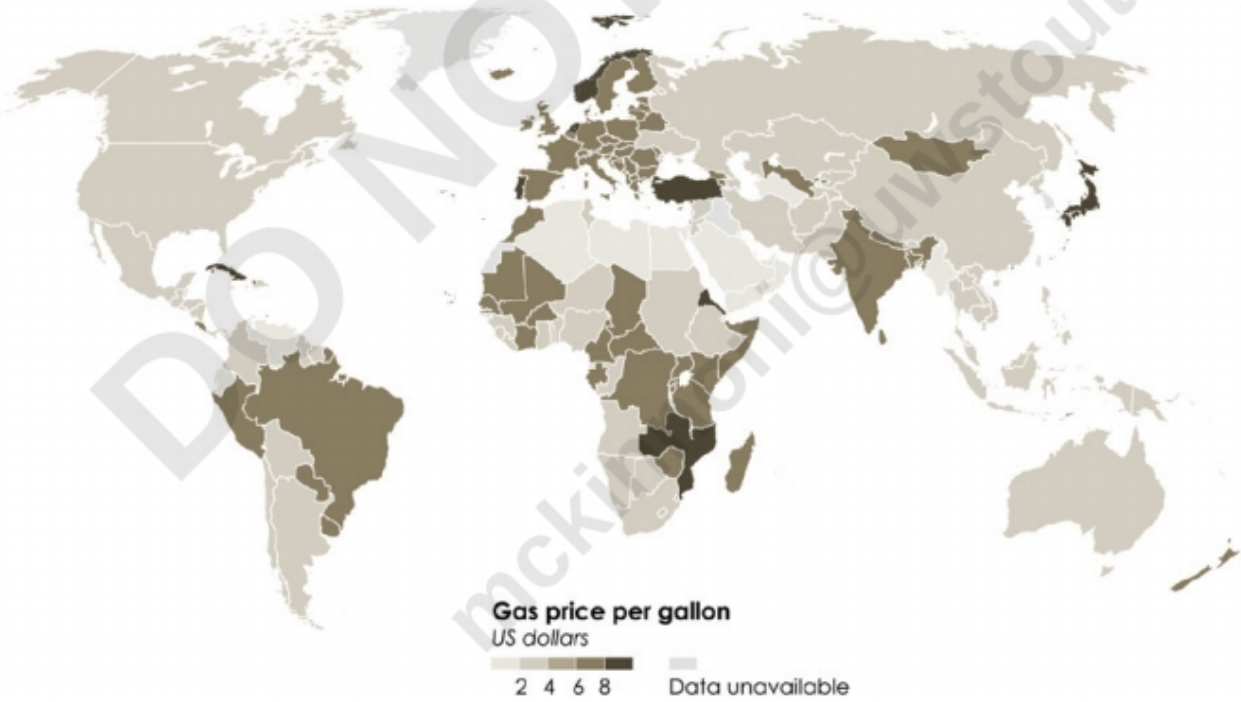


FIGURE 4-31 *Choropleth world map*

How deeply can you read into the data, though? Look at gas prices across your country, and there's variation. Heck, look at two gas stations within a few blocks of each other, and there can be a big difference. So although you can see general patterns, you shouldn't be too quick to judge as you explore. This data in particular comes from a variety of sources, such as government databases and newspaper articles, and from different years.

On the other hand, some sources use well-established methodologies and have done so for a long time. For example, the Bureau of Labor Statistics estimates the unemployment rate every month. You saw the national estimate over time in Figure 4-17, but you can also see the data by county, as shown in Figure 4-32. The map shows unemployment rate by county during August 2012. You can see high rates on the West Coast and in the Southeast and lower unemployment in the Midwest.

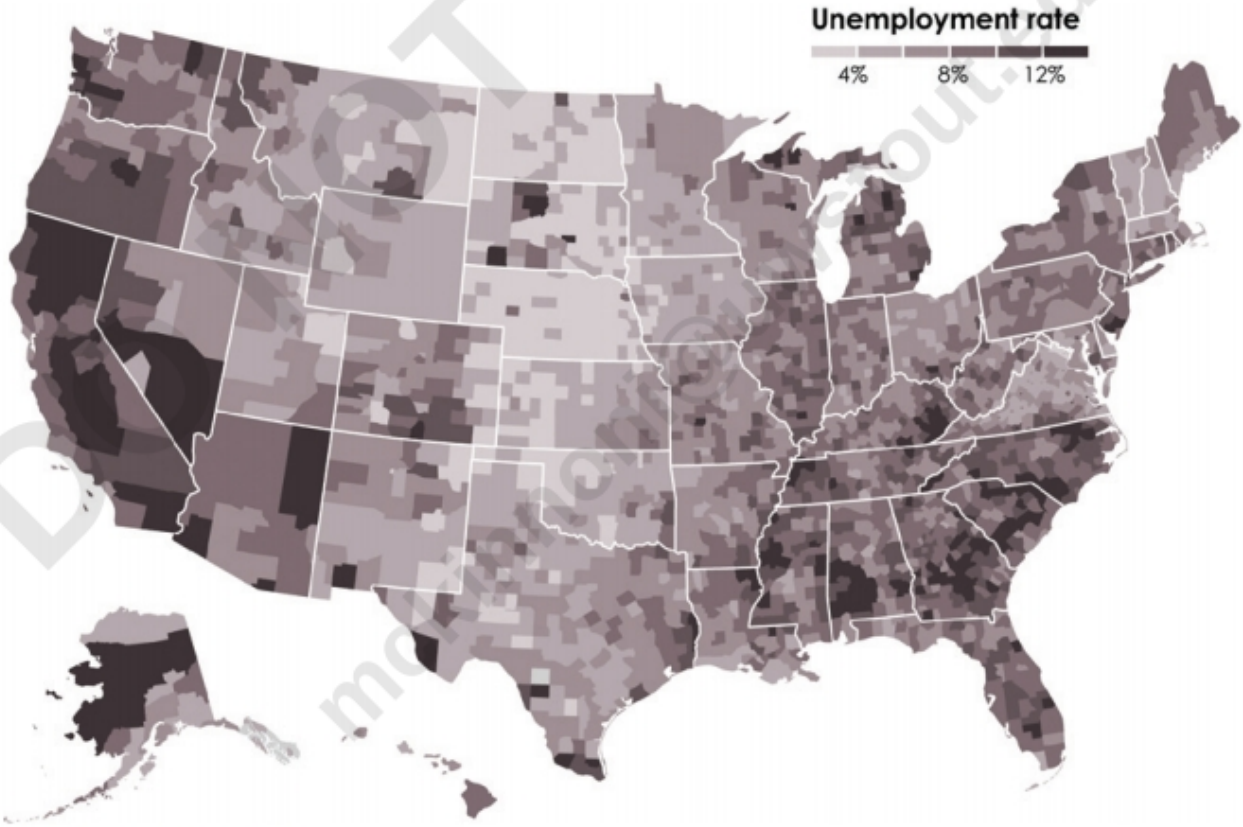


FIGURE 4-32 *County choropleth map*

There are also times when your spatial data actually does contain specific locations, but you're more interested in the aggregates. You might have a dataset with a lot of locations, and there are a lot of points in metropolitan areas. So when you map everything, points overlap, and it's difficult to tell how many observations there are in the dense areas.

For example, Figure 4-33 shows all recorded UFO sightings between 1906 to 2007, according to the National UFO Reporting Center. In areas where there were a lot of sightings (which curiously are where a lot of major airports are located), you just see a black blob, and it's hard to tell how many sightings there were when there is too much overlap.

Figure 4-34 shows the same data, but as a filled contour map. A color scale is used to show sightings density, where white means more sightings and black means none, and varying shades of red are for everything in between.

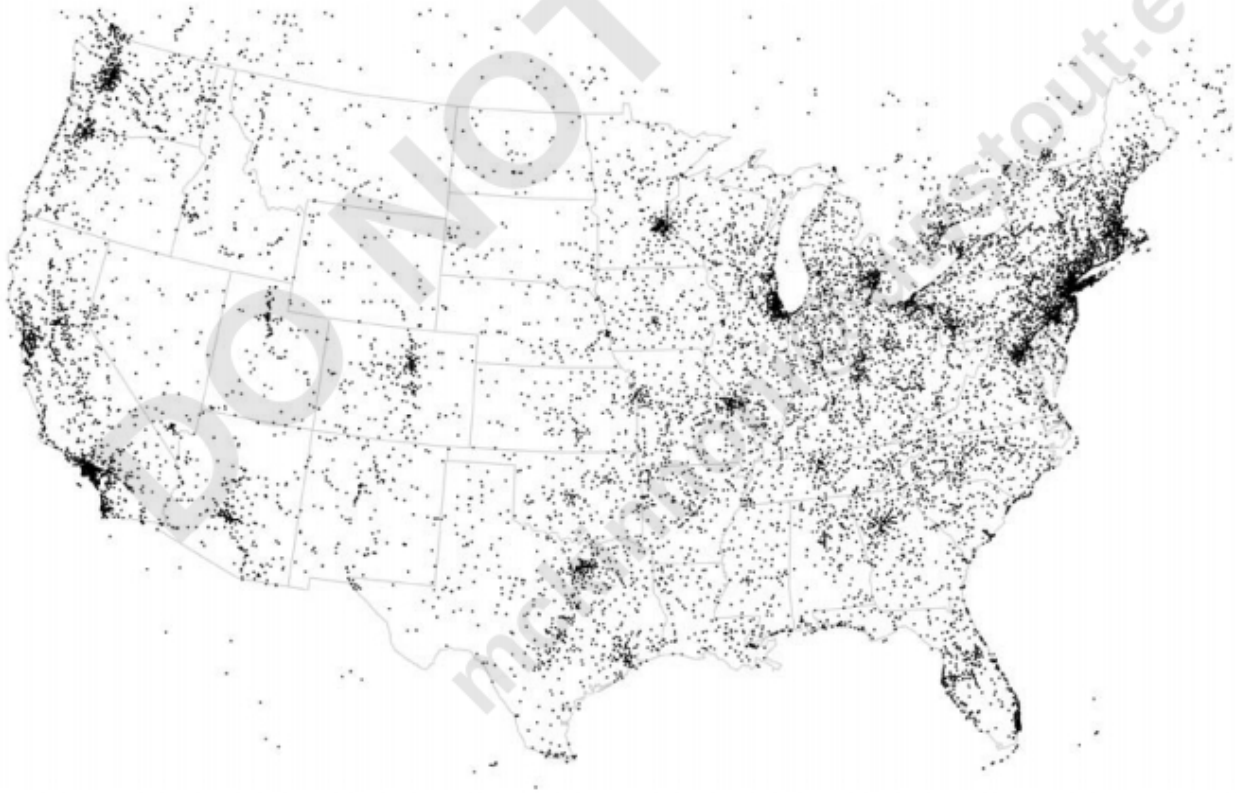


FIGURE 4-33 *Overlapping points on a map*

FIGURE 4-34 *Filled contour map*

CARTOGRAMS

A challenge with mapping regions, the choropleth map in particular, is that larger regions always get more visual attention regardless of the data. They take up more space in the physical world and on the computer screen. Cartograms are one way to remedy this. Location is somewhat preserved, but geographical areas and boundaries are not.

For example, a diffusion-based cartogram preserves boundaries but stretches them out so that the area of regions match the data. For example, Figure 4-35 shows the UFO sighting data as a cartogram. Notice the shrinking of Texas and swelling of California.

Obviously, the upside of cartograms is that areas fill the appropriate amount of space, but the trade-off is less geographic accuracy. When your data is for larger regions, with a wide range of sizes, this trade-off is worth it, but when regions are uniform in size, a choropleth map is most likely a better fit.

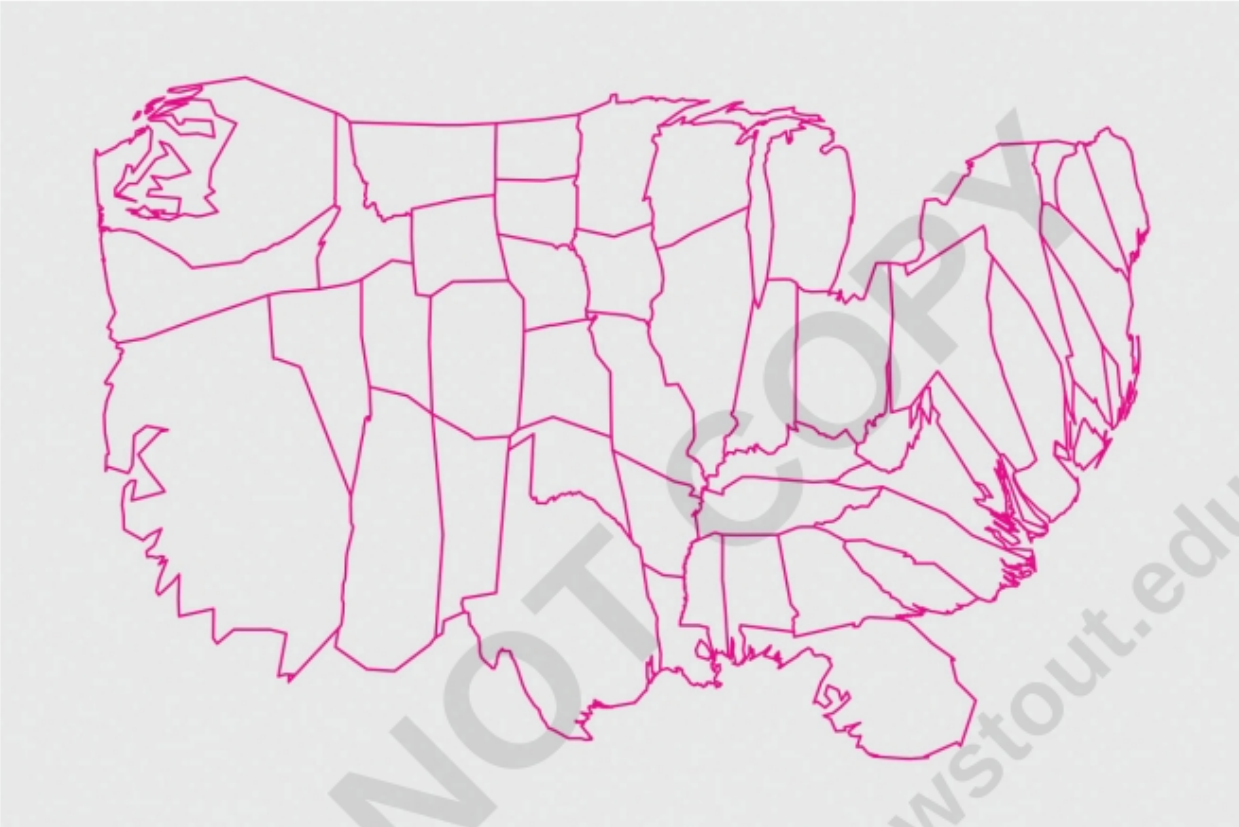


FIGURE 4-35 *Diffusion-based cartogram*

WHAT TO LOOK FOR

Spatial data is a lot like categorical data, but with a geographic component. You should know the range of the data to start with, and then look for regional patterns. Are there higher or lower values clustered in a certain area of a country or continent? Because a single value only tells you about a small part about a region filled with people, think about what a pattern implies and look to other datasets to verify hunches.

MULTIPLE VARIABLES

Data often comes in table form, with multiple columns, and each column represents a variable. You might have response data from a poll with multiple questions, results from an experiment that measured multiple aspects

of a system, or demographic data on countries that includes multiple bits of information on each.

Some visualization methods let you explore multivariate data in one view. That is, all your data might fit onto a screen, and you can interpret relationships between variables and explore trends in individual ones.

Often though, the relationships between variables aren't straightforward. There isn't always a clear increasing or decreasing trend. In these cases, multiple views using more straightforward charts and graphs can help a lot. As usual, your approach depends on the data you have.

A FEW VARIABLES

With time series data, you look for how a variable changes when another variable, time, does. Similarly, when you have two metrics about people, places, and things, you might want to know how one metric changes, given the other does. Do cities with higher burglary rates also have higher homicide rates? What is the relationship between housing prices and square footage? Do people who drink more soda per day tend to weigh more?

You can visualize relationships similarly to how you look for them with time series data. Whereas the dot plots in this chapter placed time on the horizontal axis and a variable on the vertical axis, a scatter plot replaces time with a different variable, so you have two variables plotted against each other, as shown in Figure 4-36.

Each dot represents a player during the 2008–2009 NBA basketball season. Usage percentage, an estimated percentage of possessions that a player is involved in while on the court, is plotted on the horizontal axis, and points per game is plotted on the vertical axis. As you might expect, those who spend more time with the ball tend to score more points per game.

This statistical relationship between variables is called *correlation*. As one variable increases, the other one usually does, too. In this example, the correlation is strong and obvious in the chart, but the correlation strength can vary, as shown in Figure 4-37.

For a more defined view of how two variables are related, you can fit a line through the points, as shown in Figure 4-38. You saw the same method used with time series data in Figure 4-21. The increasing curve rounds off as points per game approaches zero, but the line straightens out, showing a linear relationship. (It'd be a different story if the line resembled a sine wave.)

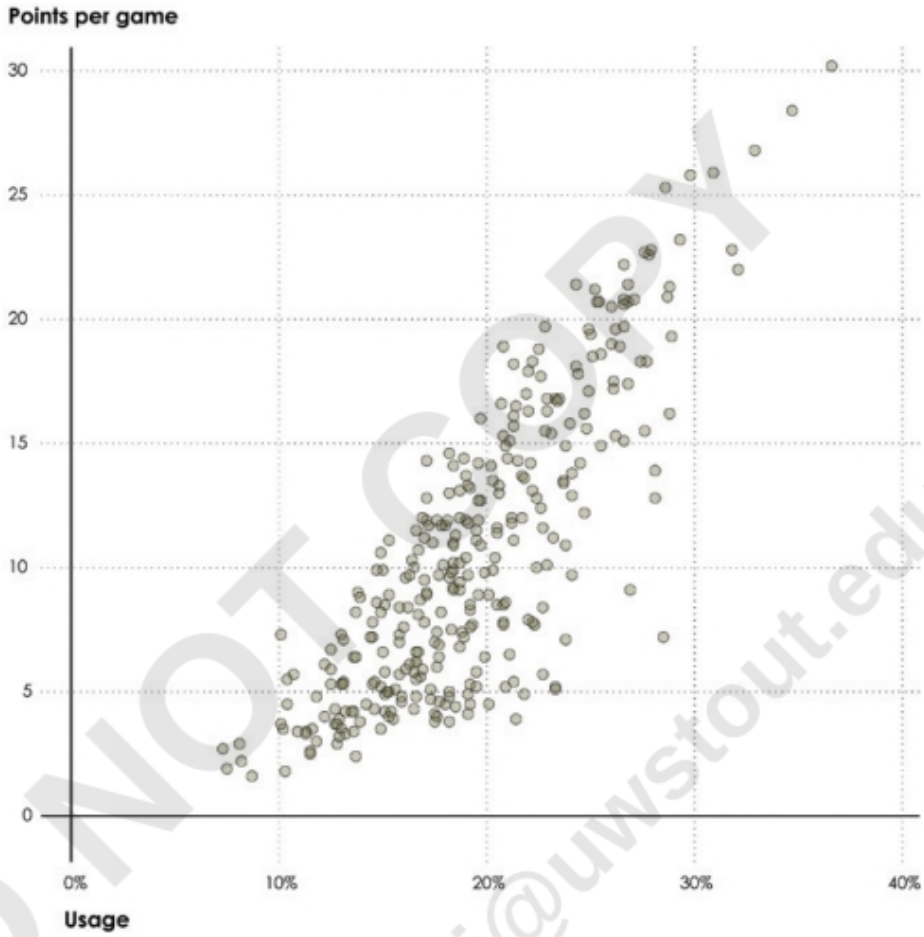


FIGURE 4-36 *Scatter plot to compare two variables*

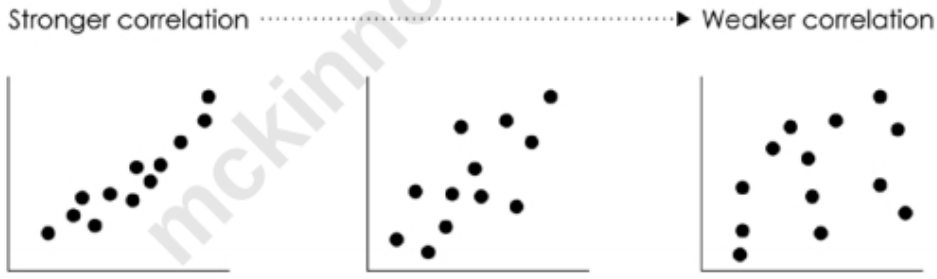


FIGURE 4-37 *Varying correlation strength*

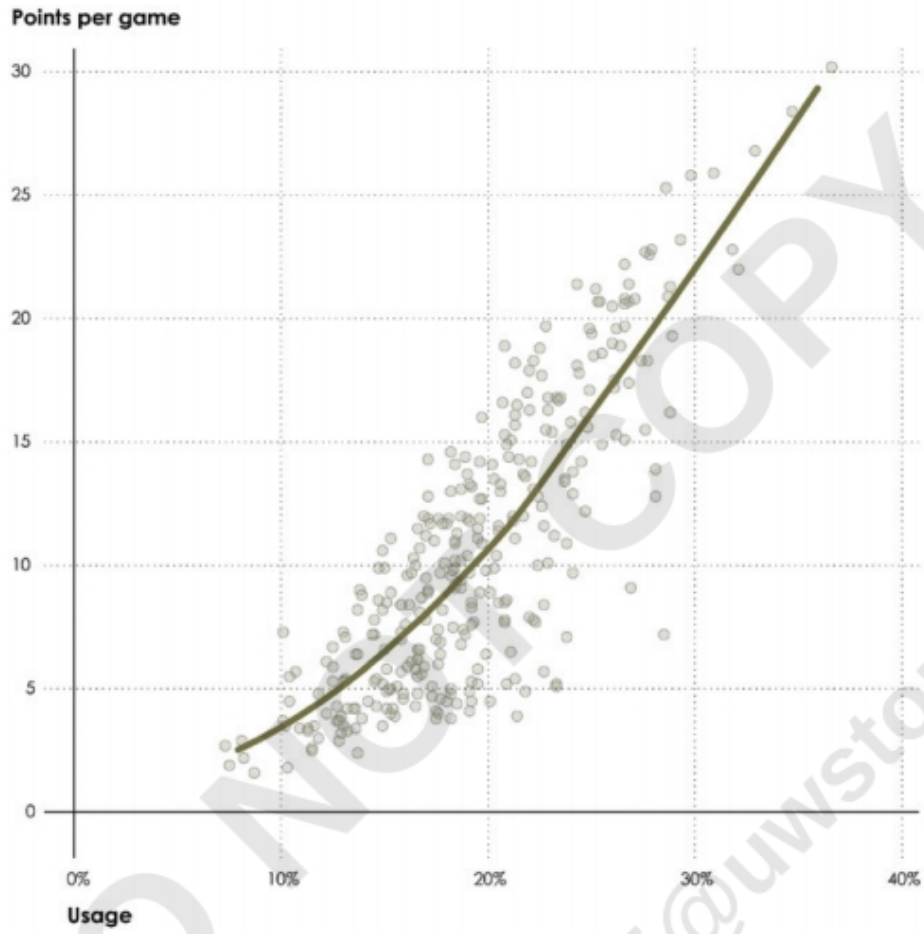


FIGURE 4-38 *Fitted LOESS curve on a scatter plot*

As you explore relationships between variables, don't confuse correlation with causation. Visualization-wise, a correlative and causal relationship between two variables will look similar, if not the same, but the latter usually requires rigorous statistical analysis and context from subject experts.

Obviously, some causal relationships are easy to interpret, such as when you place your hand over an open flame, you burn yourself. That's why you don't walk around in fire. On the other hand, the price of both milk and fuel has increased over the years. If you want to decrease the cost at the pump, should you just decrease the price of milk? Do basketball players score more points because

they handle the ball more often, or do they handle the ball more often because they are good at scoring points and the coach runs more plays for such players?

Warning: Don't mix up causal and correlative relationships. They look the same when you visualize them, but the former is more difficult to prove than the latter.

Figure 4-39 shows two ways to incorporate a third variable in a scatter plot. The symbol plot on the left should look familiar because it was used with spatial data on a geographic scale earlier in the chapter. The area of a circle represents assists per game. The scatter plot on the right uses color instead of area to show the same thing. The darker the shade, the more assists per game.

The hope is that you'll see larger circles or darker shades clustered in an area of the scatter plot. In Figure 4-40, you see assist leaders closer toward the right corner of higher usage percentage and points, but there's high variability and there isn't a clear trend. There are players with a lot of assists per game who don't score that many points, and there are others who score a lot of points, have high usage percentage, and a lot of assists. It is clear however that those who don't score many points and have lower usage percentage typically don't have many assists either.

You can also double up on encodings, using both size and color to represent a third variable, as shown in Figure 4-40. The redundant visual cues help reinforce what might be more of a challenge to see with just one visual cue.

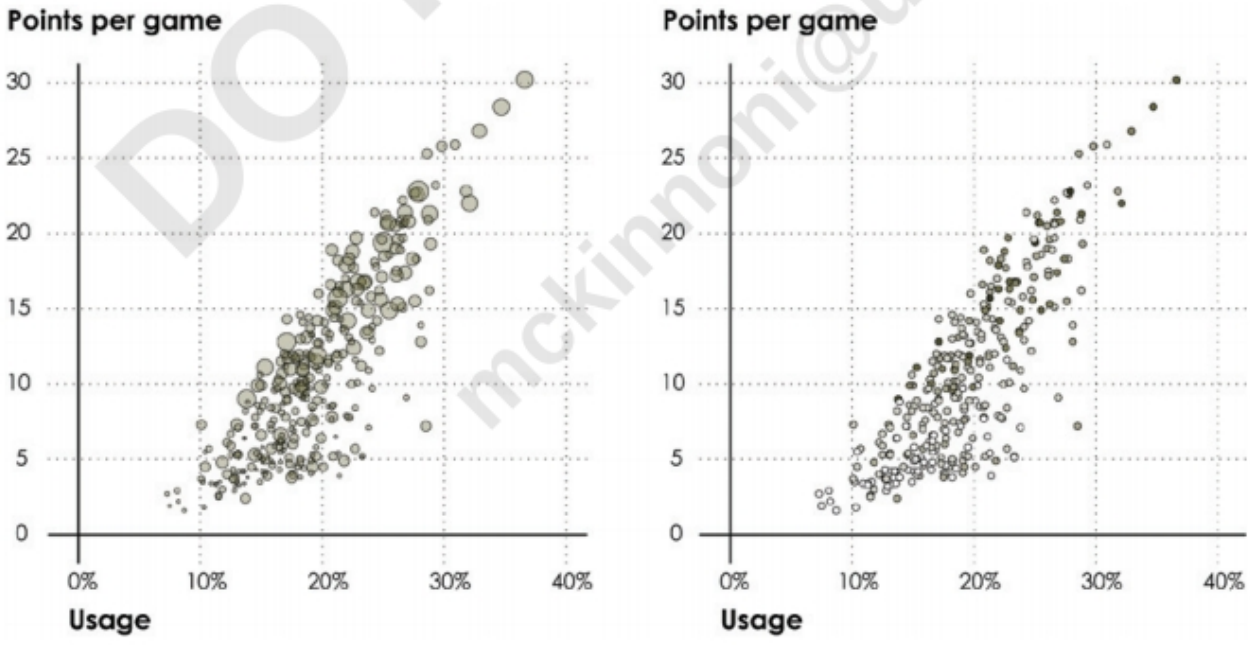


FIGURE 4-39 *Symbol plot on the left and colored scatter plot on the right*

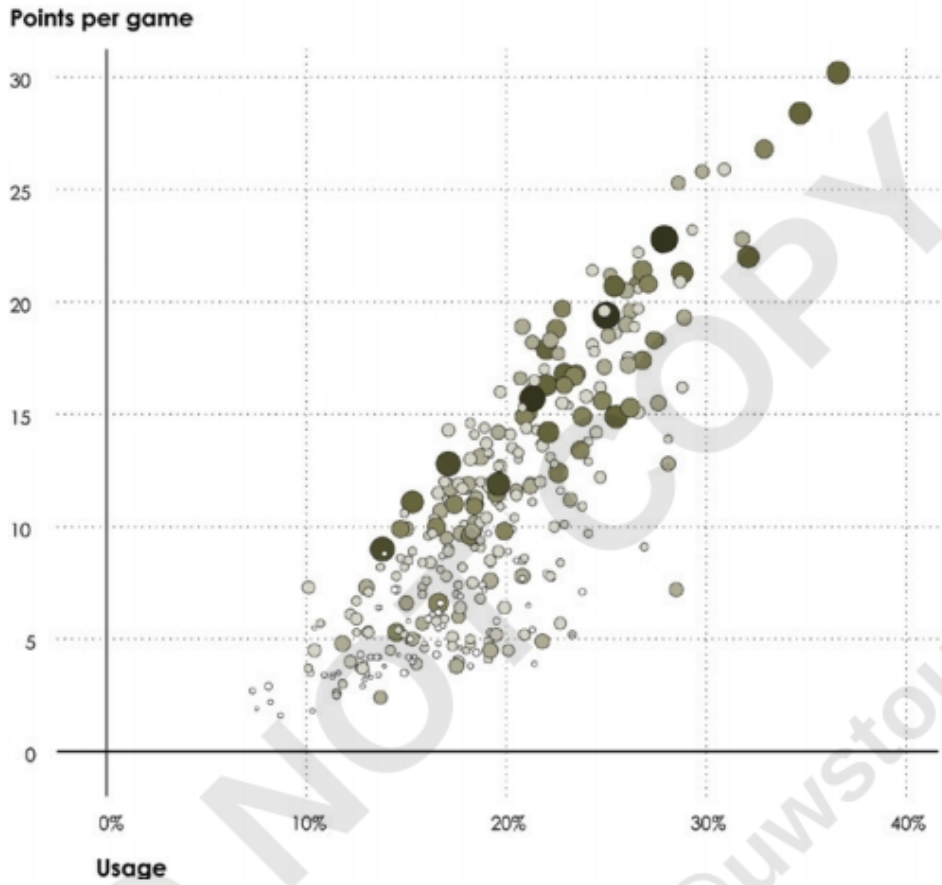


FIGURE 4-40 *Using redundant visual cues*

For example, if you were to go the opposite direction and use area and color to represent two separate metrics, the plot could be difficult to read. Figure 4-41 shows the same values on the axes, usage percentage and points per game, but uses area for rebounds and color for assists. Compare this to the previous figure, and it's clear that the additional encodings don't make anything clearer.

MANY VARIABLES

You might show four variables with a scatter plot, but what about five variables? Ten variables? There's only so much space in a scatter plot for so many visual cues. Unlike the scatter plot, there are views that are more conducive to comparing multiple variables at one time.

Note: You might be looking for a rule about how many encodings you can use at the same time before a visualization becomes useless, but I'd be overgeneralizing. It depends on the data. And the visualization. Experiment.

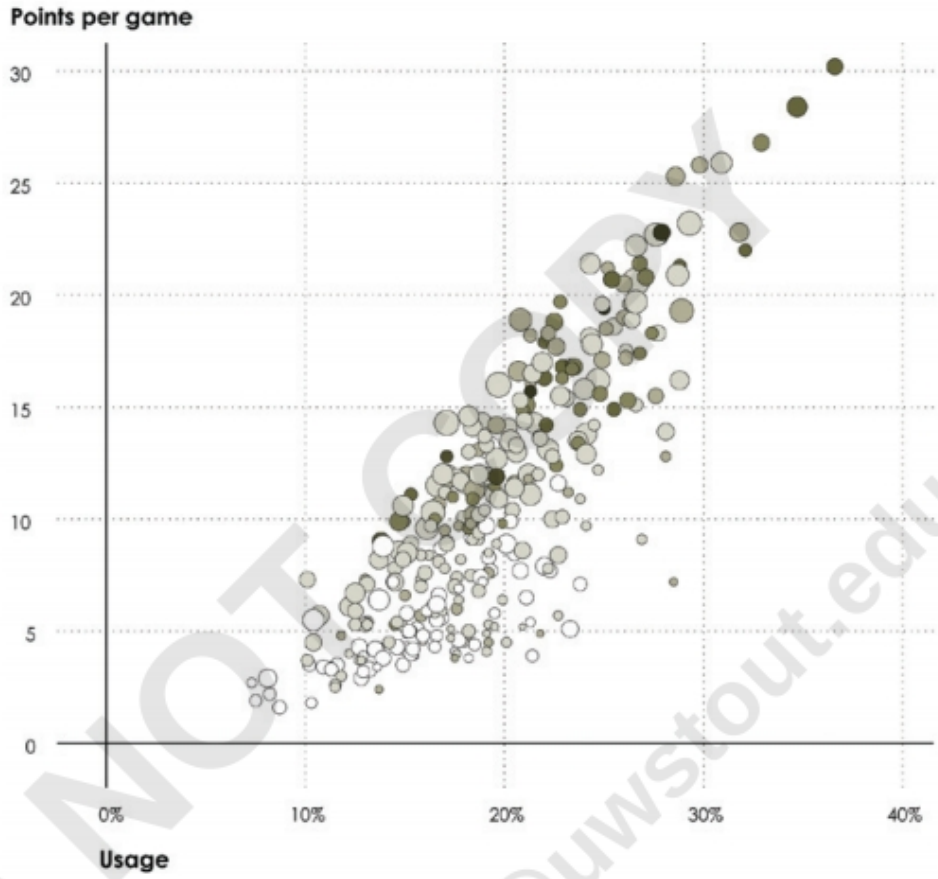


FIGURE 4-41 *Multiple encodings with a scatter plot*

A heat map, as shown in Figure 4-42, can be used to translate a table to a set of colors. It shows the same basketball player data, in addition to several other variables, including number of games played, field goal percentage, and three-point percentage. Each row represents a player, and darker shades represent relatively higher values.

With players sorted alphabetically, it's hard to see patterns, but if you sort by a column, say, points per game, as shown in Figure 4-44, relationships are easier to see. For example, usage percentage and minutes are roughly dark to light. On the other hand, the turnover rate appears to indicate a negative correlation because it goes from light to dark, and games played, field goal percentage, and three-point percentage look scattered, indicating a weak correlation, if any.

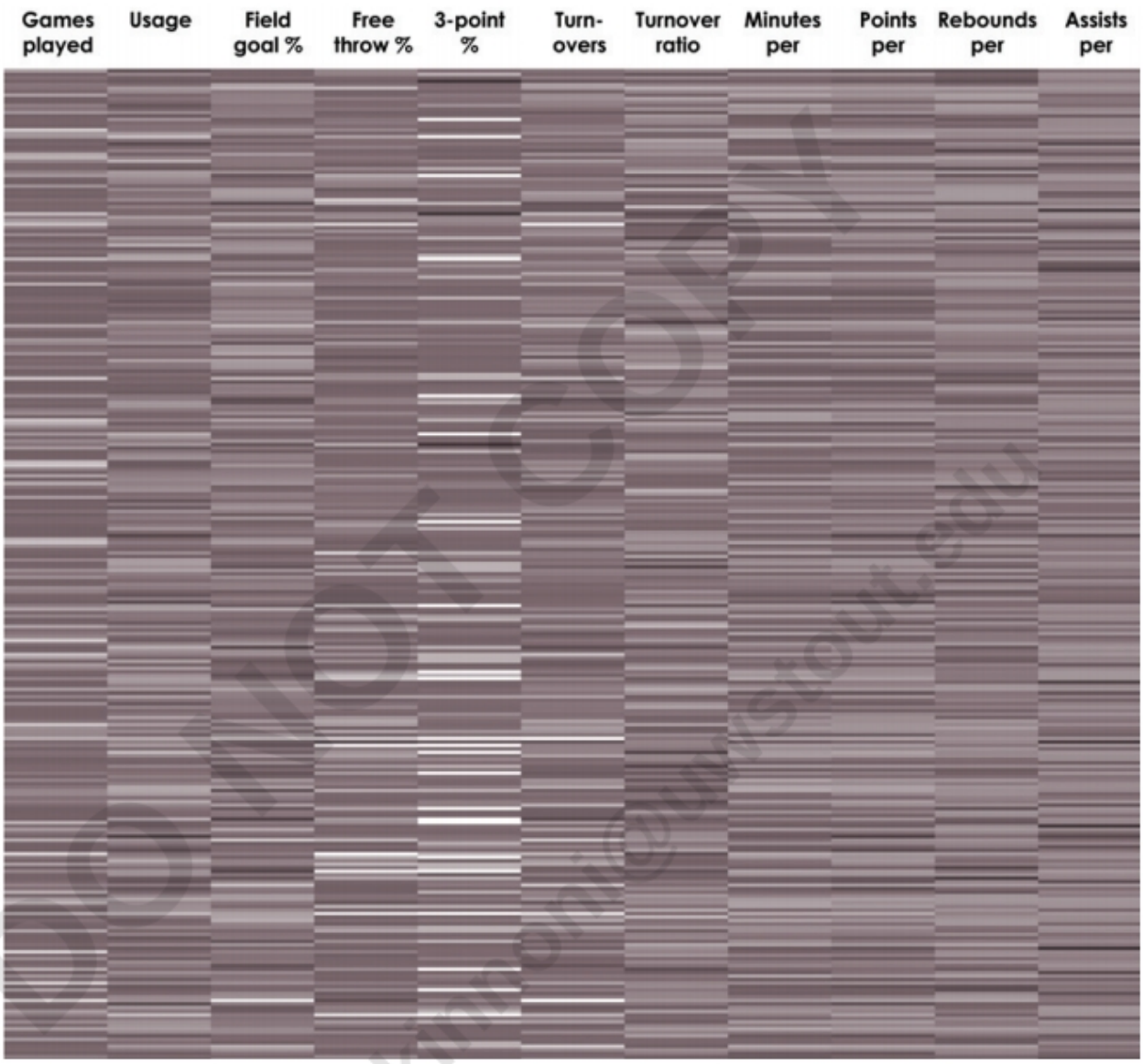


FIGURE 4-42 *Heat map to show multiple variables*

Parallel coordinates plots also arrange variables horizontally, but instead of using color like a heat map, you use vertical position, as shown in Figure 4-44. Each vertical axis represents a variable that typically ranges from the minimum and maximum of that variable, so the highest value is plotted at the top and the lowest at the bottom. Then lines are drawn left to right, positioned by the variables of each observation.

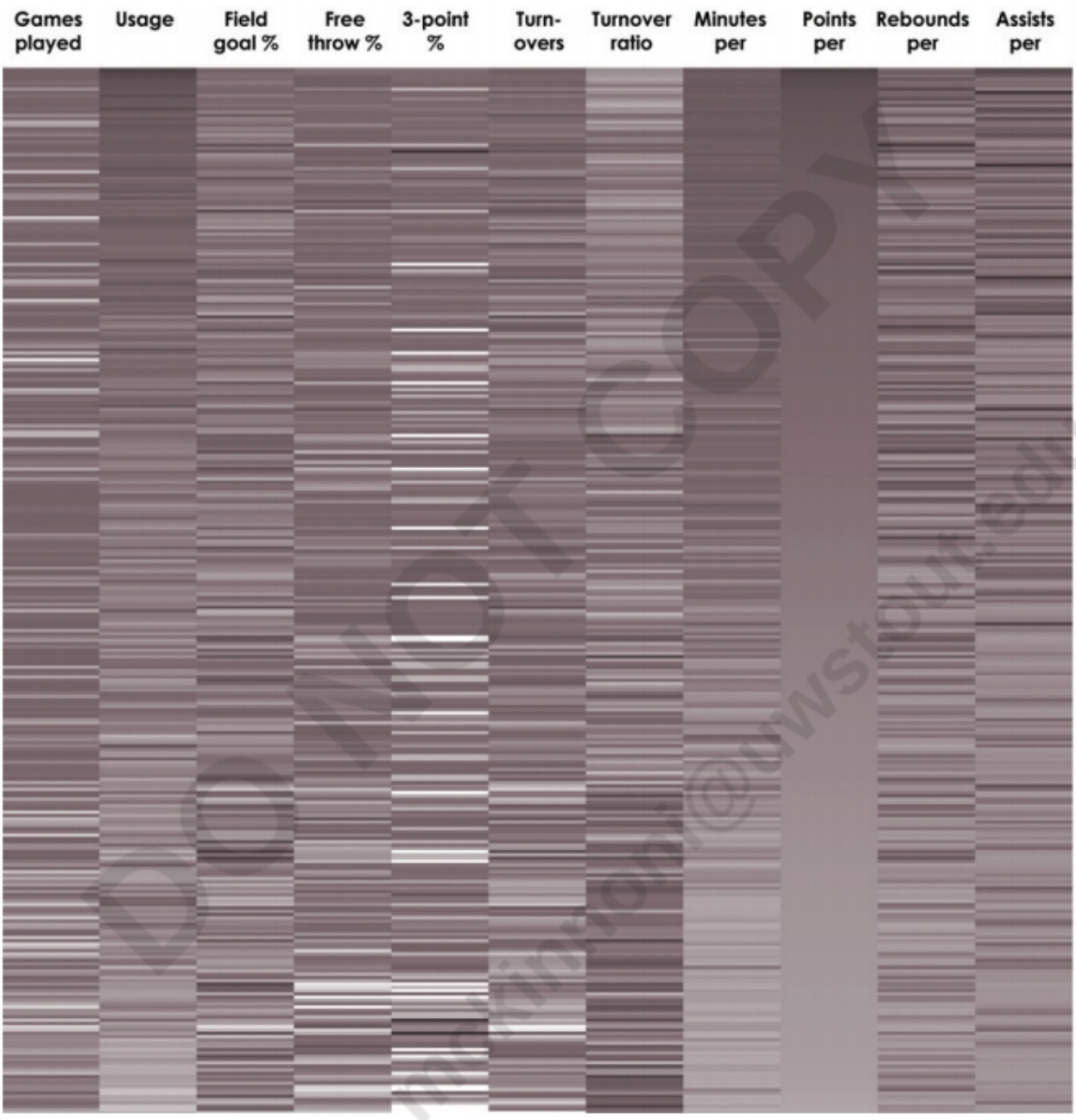


FIGURE 4-43 *Relationships with heat map*

For example, to plot a player, you start on the left, look up how many games he played, and start a line in the corresponding spot on the first vertical axis. Draw the line to the spot on the next axis that corresponds to the player's usage percentage. Do that for all the variables and all the players, and that's the parallel coordinates plot.

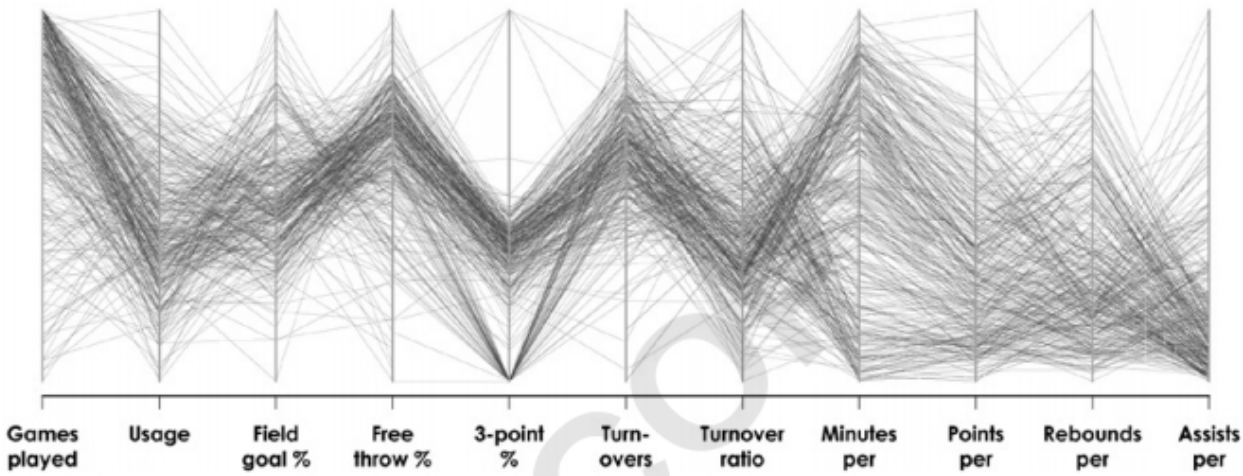


FIGURE 4-44 *Parallel coordinates plot*

If all the variables had strong positive correlations (which almost never happens), all lines would run straight across. If two variables were negatively correlated, you'd see lines on the top of one variable connect to the bottom of the axis for the other variable. Figure 4-45 shows a few more relationships.

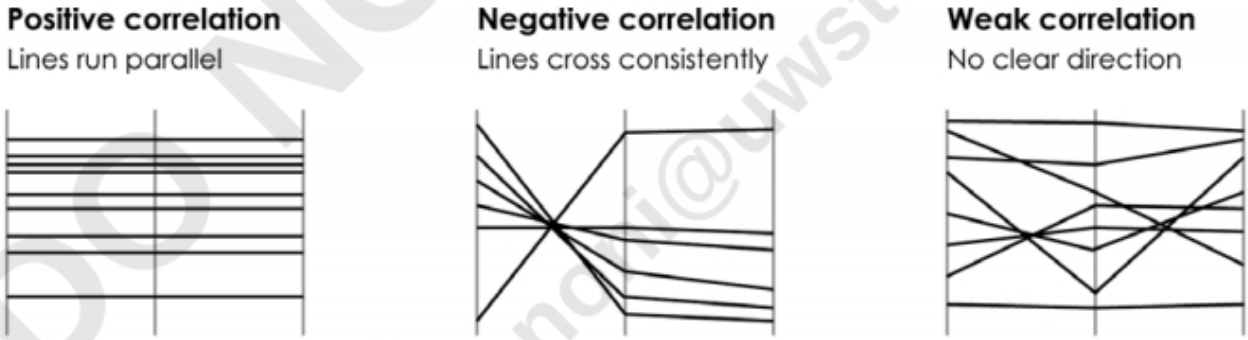


FIGURE 4-45 *Relationships with parallel coordinates plot*

When there aren't clear relationships across the board, it can be hard to see patterns. There's high variability from player to player in Figure 4-45, so you end up with a jumble of lines. You can however, highlight data based on criteria for a better view.

For example, if you highlight players who averaged five assists or more and gray out everyone else, as shown in Figure 4-46, it's easier to see how these type of players perform in other categories. Assist leaders play in more games, play more minutes, and tend to rebound less, but still vary in terms of points and field goal percentage.

Whereas the heat map and parallel coordinates plot provide an overview of the data, you might also want to look at individual data points more closely. Star plots, as you saw with time series data and shown again in Figure 4-47, present data separately. That is, you represent each row of data with its own plot.

The time series example uses the angle portion of the polar coordinate system for time. This example uses multiple variables. So in the same way that the star plot can be a polar coordinate version of a time series chart, it can also be a polar coordinate version of a parallel coordinates plot.

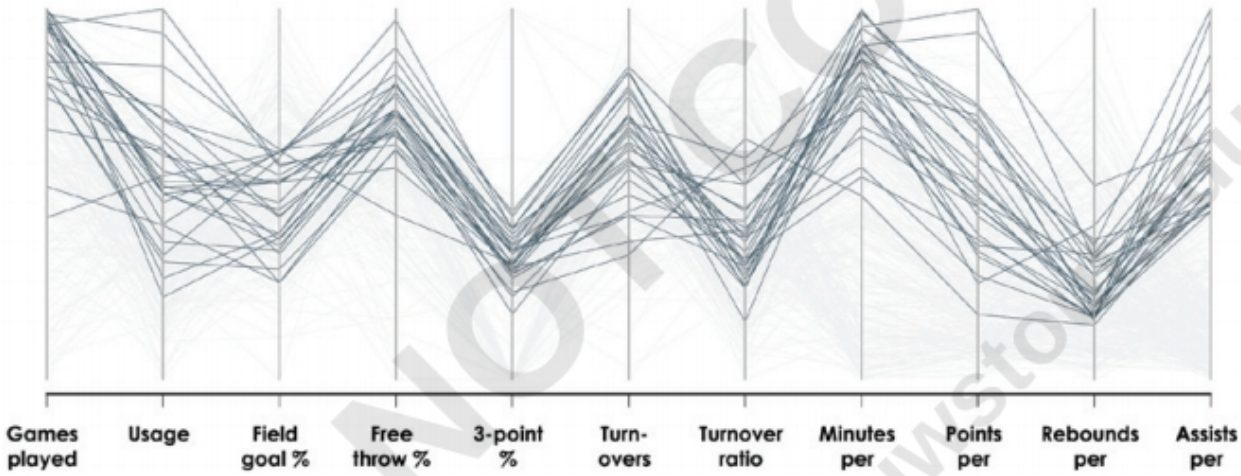


FIGURE 4-46 *Highlight data for easier reading*

USING MULTIPLE VIEWS

Note: As you gain more experience visualizing data, you'll notice that you can use many of the same type of charts for different types of data.

There's an inclination to show all the data at once, in a single view. When you have categorical data, you make a bar chart, and when you have time series data, you use a line chart. So if you have multiple variables that might be categorical, temporal, and spatial, you might also want to put it all in a single chart.

However, it can and often is better to use multiple charts instead because it lets you see the data from more angles.

You can, for example, make a lot of the same type of chart on multiple dimensions, such as the maps in Figure 4-30. The flight data is actually spatial, categorical, and temporal, so you get natural breaks in the data and a hint of places to look. Figure 4-48 explores the time series component of the flight data. Each line represents flight volume for an airline.

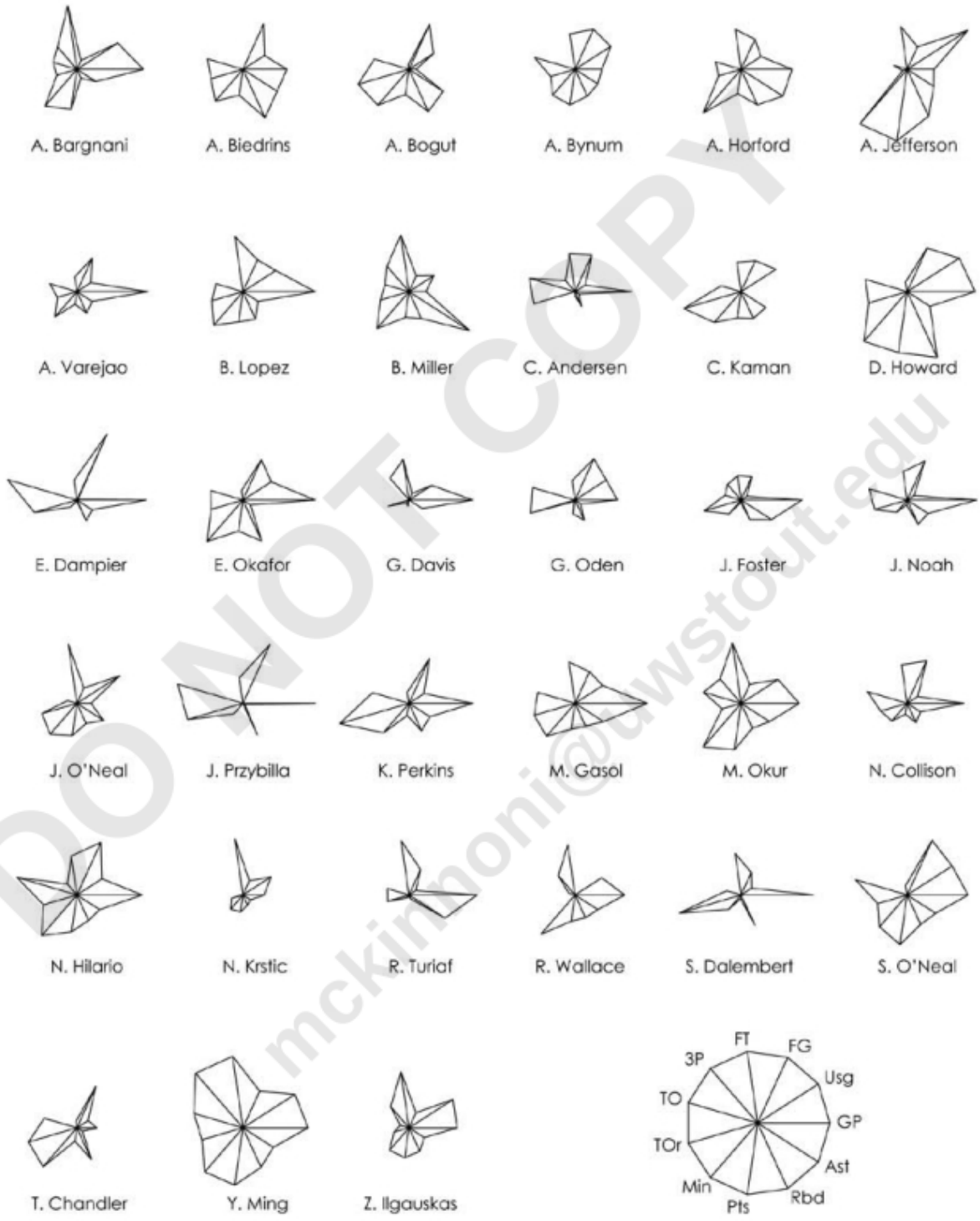


FIGURE 4-47 *Star plots*

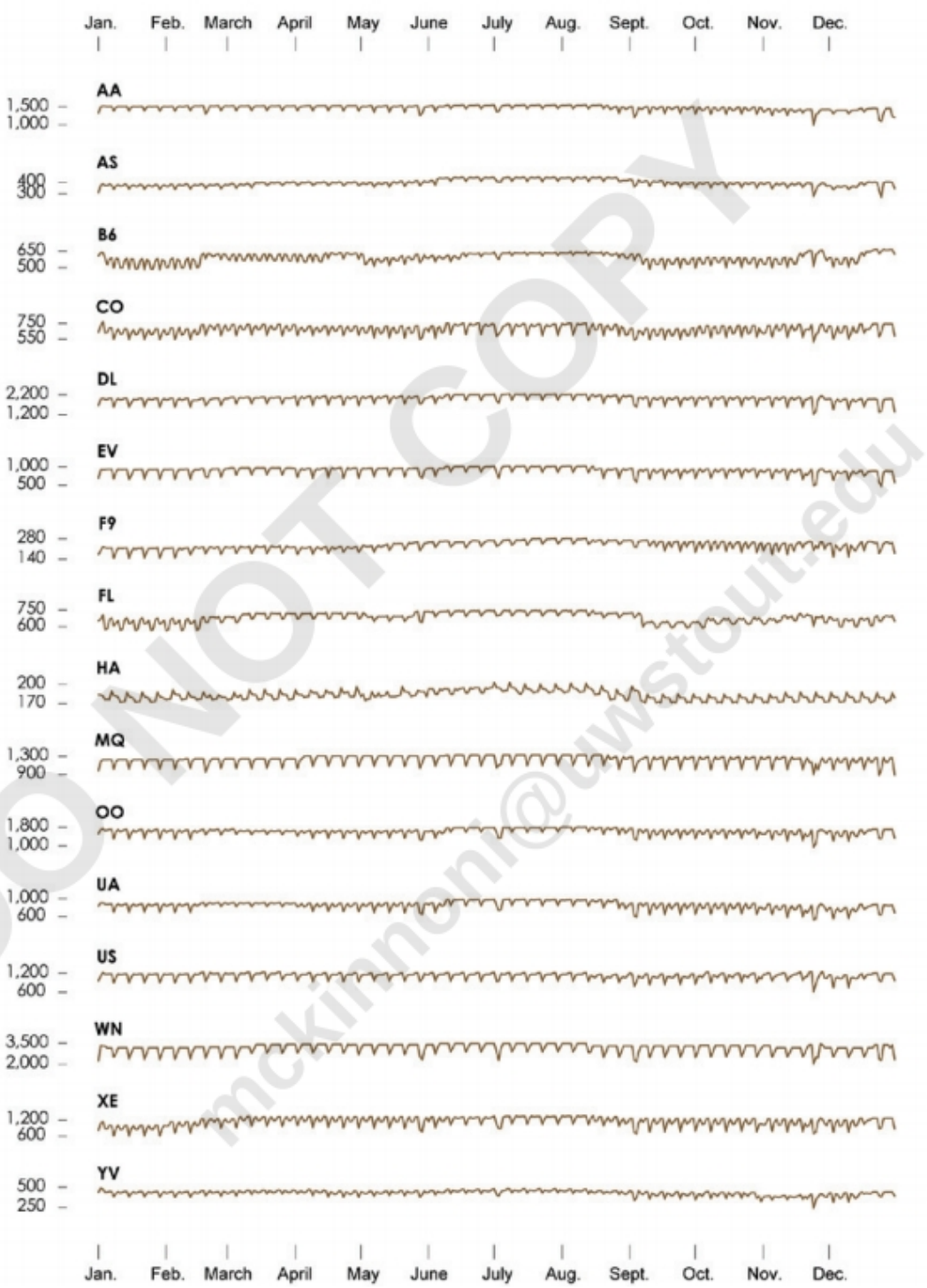


FIGURE 4-48 *Multiple time series charts for categories*

An alternative to the parallel coordinates plot, a scatter plot matrix, as shown in Figure 4-49, can show similar relationships. The relationships between variables are often easier to see in the matrix than with parallel coordinates because you can compare pairwise correlations instead of trying to decipher relationships between multiple variables at once. The latter is often complicated and hard to see.

It's also often useful to look at data with different views at the same time. For example, Figure 4-50 shows data as a heat map, bar chart, and star plot, for several players. The heat map provides detailed information about where the players shoot from; the bar graph provides an overview of the aggregates; and the star plot shows values for additional variables. Together, the views represent the playing style of several individuals, or more generally speaking, a detailed overview of several categories.

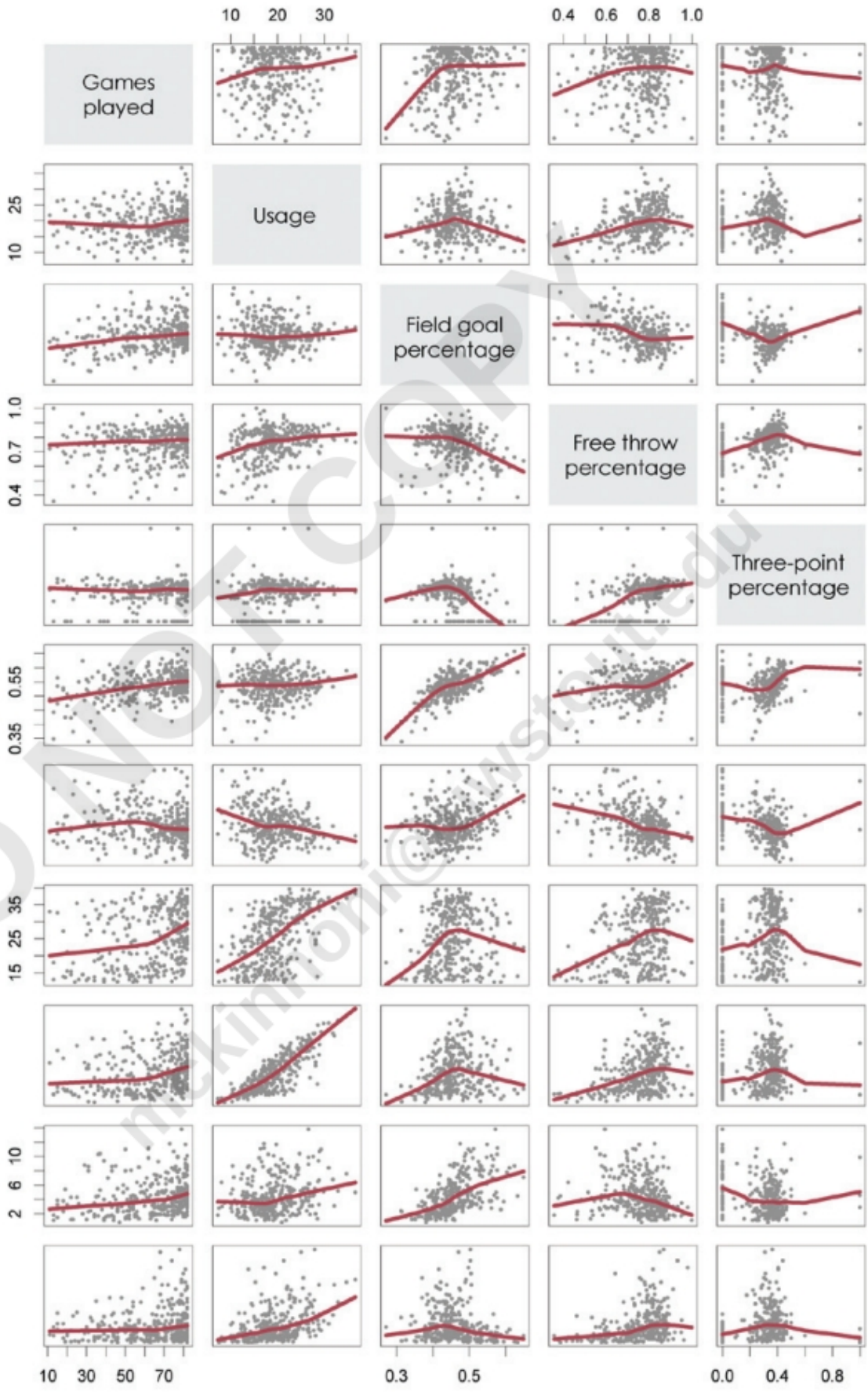
WHAT TO LOOK FOR

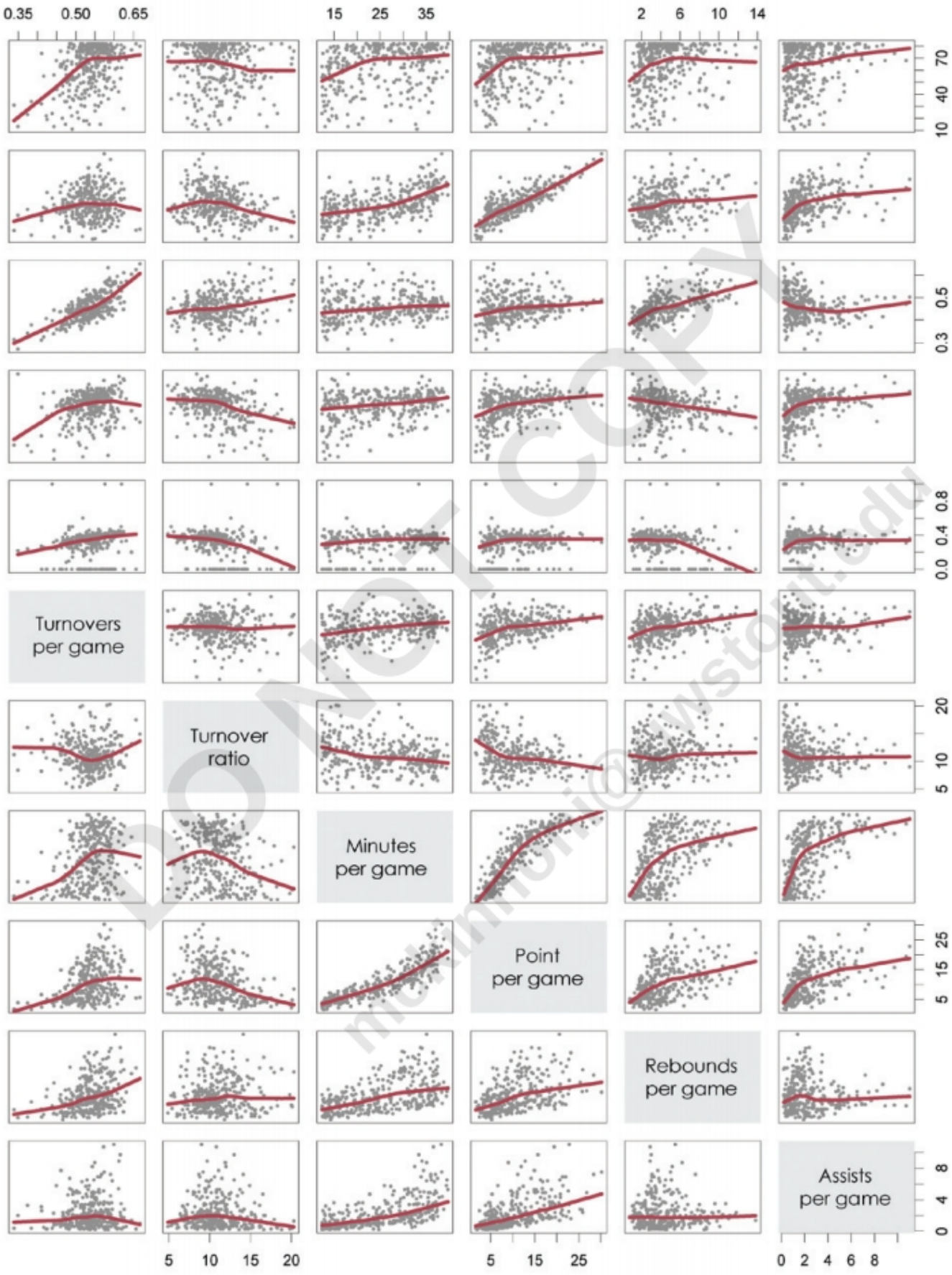
There are a lot of visualization methods that help you explore various aspects of your data, whether it is categories, time, space, or a combination of these. You can visualize the data all at once, but you can also make use of simpler, more straightforward views, which can help extract relationships. Sometimes the relationships are straightforward between two variables, but usually the relationship is complex, especially when you introduce more than two variables. Don't make assumptions as you explore relationships, and keep in mind there are variables not captured in the data that might contribute to changes. Finally, when it comes to correlation and causation, you need to take in all the context you can before you assign the latter.

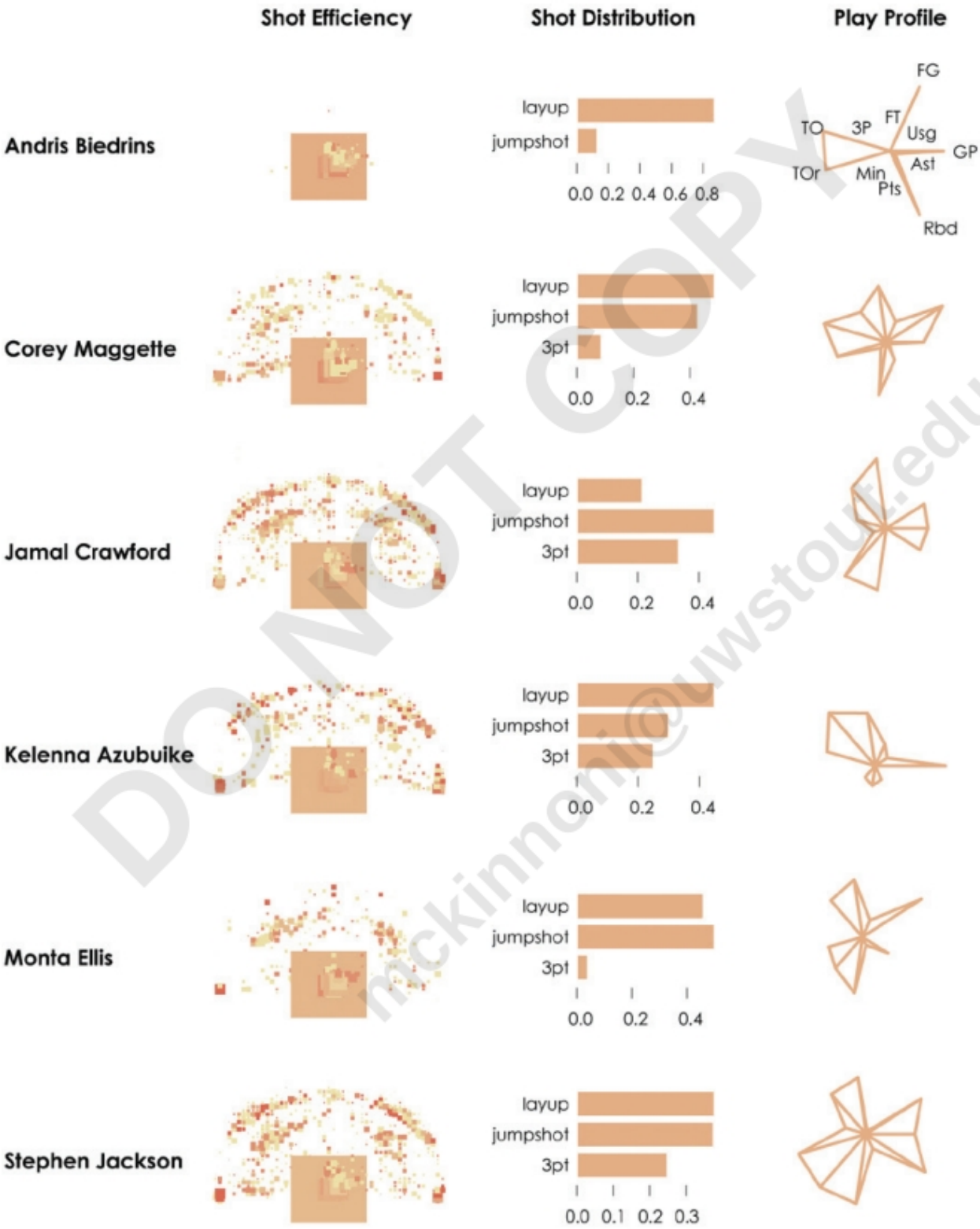
DISTRIBUTIONS

You often hear or read about means and medians. They're used to describe a group of people, places, or things, and these measurements typically imply what is "normal" or "average," and anything that is far away from these measurements is abnormal or above or below average. However, what qualifies as extremely above average or just slightly below average? Is something 10 percent greater than the median a lot or a little? To answer these questions, you must know more about the data than just where the middle is. You have to know the spread.

FIGURE 4-49 *(following page) Scatter plot matrix*







Now look at a classic example. Imagine there are 100 adults in a room. These 100 people have different heights, as shown in Figure 4-51. They range from 4 feet and 10-inches tall to 6½-feet tall, and the average height for the group is 5 feet and 4 inches.

FIGURE 4-50 *(facing page) Using multiple visualization methods to explore different dimensions*

It's hard to determine how many people there are in various height ranges without counting each dot, but you can get a better idea if you sort everyone from shortest to tallest, as shown in Figure 4-52. There are a few relatively tall people and a few short people, but most heights are around the 5 to 6-foot range. The median line at 64 inches is in the middle, where 50 people are shorter and 50 people are taller.

You get a better sense of the heights in the room, but there's a better way to see the distribution. You can group them into height categories or bins, such as those in between 4 feet and 4½- feet, as shown in Figure 4-54.

Now it is easy to see where most people are centered and to see the spread across a range. However, the dot plot can take a lot of space, especially if you had a lot more heights to show. So instead of dots, you could use bars, as shown in Figure 4-54. This chart is called a *histogram*, which you'll see more of soon. This counting and binning process is the basis for visualizations used to explore distributions.

As shown in Figure 4-55, you can visualize distributions with varying levels of granularity. Some views show only summary statistics, such as median, whereas other views, such as the histogram, show distribution in greater detail.

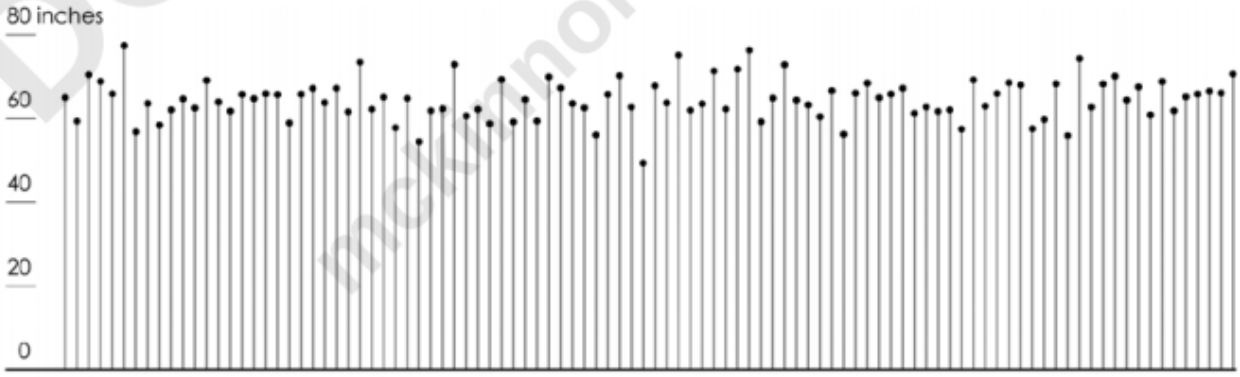


FIGURE 4-51 *Heights of 100 imaginary people*

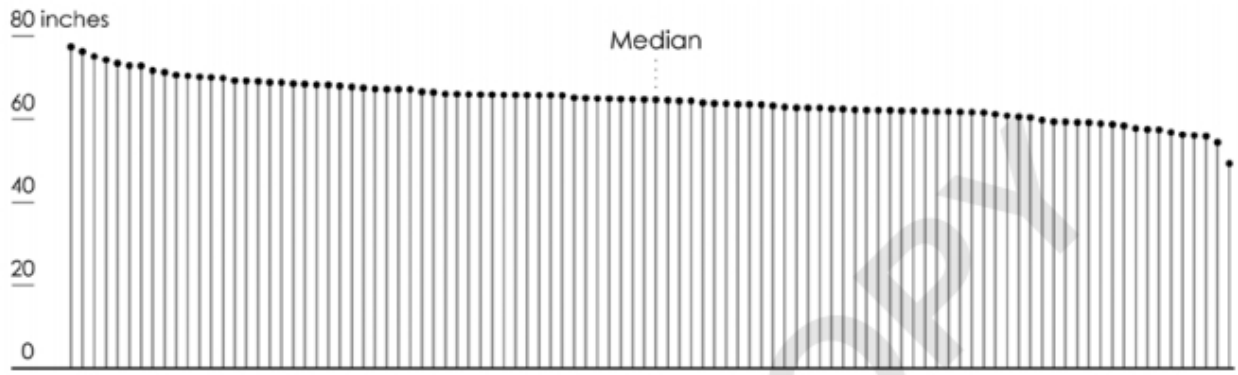


FIGURE 4-52 *Heights of imaginary people, sorted from shortest to tallest*

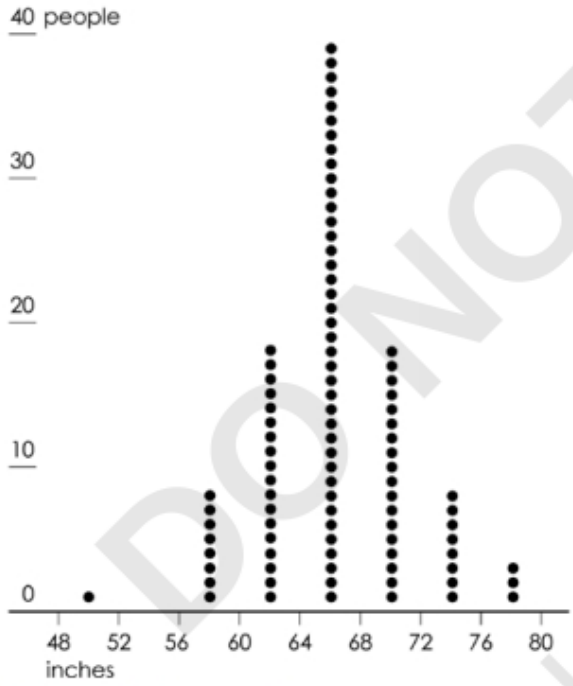


FIGURE 4-53 *Heights arranged in bins*

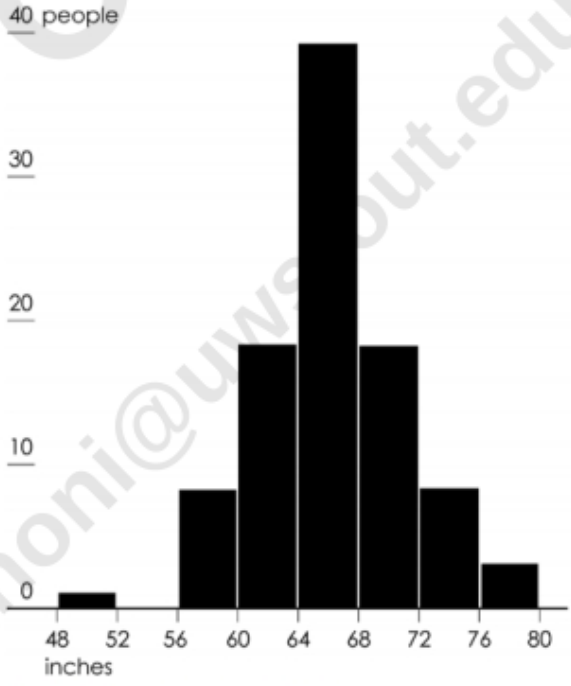


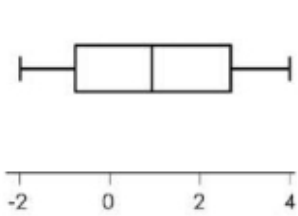
FIGURE 4-54 *Histogram showing heights*

The box plot, as shown in Figure 4-56, is an overview visualization that provides a general sense of distribution. The box in the middle is defined by the lower and upper quartiles. That is, whereas the median (the line in the middle) represents the halfway point, the lower quartile represents where one-quarter of the values are lower, and the upper quartile represents where one-quarter of the values are higher.

Distribution Summary

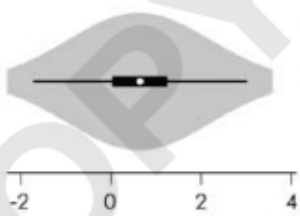
You can visualize data at different granularities with the charts above. These show key values for a less specific view of distributions.

Box plot



Shows range, median and quartiles

Violin plot

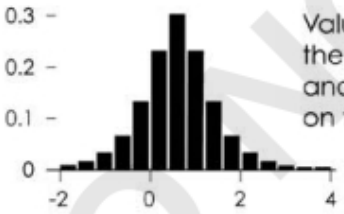


A combination of a box plot and density plot

Distribution of one variable

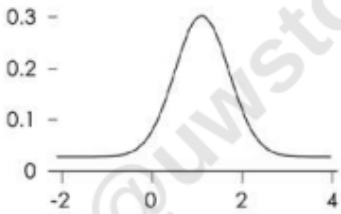
You can see where data is clustered and see any outliers by keeping track of where they sit on a value axis.

Histogram



Value axis on the horizontal and density on the vertical

Density plot

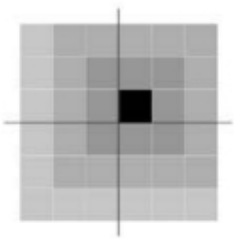


Like histogram but continuous instead of bins

Distribution of multiple variables

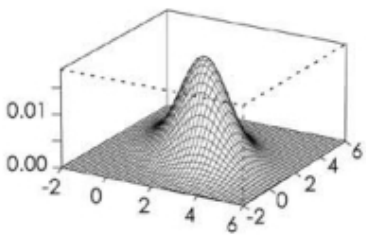
Sometimes values come as pairs, and it makes sense to show both values at the same time.

Heat map



Density on a a 2-D plane, using color as visual cue

Surface plot



Shows same patterns as heat map, but uses height instead of color

FIGURE 4-55 *Visualizing distributions*

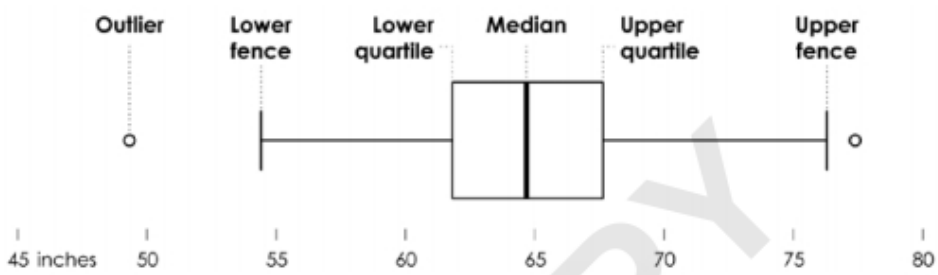


FIGURE 4-56 *Box plot*

The range in between the upper and lower quartiles is called the *interquartile range*. The outer lines are the lower and upper fences, defined by subtracting and adding 1½ times the interquartile range from the lower and upper quartiles, respectively. If the maximum and minimum values are within the upper and lower fences, the outlines are only drawn to the extremes. Otherwise, dots are used to represent any points that fall outside the upper and lower fences and are considered outliers.

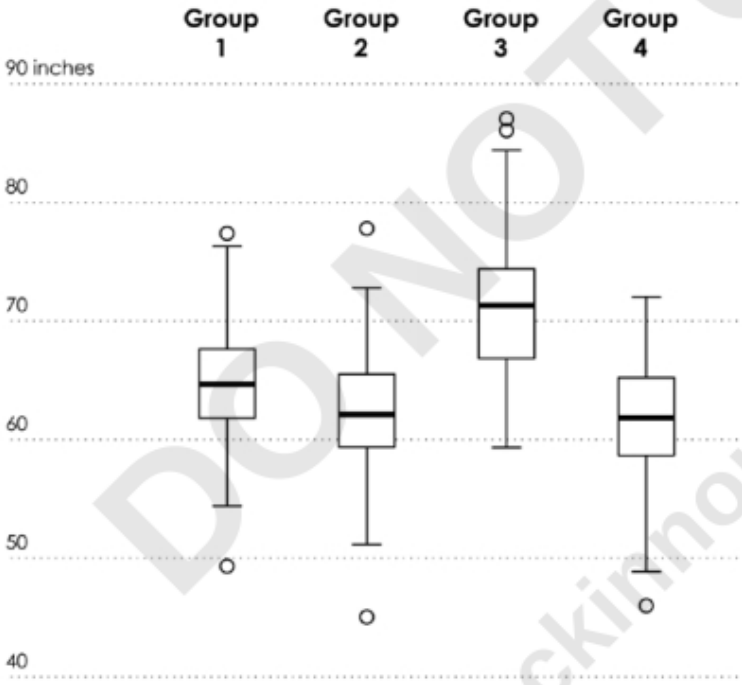


FIGURE 4-57 *Multiple box plots for comparison*

That said, the terminology makes the chart more confusing than it actually is. The main point: You can see a general distribution with a box plot. You can also use multiple box plots to compare distributions, as shown in Figure 4-57.

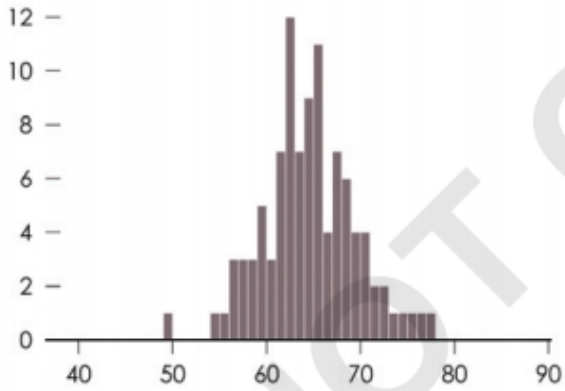
The histogram provides a more detailed view, which you saw in Figure 4-54. Bar height represents the proportion of values within a corresponding range, and when you change bin sizes, you change how much variability is visible. Figure 4-58 shows how the same height data can be represented with different bins.

Like box plots, you can also use multiple histograms to compare distributions. In one last return to the flight data, Figure 4-59 shows the distributions of arrival delays for major airlines. Delays of more than 15 minutes are highlighted in orange.

Note: Bin size changes by dataset, but you want it to be big enough so that you can see the variability over the range of values, but not so small that the histogram is too noisy to interpret.

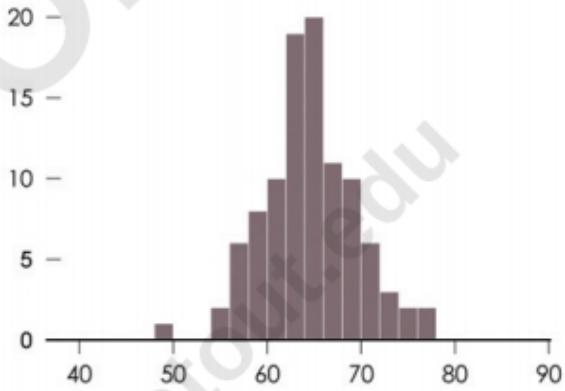
One-inch bins

Small bins shows variations at higher granularity.



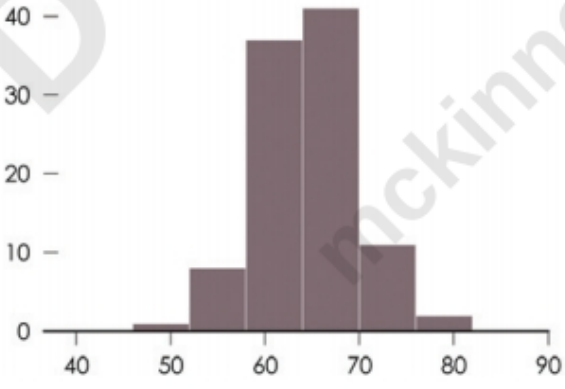
Two-inch bins

You see less variation, but the distribution around the median is more obvious.



Half-foot bins

You can see distribution around the median, but you can only see some variation.



One-foot bins

The spread of the data isn't as obvious. because the larger bins show less detail.

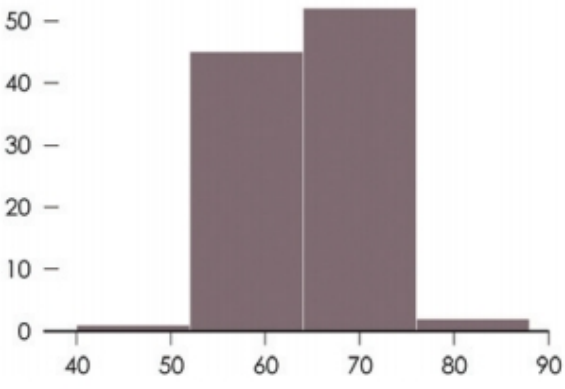


FIGURE 4-58 *Varying bin sizes with histograms*

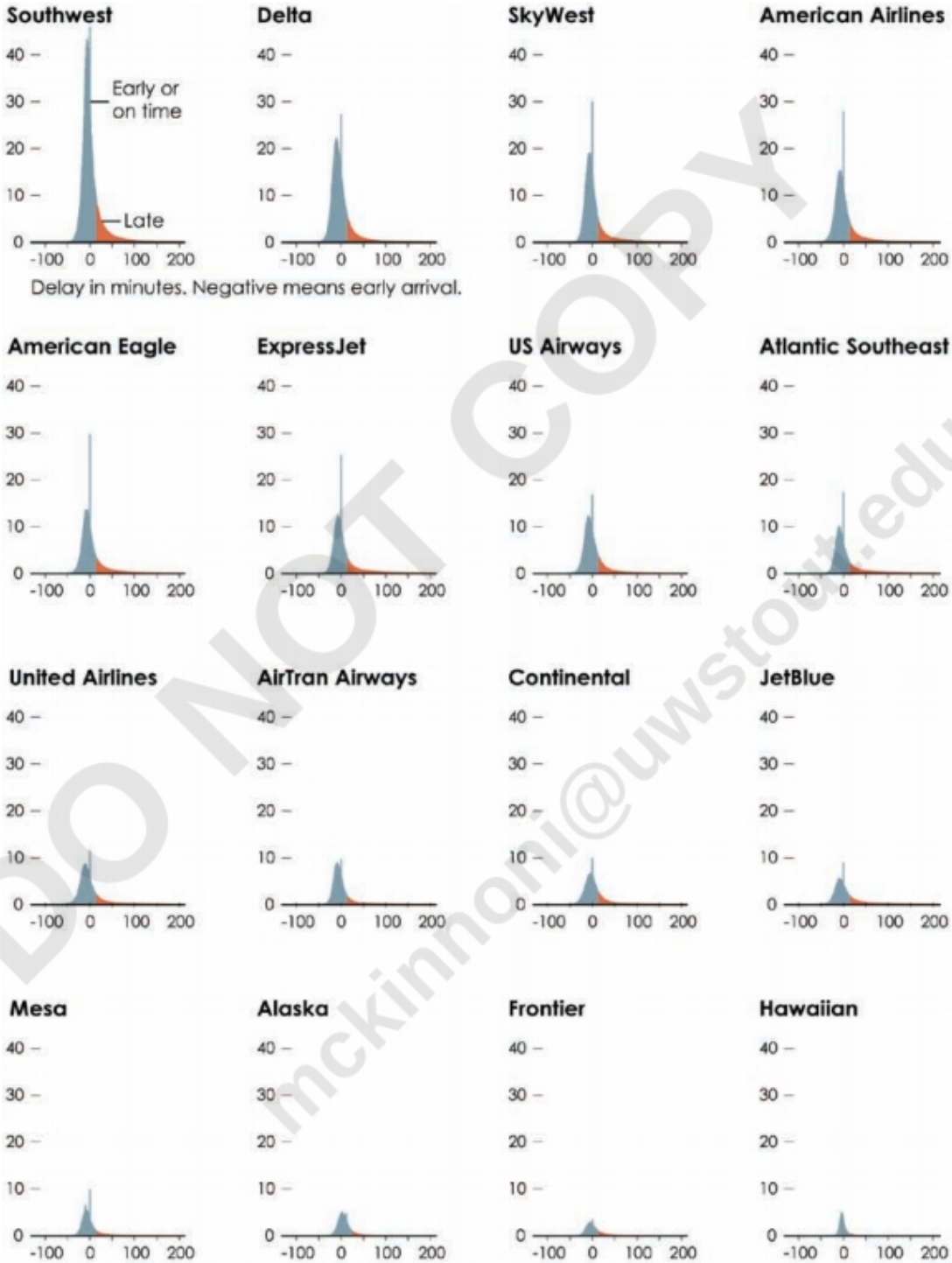


FIGURE 4-59 *Multiple histograms to compare distributions*

Notice the spike at the zero-minute mark, where airlines aim to be right on time? It looks like airlines either arrive right on time or whoever records the data rounds to on-time arrivals. You wouldn't see that with just means and medians.

WHAT TO LOOK FOR

Regardless of the type of visualization you use to explore distributions, look for peaks and valleys, range, and the spread of your data, which tell you a lot more than just the mean and median would. The visual analysis of raw data and the variation in between the summary statistics are almost always more interesting, so make use of the opportunity when you get it.

WRAPPING UP

Visualization can be a great tool to explore your data, and with advancing technologies, computers are less of a limiting factor than they were just a few years ago. So the key to getting the most out of your data—to understand what it represents and what it means—isn't so much about finding the right software than it is to learn how to use the tools you have and to know what questions to ask.

Consider what data you have and what you can get, where the data is from, how it was derived, and what all the variables mean, and let that extra information guide your visual exploration. If you use visualization as an analysis tool, you must learn as much as you can about your data. Even if your goal is to visualize data for presentation, exploration can lead to unexpected insights, which makes for better graphics.