



# Data Visualization

# Forces of Change

- First, there's a lot more data available in the world
- Second, software to analyze and visualize data is ubiquitous: ggplot2, plotly, matplotlib, ...
- Third, the cost of hardware is decreasing while computing power is increasing

# Data Visualization—Storytelling

- The use of data graphics for storytelling is a popular technique employed to engage an audience.

MAP A



MAP B

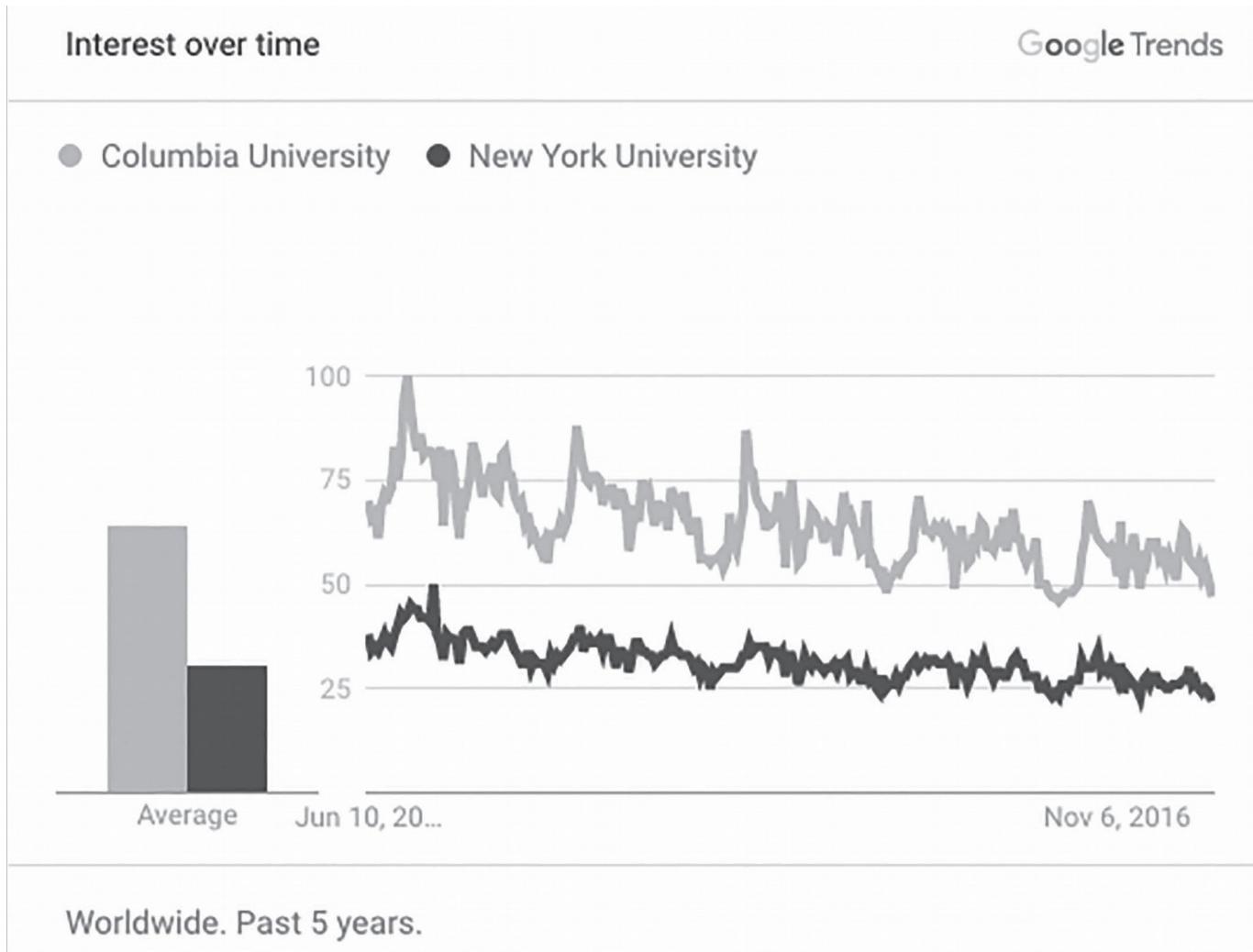


MAP C



**Figure 1.1** Viewing Manhattan through the lens of taxi hails, motor vehicle accidents, and Wi-Fi hotspots.

# Data Visualization—Storytelling



**Figure 1.2** Google search trends for New York University and Columbia University

# Charts

- Each type of chart is designed to show a type of data in a particular way.

For example:

- Horizontal bar charts show rank well by ordering bars from largest to smallest.
- Line charts convey a change over a specified period of time, such as the unemployment rate per month over a 12-month period.
- Point maps effectively demarcate precise locations, such as the address of each public school in a district.
- Filled or choropleth maps allow for the comparison of regions, such as the GDP of each African country. Each region is filled with a shade. The darker the shade the higher the value.

# Which map communicates the data best? The short answer is: it depends.

Chart A: Recycling bins grouped by zip code

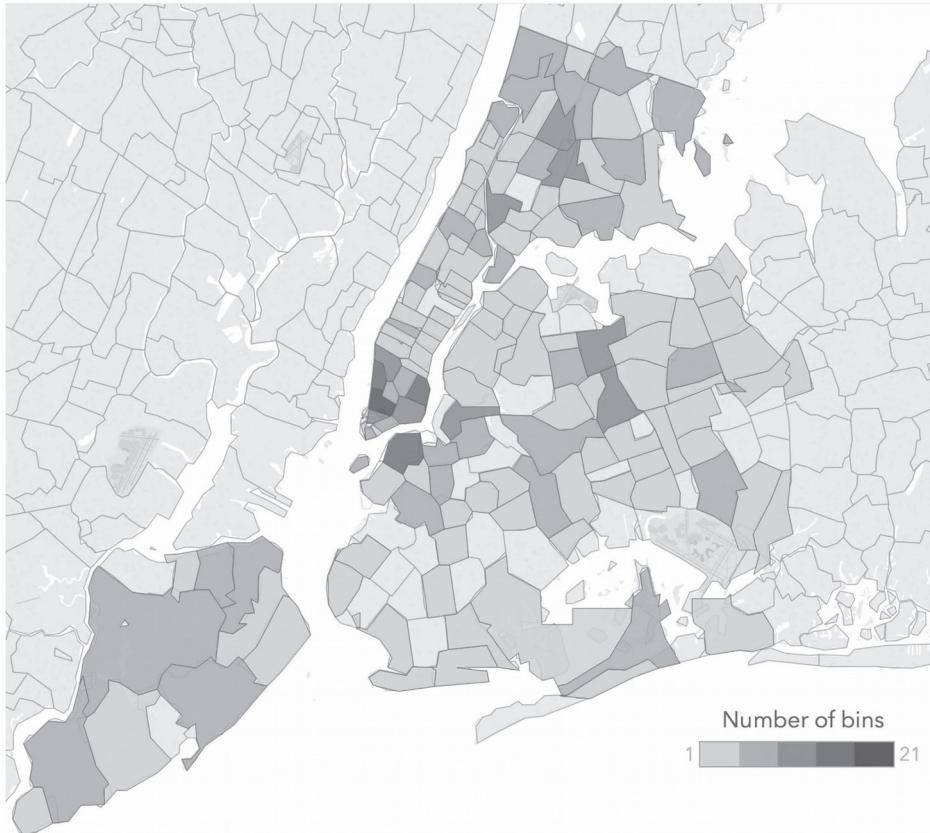
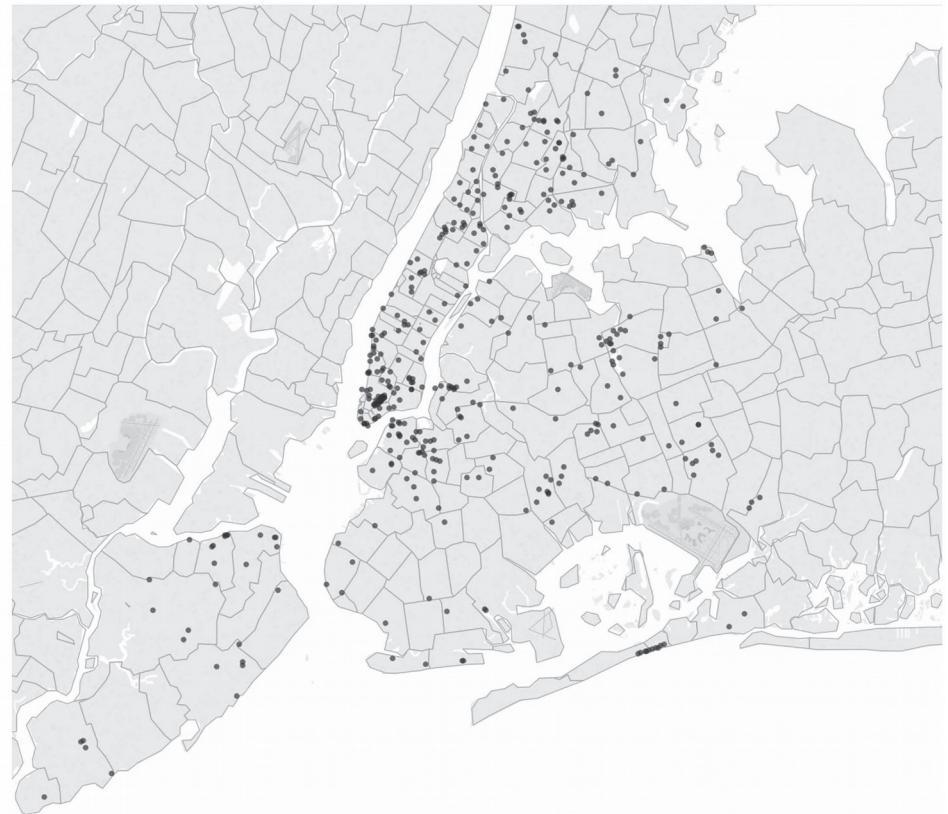


Chart B: Recycling bins grouped by individual locations



The type of chart you select is limited by the type of data.

**Table 3.1** The general data insight that corresponds to each data classification

Data	Example	Insight	Chart type
Categorical	Non-numeric data such as types of movies, books, or authors.	Comparisons, proportions	Vertical bar, column bar, horizontal bar, and bullet charts Pie, stacked bar, stacked 100% bar, stacked area, stacked 100% area, and a tree map
Univariate	One numeric variable, such as book price	Distributions, proportions, frequencies	Histogram, density plot, and a boxplot
Geospatial	Specific locations marked by the latitude and longitude, regions coded by zip code, city, state, country, or county boundaries	Locations, comparisons, trends	Choropleth filled-map, bubble map, point map, connection map, and isopleth map
Multivariate	Two or more numeric variables, for example, weight, height, and IQ	Relationships, proportions, comparisons	Scatterplot, scatterplot matrix, bubble, parallel coordinates, radar, bullet, and a heat map.
Time series	Years, months, days, hours, minutes, seconds, or date	Trends, comparisons, cycles	Line chart, sparkline, area, stream graph, as well as bubble, stacked-area, and vertical bar charts.
Text	Single words or phrases, such as keywords from restaurant reviews on Yelp	Sentiment, comparisons, frequency	Word cloud, proportional area chart using size bubbles or squares, histogram, and bar chart
Edge lists or adjacency matrices	Who contacts whom or who knows whom in a network	Connections, relationships, tie strength, centrality, interactions	Undirected network diagram and directed network diagram

# Python visualization Tools

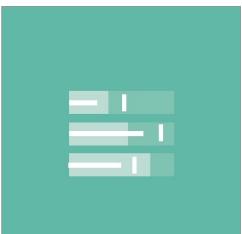
- **matplotlib** package is used to plot basic charts.
- **Seaborn** yield high-quality data graphics.
- **geoplotlib**
- **Bokeh**
- **Pandas**
- **Altair**
- **ggplot**
- **pygal**
- **plotly**

# Comparisons of Categories and Time

- **Questions:**
  1. What's the best? What's the worst?
  2. Who's ranked the highest? The lowest?
  3. How does performance compare to the target or goal? For example, did total sales exceed the forecast?
- **Insight:**
  - use comparisons to illustrate the similarities and differences among categories. This includes the minimum value, maximum value, rank, performance, sum, totals, counts, and quantities.
- **Data:** aggregated categorical data.

# Chart options:

**Table 3.2** Chart types to present categorical data

Chart type	Description and design considerations
Vertical bar	 <p>Bars are arranged vertically on the x-axis. Each bar represents a category or sub-category. The bar height measures the quantity (count) or sum.</p> <ul style="list-style-type: none"><li>• Keep bars the same color and shade when they measure the same variable (Wong, 2010).</li><li>• Use a zero baseline for the y-axis.</li><li>• Show negative values below the baseline.</li><li>• Keep the width of the bar about twice the width of the space between the bars (Wong, 2010).</li></ul>
Column bar	 <p>Column bar charts present two series for each category.</p> <ul style="list-style-type: none"><li>• Use different color shading for each series.</li><li>• Shade bars from lightest to darkest (Wong, 2010).</li></ul>
Horizontal bar	 <p>Bars are arranged horizontally, rather than vertically.</p> <ul style="list-style-type: none"><li>• Best used for ranking, such as first place, second place, third place.</li><li>• Arrange bars in descending order, from largest to smallest.</li></ul>
Bullet	 <p>Bullet charts display performance of a variable as a horizontal bar compared to a target or goal, represented by a vertical line. For example, a bullet chart could show whether the actual sales for a given period(s) are above/ below target sales.</p> <p>The performance measure (horizontal bar) overlays several shaded rectangles that represent qualitative ranges (e.g., 40% to the target goal, to indicate the performance progress).</p>

# Distributions

- **Questions:**

1. What are the highest, middle, and lowest values?
2. Does one thing stand out from the rest?
3. What does the shape of the data look like?

- **Insight:**

Use distribution charts to reveal outliers, the shape of the distribution, frequencies, range of values, minimum value, maximum value, and the median.

- **Data:** univariate or a single numeric variable.

# Chart options:

**Table 3.3** Chart types for showing distributions

Chart type	Description and design considerations
Histogram	<p>Histograms show frequencies of a single variable grouped into bins or frequency ranges on the x-axis. The y-axis of the histogram shows the frequency count or percentage.</p> <ul style="list-style-type: none"><li>• A large bin size can obscure the data.</li><li>• Adjust the size of the bins to best reveal the shape of the frequency distribution.</li></ul>
Density plot	<p>Density plots show probability densities and the distribution of a single variable. The area under the curve emphasizes the shape of the distribution of data.</p> <p>Annotate the mean to draw attention to the center of the distribution.</p>
Boxplot	<p>Boxplots show the range of a single variable including the minimum, 25th percentile, 50th percentile, median (not the average), 75th percentile, and the maximum value. Boxplots are helpful to spot outliers.</p>

# Proportions

- **Questions:**

1. What are the parts that make up the whole?
2. What part is the largest or smallest?
3. What parts are similar or dissimilar?

- **Insight:**

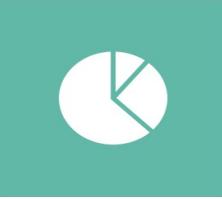
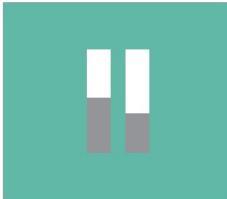
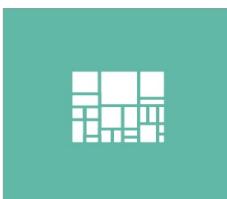
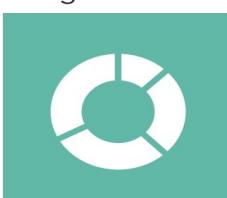
Use to show summaries, similarities, anomalies, percentage related to the whole (by category, subcategory, and over time).

- **Data:**

Single categorical variable with subcategories, two or more variables. A time dimension can also be included.

# Chart options:

**Table 3.4** Chart options for showing proportions

Chart type	Description and design considerations
Pie	 Pie charts show proportions within a whole. The slices are subcategories of a single category. Slices add up to 100% or 1. <ul style="list-style-type: none"><li>• Avoid using pie charts if all the slices are similar in size.</li><li>• Limit pie charts to eight slices or less (Wong, 2010).</li><li>• Label directly on the pie slices, rather than using a legend.</li><li>• Keep pie slices the same color. Use the whitespace between slices to differentiate the slices.</li></ul>
Stacked bar	 Stacked bar charts show proportions and quantities within a whole category. They show absolute and relative differences. <ul style="list-style-type: none"><li>• Limit the number of subcategories to four or less.</li><li>• Use stacked bars that add up to 100% to show the relative differences between quantities within each group.</li></ul>
Stacked area	 Stacked area charts highlight the absolute and relative differences between two or more series. They are line charts with the area below the line filled in with color. <p>To show relative differences use a 100% stacked area chart. Label each series directly, if possible over using a legend.</p>
Tree map	 Tree maps show parts of the whole by using nested rectangles. Each rectangle is designated a size and a shade of a color. This enables you to emphasize both the importance (usually shown by size) and urgency (usually represented by color) of a data point. <ul style="list-style-type: none"><li>• Used often for portfolio analysis to highlight similarities and anomalies.</li><li>• Usually require interactivity such as mouse-over, to read the subcategory labels for the smallest rectangles.</li><li>• This chart type is best used for analysis and exploration rather than presentation.</li></ul>
Doughnut	 Doughnut charts present proportions of a whole through slices of a doughnut shaped graphic. It is just a pie chart with the center missing. This type of chart can contain multiple series, represented as doughnuts arranged inside one another.

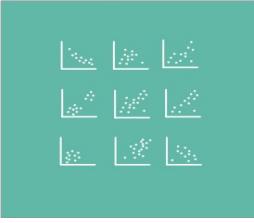
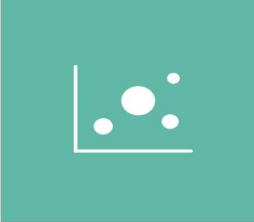
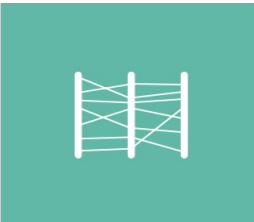
# Relationships

- **Questions:**
  1. Is the relationship positive, negative, or neither?
  2. How are x and y related to each other?
  3. What makes one group or cluster different from another?
- **Insight:**

Use to show outliers, correlations, positive, and negative relationships among two or more variables.
- **Data:** two or more numeric variables.

# Chart options:

**Table 3.5** Chart options for showing relationships between two or more variables

Chart type	Description and design considerations
Scatterplot	 <p>Scatterplots show relationships between two variables. For example, they show the change in x given y.</p> <ul style="list-style-type: none"><li>• Use to show positive or negative correlations, or linear and nonlinear relationships between variables.</li><li>• Labeling of every data point reduces readability but increases interpretation.</li></ul>
Scatterplot matrix	 <p>Scatterplot matrices help identify a correlation between multiple variables. It makes it easy to observe the relationship between pairs of variables in one set of plots.</p> <p>This chart type is best reserved for exploration versus presentation.</p>
Bubble chart	 <p>A bubble chart is a scatterplot that shows relationships between three or four variables. The position of the bubble shows the relationship between the x and y variables.</p> <ul style="list-style-type: none"><li>• The bubble size is based upon a numerical variable, such as population, or sales.</li><li>• The bubble color is best reserved for categorical data, such as region.</li><li>• Bubble charts are best when the bubble sizes vary significantly.</li></ul>
Parallel coordinates	 <p>Parallel coordinates map each column in a data table as a vertical parallel line with its own axis. Each observation (row) is represented by a point on the parallel line. That point is then connected to the next point on the next parallel line by a horizontal line.</p> <ul style="list-style-type: none"><li>• Use the technique of highlighting the lines that touch any number of values in either of the categories, called brushing, to provide data context while focusing on select series.</li><li>• This chart type is best reserved for exploration over presentation.</li></ul>

# Chart options:

Table 3.5 (Continued)

Chart type	Description and design considerations
Radar	<p>Radar charts compare multiple numerical variables. They show which variables have similar values, and to spot outliers, high values, and low values. Each variable is provided its own individual axis, but the axes are arranged radially. Every observation connects to form a shaded polygon.</p> <ul style="list-style-type: none"><li>• Limit the number of variables to reduce the number of axes to increase readability.</li><li>• Scaling is affected when variables have dissimilar minimum and maximum ranges.</li></ul>
Heat map	<p>A heat map is a graphical representation of a table of data. The individual values are arranged in a table/matrix and represented by colors. Use grayscale or gradient for coloring. Sorting of the variables changes the color pattern.</p>
Small multiples	<p>A series of similar graphs that use the same scale. This allows for easy comparisons between variables. A single chart represents each categorical variable, such as sales personnel; the individual charts are grouped together on a single display.</p> <ul style="list-style-type: none"><li>• Allows easy comparisons by using the same scale for each chart</li><li>• Avoid showing too much detail in any individual chart.</li></ul>

# Locations

- **Questions:**
  1. Where can the most or least be found?
  2. How does one area compare to another?
  3. What is the distance from one place to another?
  4. How does a variable change by location?
- **Insight:**

Use to demonstrate similarities and differences by location, density, distance, and counts (such as population).
- **Data:**

latitude and longitude, zip codes, census tracks, cities, states, countries, and regions.

# Chart options:

**Table 3.6** Chart options for showing locations

Chart type	Description and design considerations
Choropleth or filled maps	<p>Choropleth maps fill regions with color. A color gradient and density distinguishes regions from one another. Use to compare different regions such as continents, countries, states, territories, zip codes, or census tracks.</p> <p>Provide a legend. Keep the gradient of colors within a limited range. This will allow the reader to easily compare the regions.</p> 
Point map	<p>Point maps show a specific location. These dots can vary in size, form, or color.</p> <p>Point maps illustrate density when the individual locations are easily distinguishable. Too many points can obscure the location. Consider the size of the points and the labeling of the points.</p> 
Symbol or bubble map	<p>Symbol maps are point maps that use different sized bubbles or shapes to mark a location. These symbols are sized by a certain variable.</p> <p>Too many or too large bubbles can obscure the locations referenced.</p> 
Connection or path maps	<p>Connection maps graph a line from one or more points to another. Use to show distances or pathways between one or more locations.</p> <p>Use high contrasting colors for the map projection and the lines that connect the points. Avoid too many overlapping lines.</p> 
Geographic heat map (Isopleth)	<p>Isopleth maps show gradual change over geography. This technique uses a color value (lightness/darkness) and hue to show density. The color value is not constrained by boundary lines (e.g., such as zip code).</p> <p>Use for events that are continuous and unbounded (e.g., such as temperature).</p> 

# Trends

## Showing Comparisons or Composition Over Time

- **Questions:**

- What changed today from yesterday?
  - How does time of year affect sales, results, outcomes, etc.?
  - What times are the most popular? Least popular?

- **Insight:**

- Change over time, cycles, or comparisons over time.

- **Data:**

- Time dimension such as year, month, day, hour, minute, second, date, quarter, season, century, decade, etc

# Chart options:

**Table 3.7** Chart options for showing trends

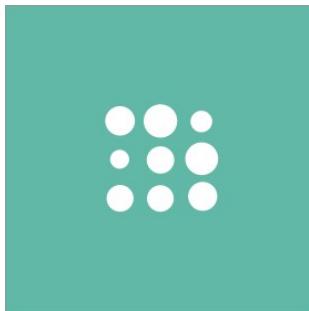
Chart type	Description and design considerations
Line chart	<p>Line charts show the change over time for one or more series (sales per hour). The line connects each data point in the series (shown or not). The y-axis baseline should be equal or less than the minimum value in the data.</p> <ul style="list-style-type: none"><li>• Show four or fewer series of lines on a line chart (Wong, 2010).</li><li>• Label each series directly or use an ordered legend.</li></ul>
Sparkline	<p>A sparkline is a line chart without axes or much detail. It is a small graphic designed to give a quick representation of change over time.</p> <ul style="list-style-type: none"><li>• Not intended to provide the quantitative precision of a normal line graph.</li><li>• Label the last data point to provide additional information.</li></ul>
Area graph	<p>Area graphs are line charts with the area below the line filled in with color. They can show a single series or multiple time series using stacked areas.</p> <p>Use the same color for the line and the area beneath it. See Table 3.4 for use of stacked areas to show proportional change over time.</p>
Stream graph	<p>Stream graphs show changes over time for different data series. Color is used to distinguish the categories. Each stream represents a single category proportional change over time. Stream graphs are used to provide a general overview, not when accuracy is important.</p> <p>Use for large time series data sets with five or fewer categories.</p>

# Word Frequency and Sentiment

- **Questions:**
  - How many times does a given word or phrase appear?
  - What words or phrases appear most often? Least often?
  - What words appear together?
  - Are most words or phrases positive or negative?
- **Insight:**
  - Frequency or counts of words and phrases. The count of the positive or negative direction of the sentiment of the words or phrases.
- **Data:**
  - Text as single words, or n-grams (one or more words that appear together in text).<sup>22 / 36</sup>

# Chart options:

**Table 3.8** Chart options for showing sentiment and frequency words

Chart type	Description and design considerations
Word cloud 	<p>Words are arranged in a cluster or cloud of words. Words can be arranged in any format: horizontal lines, columns, or within a shape.</p> <p>Color is used to categorize words by sentiment, or another categorical variable.</p>
Proportional bubble area chart 	<p>Words are ranked by their frequency. The frequency is represented by sized bubbles or squares. The bubbles /squares are arranged in a grid with words on the x-axis and observation on y.</p> <p>Works well for the top 10 words (difficult to view beyond that).</p>

# Connections and Networks

- **Questions:**

- Who is closest to whom? Who is connected to whom?
  - Who is the most popular? Who is the least?
  - What communities exist and who are their members?
  - What is the strength of the relationship between two entities?

- **Insight:**

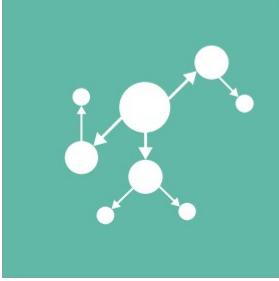
- see relationships, patterns, centrality, or interactions. This is shown by the width, color, or arrowheads on a line to communication relationships.

- **Data:**

- Edge lists or adjacency matrices show relationships between entities.

# Chart options:

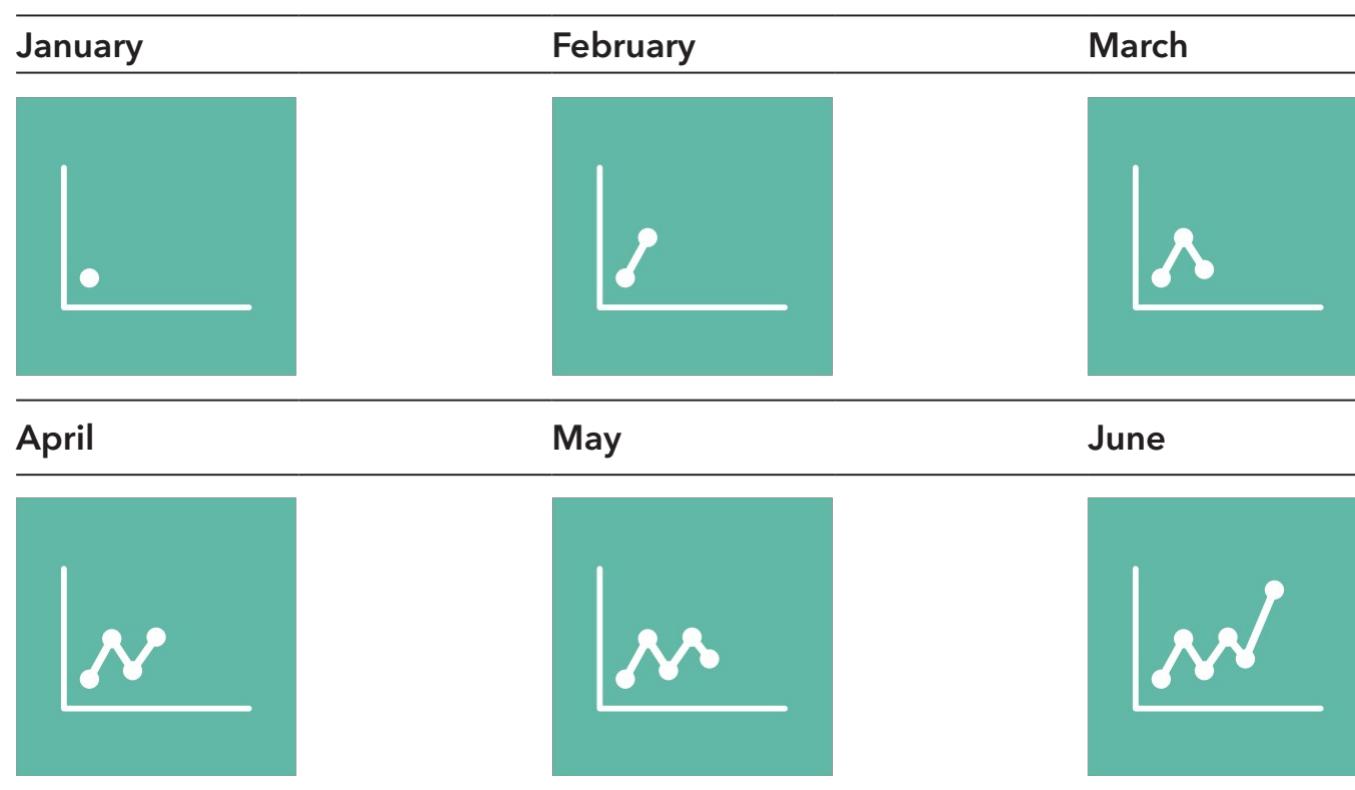
**Table 3.9** Chart options for showing networks

Chart type	Description and design considerations
Network diagram (undirected)	<p>Undirected network diagrams depict equal relationships between individual entities. Each entity is referred to as a node, represented as a bubble. The relationships between the entities are shown as lines known as edges. The thicker the line, the stronger the relationship. The position of the nodes shows centrality in the network and distance from other nodes.</p> <p>Minimize edge (line) crossing (Ognyanova, K, 2016).</p> 
Network diagram (directed)	<p>Use directed layout to show the orientation of the relationship between nodes. Communicate the strength of a relationship by the width and the direction of using arrow heads.</p> 

# ANIMATION

- Trend Animation

**Table 3.10** A trend animation shown frame by frame

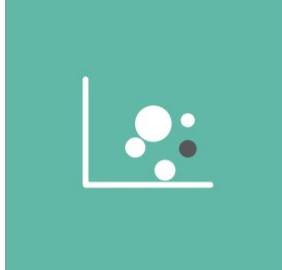
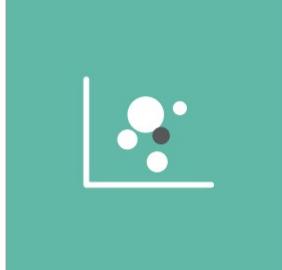
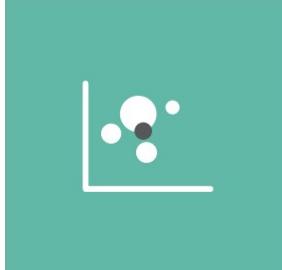
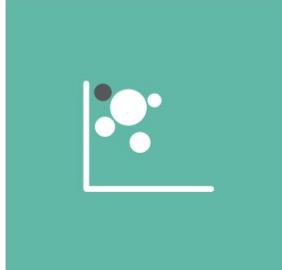


# ANIMATION

- Transition Animation

Table 3.11 shows a transition animation of city budget by city and population and budget subcategory.

**Table 3.11** A transition animation shown frame by frame

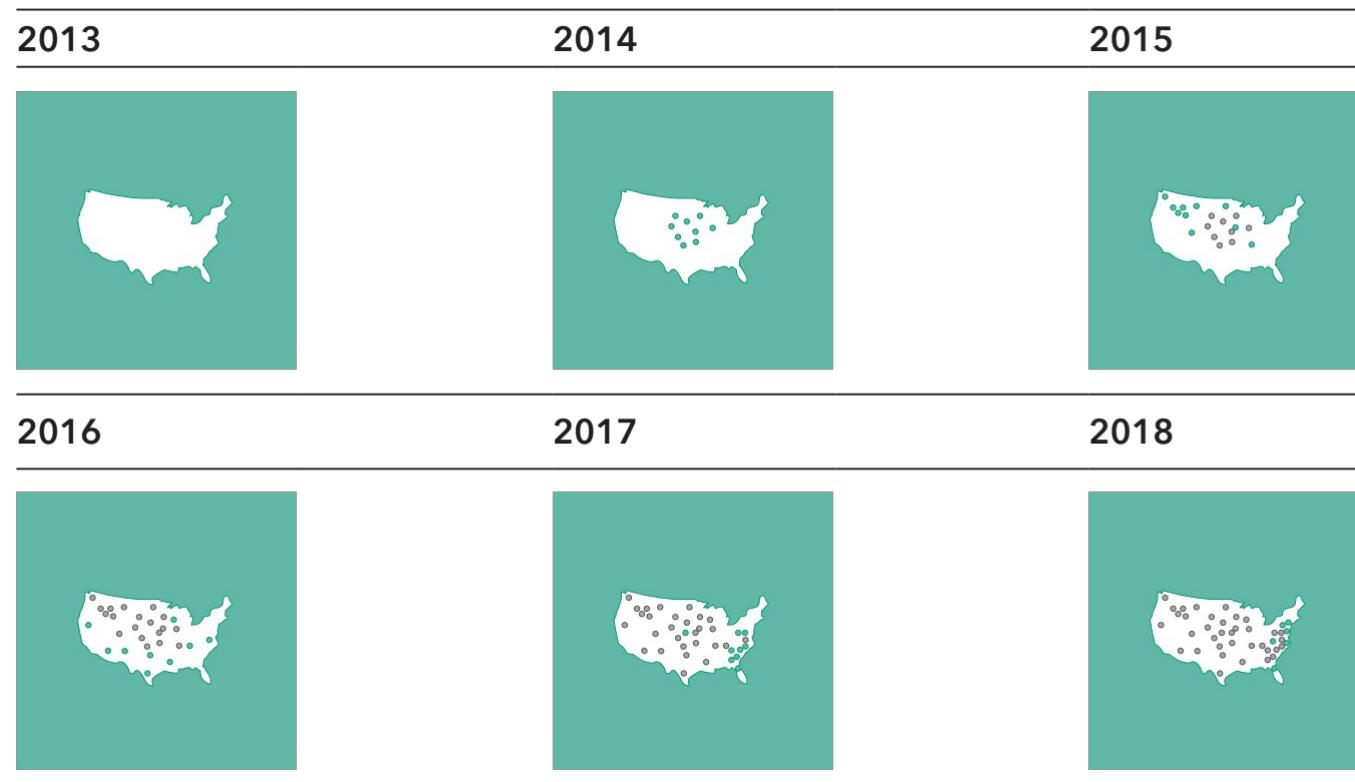
Public Safety	Admin.	Community
		
Government	Public Works	Development
		

# ANIMATION

- Trace Animation

Table 3.12 shows a map of store openings over a six-year period.

**Table 3.12** A trace animation shown frame by frame



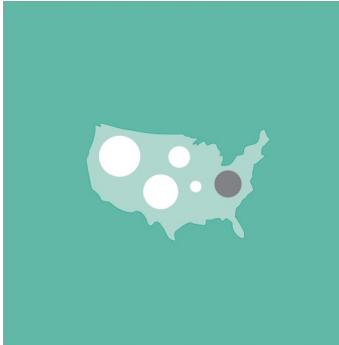
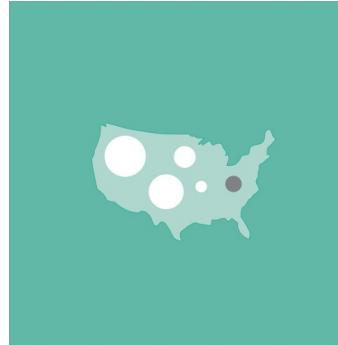
# Interactive visualization

- Interactive data visualizations allow the audience to explore a data set through a visual interface.
- Usually used in dashboards.

# Interactive visualization

- Select Interaction
  - Click on a data point to mark or highlight it. This is used for tracking data points.

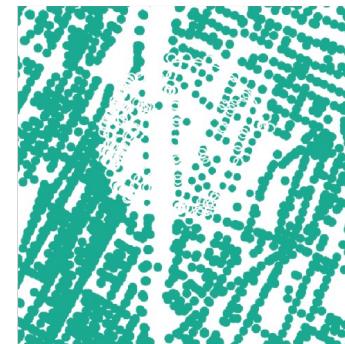
**Table 3.13** Marking a point of interest to highlight and observe

Click on data point	Data point changes color	Point size changes based on time or another attribute
		

# Interactive visualization

- Explore Interaction
  - Table 3.15 presents a scatterplot with a point selected and a label with a description of the data point.

**Table 3.14** Exploring a point map using panning and zooming

Original view	Panning	Zooming	Selecting and zooming
			

# Interactive visualization

- Explore Interaction
  - Table 3.14 shows panning and zooming interactions on a point map of New York City

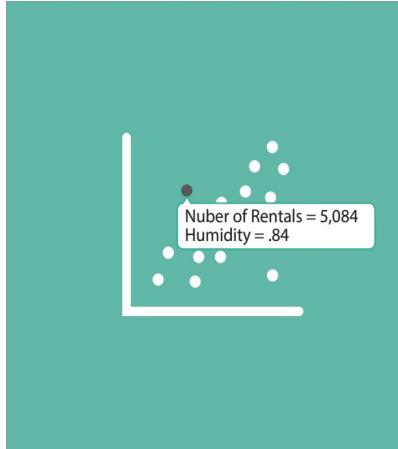
**Table 3.14** Exploring a point map using panning and zooming

Original view	Panning	Zooming	Selecting and zooming
			

# Interactive visualization

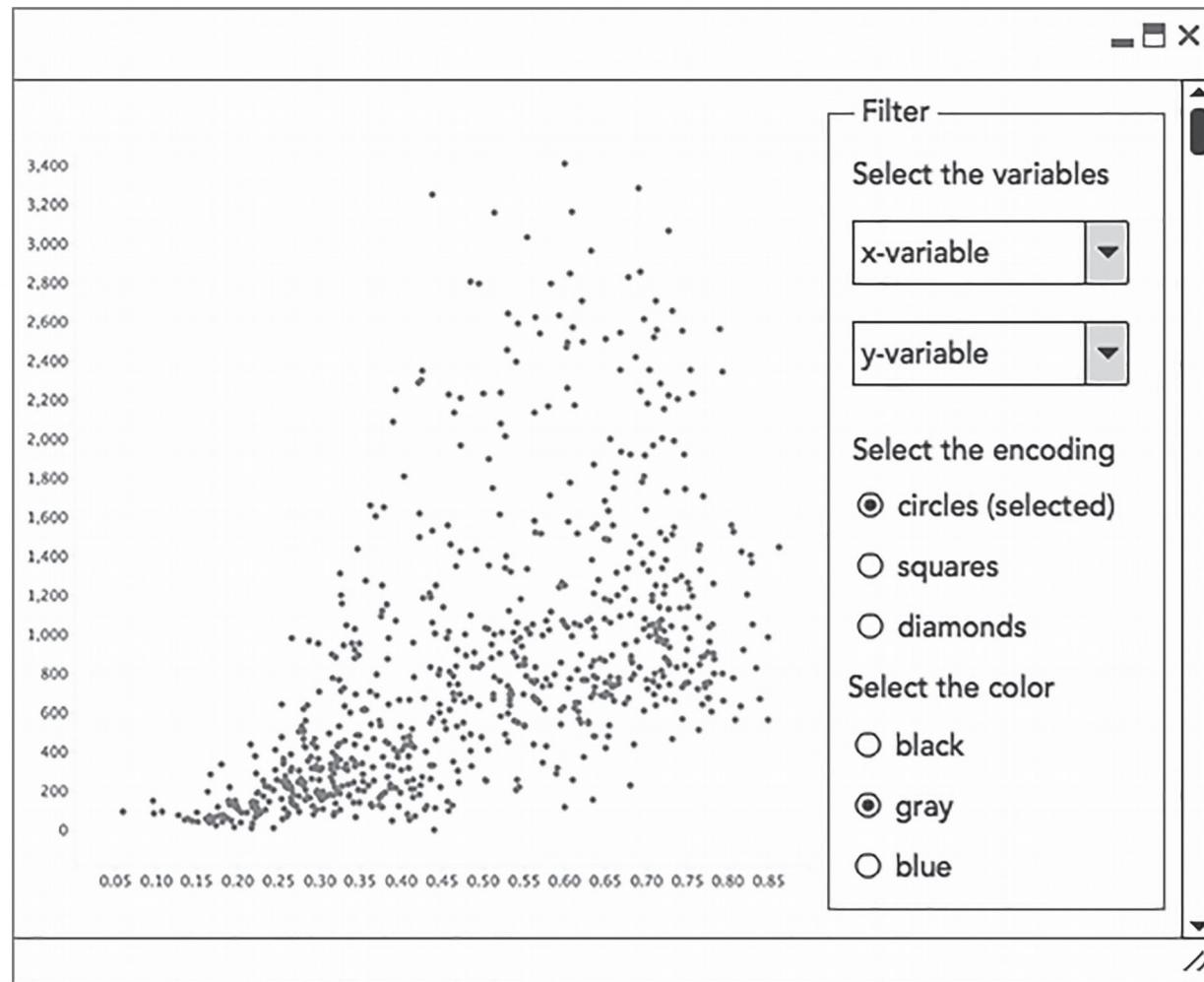
- Abstract and Elaborate Interaction

**Table 3.15** Clicking on or hovering over a data point reveals information

Original view	Hover or click
	

# Interactive visualization

- Reconfigure and Encode interaction

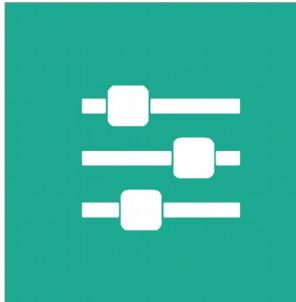
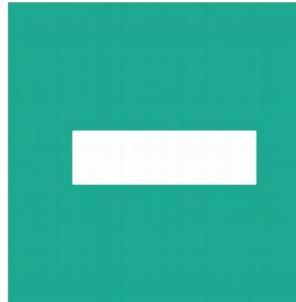


**Figure 3.2** An interface for selecting the variables and encoding attributes

# Interactive visualization

- Filter interaction

**Table 3.16** Six types of filters

<b>Single value list</b> 	<b>Radio buttons</b> 	<b>Drop down list</b> 
<b>Checkbox</b> 	<b>Slider</b> 	<b>Search</b> 

- Reconfigure and Encode