Creating Good Graphics

Identifying the problem

Let's start out by looking at some examples of less-than-effective charts.

Example 1: Pie Chart Poll Results

See if you can spot the problem with this one, published in the March 16, 2021 Scottsbluff, NE Star Herald (wtfViz 2021).

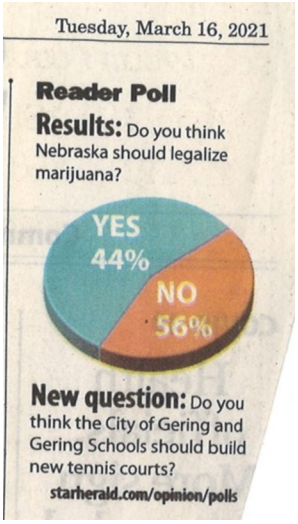


Figure 1: Scottsbluff Star Herald Reader poll.

Discuss:

- What is wrong with this chart?
- Do you think it might be misleading? If so, how?
- Do you think the mistakes were intentional?

Example 2: High Support

While I didn't intend this section to have a theme, here's another chart on a similar topic from CBS News (wtfViz 2022).

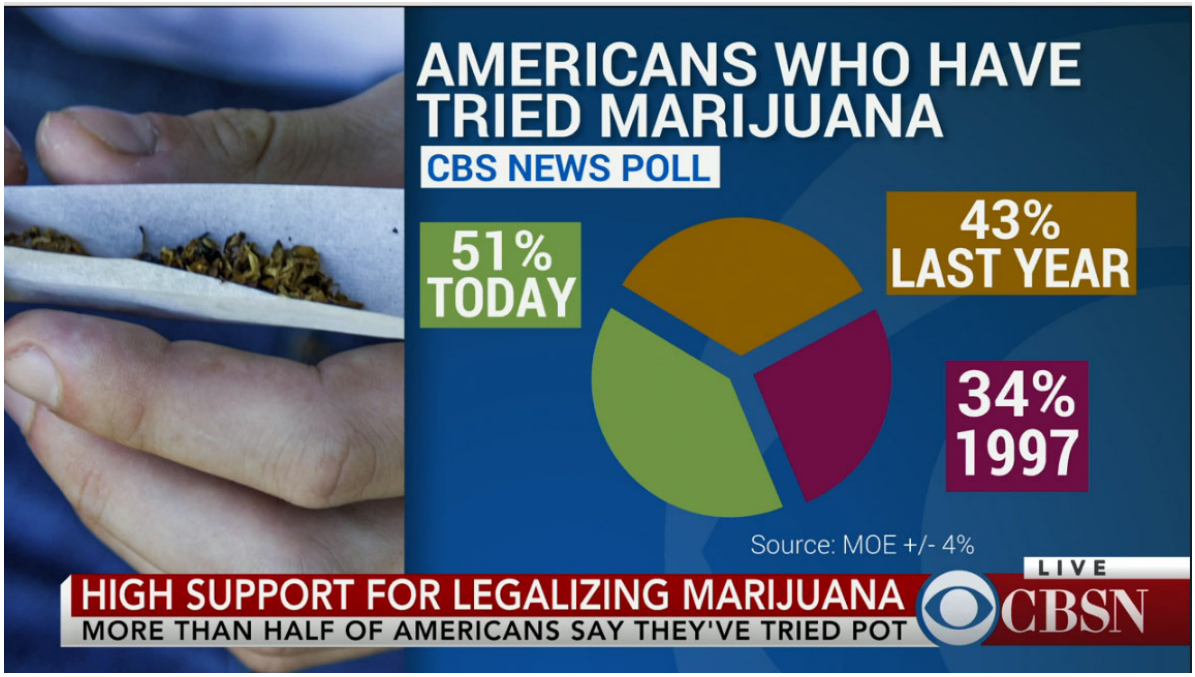


Figure 2: High support.

Discuss:

- What is wrong with this chart?
- What would you change to more accurately represent the data?
- Do you think the mistakes were intentional?

Example 3: Gas Prices

Pie charts aren't the only chart type that commonly are presented wrong. Here's a bar chart that generated a lot of conversation online, from Express Web Desk (2018).

Discuss:

- What is wrong with this?
- What design choices contribute to the problems?

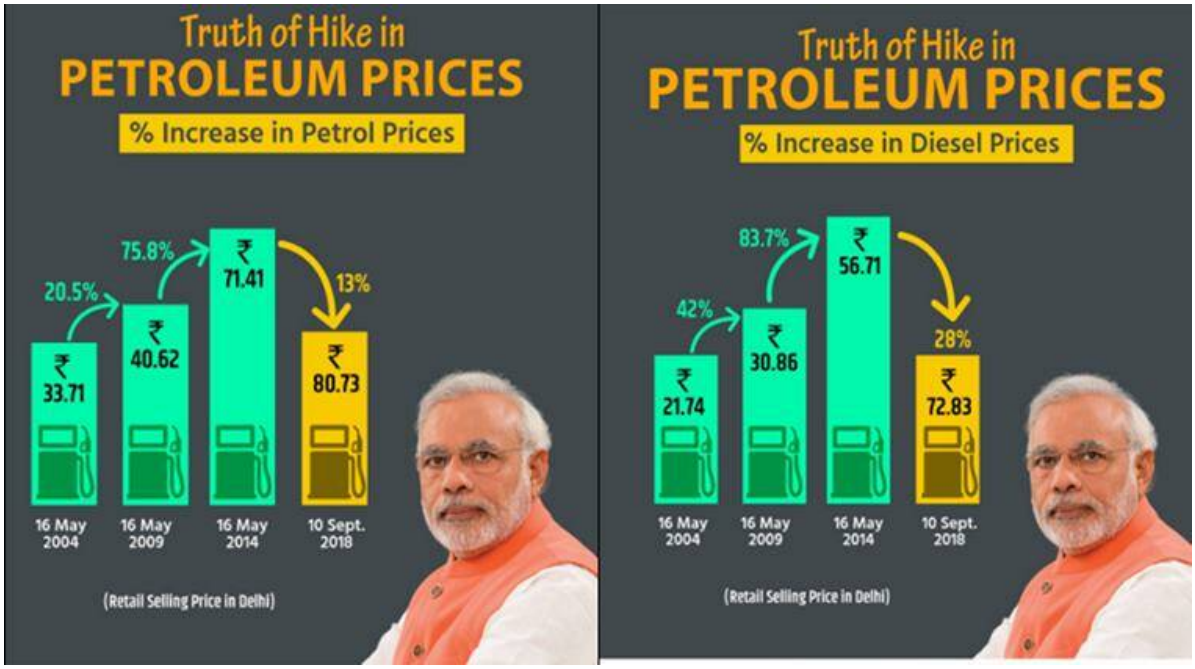


Figure 3: Gas and Diesel price changes in India (2004 - 2018).

- Do you think this was intentionally designed to be misleading? Why or why not?

Example 4: Information Overload

Not all charts are intentionally designed to be misleading. Sometimes, the desire to show all of the data goes awry. Here is an attempt to show 6 variables using three location variables, color, column length, and column width (Pies 2013). The original source doesn't specify what variables are plotted, so analyze this based on its' form, rather than the data it shows.

Discuss:

- What problems do you have reading this chart?
- Can you compare the quantities of all 6 variables shown? Why or why not?

(Yes, the blog this chart is taken from is satirical. This is *not* a recommended graphical form.)

These are some of my favorite examples, but of course, there are also bad charts in the scientific literature (Broman 2018). The goal of this module is to ensure that as you work on research, you will create effective graphics that are accessible and well-designed.

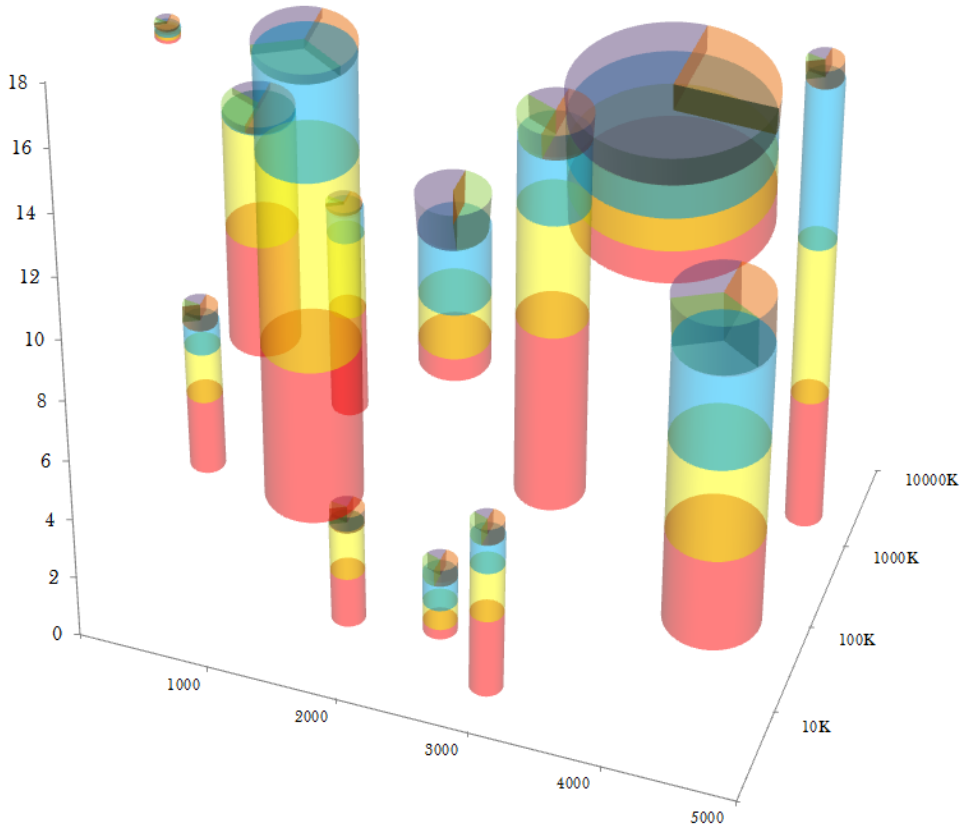

18
16
14
12
10
8
6
4
2
0
1000
2000
3000
4000
5000
10000K
1000K
100K
10K

Figure 4

Designing Good Charts

Why Graphics Matter

Graphics are a form of **external cognition** that allow us to think about the *data* rather than the *chart*.

That is, graphics are a tool to make it easier for us to think about what the data means.

Good graphics take advantage of how the brain works, leveraging

- preattentive processing
- perceptual grouping
- awareness of visual limitations

Good graphics also depend on the data: the chart type should be chosen based on the types of variables you want to display, the amount of data you have, and the results you want to highlight.

Example: Hertzsprung Russell Diagram

Our first example of a good chart is the Hertzsprung-Russell Diagram (Wikipedia contributors 2023a).

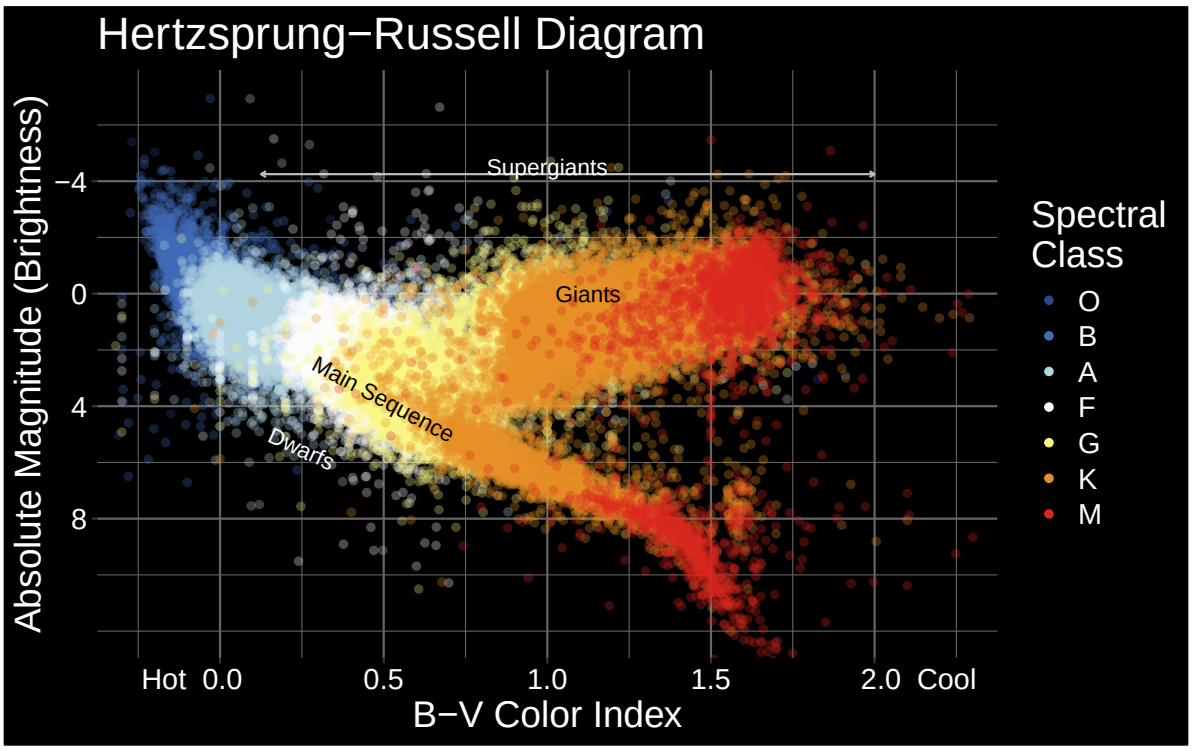


Figure 5: The Hertzsprung Russell diagram. Discovered independently by Ejnar Hertzsprung (1873–1967) and Henry Norris Russell (1877–1957). The diagram plots the color index of the star against the brightness (absolute magnitude) of the star. As a result, it is possible to discern that these two variables are related and change together over a star's life cycle: a hypothesis that only came to be because of this chart.

John Tukey, a famous statistician often considered the father of statistical graphics, wrote in Exploratory Data Analysis (1977):

> The greatest value of a picture is when it forces us to notice what we never expected to see.

This chart is an excellent example of the value that good graphics create in research: they can help us understand our data in a new way, leading to innovations and new research directions.

Discuss:

- What variables are mapped to the following chart dimensions?
  - X location
  - Y location
  - color

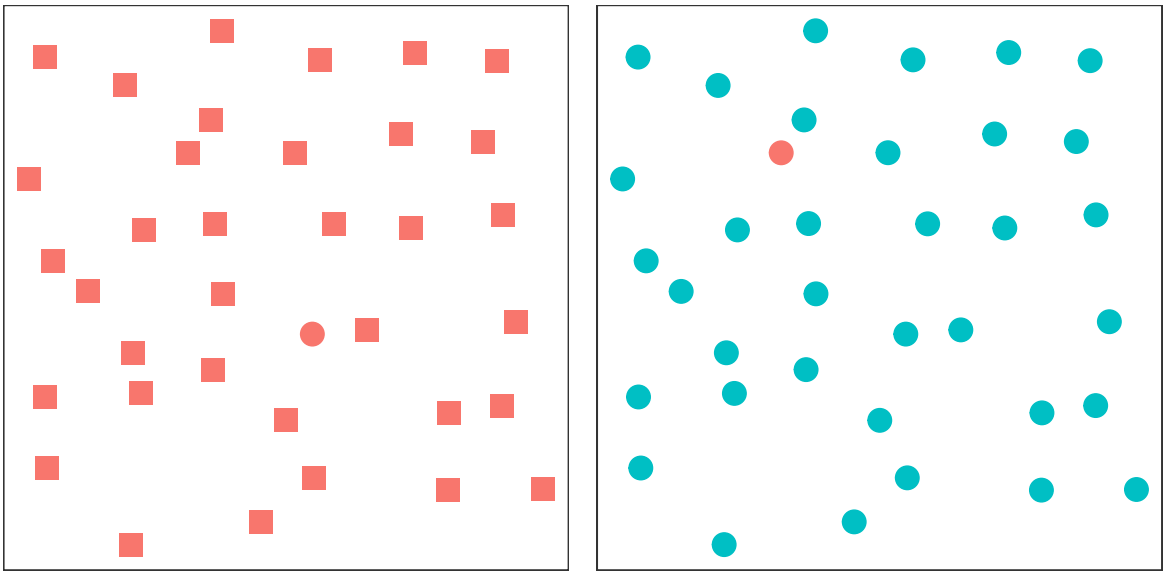
- What other information is present on the chart that is not specifically a data value?
- What does this chart do well?
- What design features "work"?
- What don't you like?

I've used data from the HYG Database to generate this chart. Only stars within 500 AU are shown.

Perceptual Principles

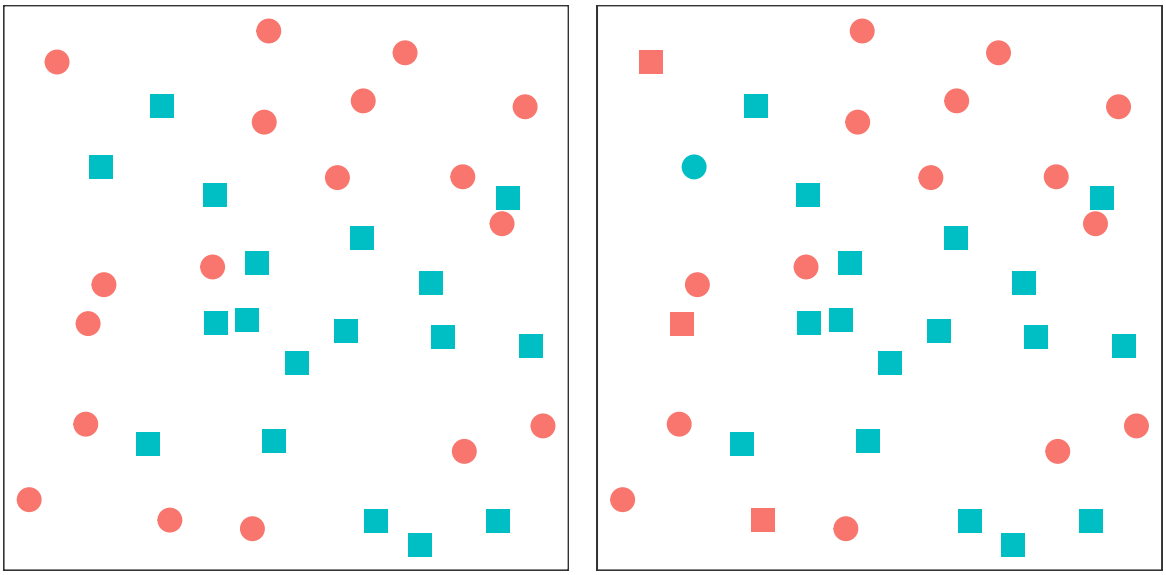
Preattentive Perception

- Occurs automatically (no effort)
- Color, shape, angle
- Combinations of preattentive features require attention
  - Double-encoding (using multiple features for the same variable) is ok



(a) Shape (b) Color

Figure 6: Two scatterplots with one point that is different. Can you easily spot the different point?



(a) Shape and Color (dual encoded)

(b) Shape and Color (different variables)

Figure 7: Two scatterplots. Can you easily spot the different point(s)?

Perceptual Grouping

Perception is interesting, because when faced with ambiguous images, we can learn something about how our brains work.

In Figure 8, the image is ambiguous, and depending on what orientation we use, the figure can be either a rabbit or a duck. That is, the same image can be interpreted in two different ways, depending on the contextual information we have.

How do you describe the components of Figure 9? Something like "Three circles, a black triangle outline, and a white triangle over top?" The components of Figure 9 are 3 pac-man shapes and 3 angles - that's what's actually there. The appearance of triangles and depth information is a construction that occurs within the brain.

Our brains use past experience to simplify the visual input from the world. It's much easier to describe Figure 9 if you see triangles and circles rather than angles and pac-men shapes.

The perceptual rules that describe how we make sense of the world are the **Gestalt laws**, but the general principle is that the whole is more than the sum of the parts - as in Figure 9, the brain constructs meaning from individual pieces that, when combined, produce greater meaning - for example, a white triangle, a black triangle outline, and 3 circles.

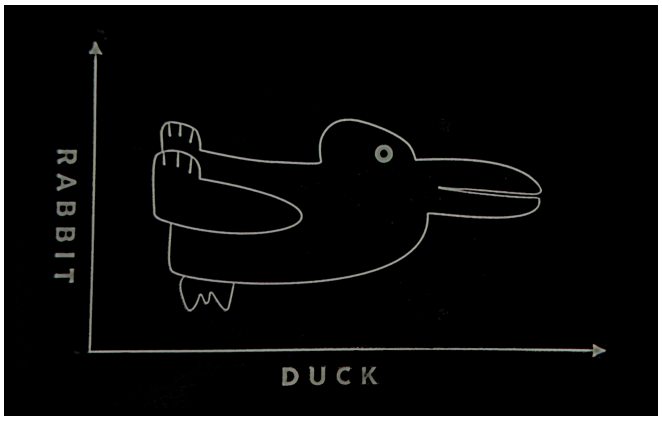


Figure 8: Is this a rabbit or a duck?

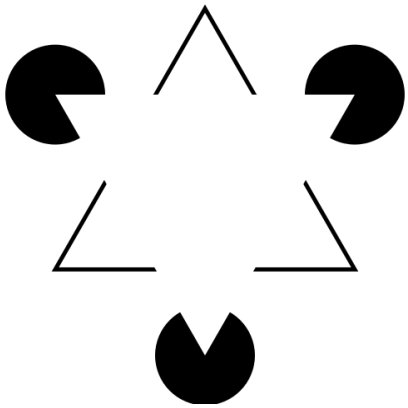
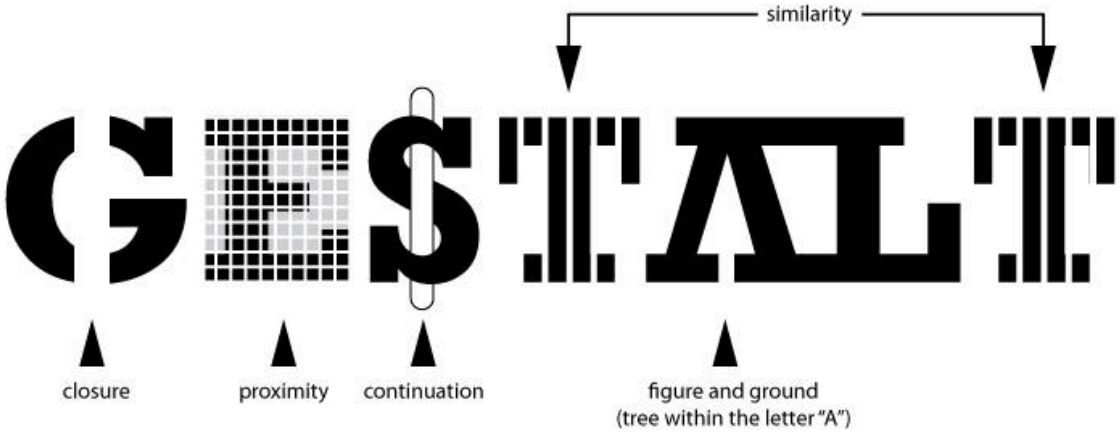


Figure 9: What do you see in this image?



You can read about the gestalt rules here (Wikipedia contributors 2023b), but they are also demonstrated in the figure above.

In graphics, we can leverage the gestalt principles of grouping to create order and meaning. If we color points by another variable, we are creating groups of similar points which assist with the perception of groups instead of individual observations. If we add a trend line, we create the perception that the points are moving "with" the line (in most cases), or occasionally, that the line is dividing up two groups of points. Depending on what features of the data you wish to emphasize, you might choose different aesthetics mappings, facet variables, and factor orders.

Suppose I want to emphasize the change in state population between 2010 and 2019. I could use a bar chart (showing a few states bordering Nebraska for space), a line chart (showing one line per state), or a box plot (showing variability over time).

```
library(readr)
```

```
Attaching package: 'readr'

The following object is masked from 'package:scales':

    col_factor
```

```
library(dplyr)
library(tidyr)
```

```
Attaching package: 'tidyr'

The following object is masked from 'package:magrittr':

    extract
```

```
library(ggplot2)
library(geomtextpath)
library(stringr)
theme_set(theme_bw())
counties <- read_csv("https://raw.githubusercontent.com/evangambit/JsonOfCounties/master/co
```

```
Rows: 3142 Columns: 237

-- Column specification --------------------------------------------------------
Delimiter: ","
chr   (4): name, fips, state, zip-codes
dbl (233): land_area (km^2), area (km^2), longitude (deg), latitude (deg), n...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
popsubset <- counties |> select(name, fips, state, starts_with("population")) |>
  pivot_longer(starts_with("population"), names_to = "year", values_to = "population") |>
  mutate(year = parse_number(year))

state_pop <- popsubset |> group_by(state, year) |>
  summarize(population = sum(population))
```

```
`summarise()` has grouped output by 'state'. You can override using the
`.groups` argument.
```

```r
pal <- viridis::viridis(10)
bar_state_by_year <- state_pop %>%
  filter(state %in% c("NE", "IA", "CO", "MO", "KS")) %>%
  ggplot(aes(x = state, y = population, fill = factor(year))) +
  geom_col(position = "dodge", color = "black") +
  scale_fill_manual("Year", values = pal) +
  scale_x_discrete("State") +
  ylab("State Population, 2010-2019") +
  theme(legend.position = c(1, 1), legend.justification = c(.95,.95))
```

```
Warning: A numeric `legend.position` argument in `theme()` was deprecated in ggplot2
3.5.0.
i Please use the `legend.position.inside` argument of `theme()` instead.
```

```r
line_state_by_year <- state_pop |>
  filter(state %in% c("NE", "IA", "CO", "MO", "KS")) |>
  mutate(label = str_replace_all(state, c("NE" = "Nebraska", "IA" = "Iowa", "CO" = "Colorad
  ggplot(aes(x = year, y = population, color = state)) +
  geom_textpath(aes(label = label), text_only = F) +
  scale_x_continuous("Year", breaks = (1005:1010)*2) +
  guides(color ="none") +
  ylab("Population")

box_state <- state_pop %>%
  ggplot(aes(x = factor(year), y = population)) +
  geom_boxplot() +
  xlab("Year") +
  ylab("Population")
```

Three versions of the same data that emphasize different aspects of the dataset.

Which one best demonstrates that in every state and region, the murder rate decreased?

Perceptual and Visual Limitations

Our perceptual system is not infallible, and some people have additional challenges to work with.

Color

About 10% of the XY population and 0.2% of the XX population has some form of colorblindness or color deficiency.

Here are some basic tips for choosing color schemes for your charts.

- Do not use rainbow color gradient schemes
  - because of the unequal perception of different wavelengths, these schemes are *misleading* - the color distance does not match the perceptual distance.
- Avoid any scheme that uses green-yellow-red signaling if you have a target audience that may include colorblind people.
- To "colorblind-proof" a graphic, you can use a couple of strategies:
  - double encoding - where you use color, use another aesthetic (line type, shape) as well to help your colorblind readers out
  - If you can print your chart out in black and white and still read it, it will be safe for colorblind users. This is the only foolproof way to do it!
  - If you are using a color gradient, use a monochromatic color scheme where possible. This is perceived as light -> dark by colorblind people, so it will be correctly perceived no matter what color you use.
  - If you have a bidirectional scale (e.g. showing positive and negative values), the safest scheme to use is purple - white - orange. In any color scale that is multi-hue, it is important to transition through white, instead of from one color to another directly.
- Be conscious of what certain colors "mean"
  - Leveraging common associations can make it easier to read a color scale and remember what it stands for (e.g. blue for cold, orange/red for hot is a natural scale, red = Republican and blue = Democrat in the US, white -> blue gradients for showing rainfall totals)

  - Some colors can can provoke emotional responses that may not be desirable.[1]
  - It is also important to be conscious of the social baggage that certain color schemes may have - the pink/blue color scheme often used to denote gender can be unnecessarily polarizing, and it may be easier to use a colder color (blue or purple) for men and a warmer color (yellow, orange, lighter green) for women[2].

Working Memory

We can hold about 7 items in "working memory" and maintain these by rehearsing the content. As a result, using a legend with more than 7 items will create additional cognitive load on those viewing the visualizations. Wherever possible, keep cognitive limitations in mind

Alt Text and Accessibility

Some individuals may have limited vision or visual processing ability. To make your charts accessible, you should always provide alt-text for your graphics.

It can also help to use larger text size and/or fonts that are easier to read for individuals with e.g. dyslexia (Zorzi et al. 2012). You can customize your charts to make these changes, though instructions will vary based on which plotting system you are using.

Chart Types

- Graph Galleries
  - Python
  - R
  - D3.JS
  - Matlab
  - SAS
  - Stata
  - Excel

Choose your chart type based on your data types as well as what you want to show:

- If you have continuous data to show in X and Y, a scatterplot is a great idea.
- If you have categorical data in X and continuous data in Y, you may want to consider a boxplot, violin plot, or jittered scatterplot.

[1] When the COVID-19 outbreak started, many maps were using white-to-red gradients to show case counts and/or deaths. The emotional association between red and blood, danger, and death may have caused people to become more frightened than what was reasonable given the available information.

[2] Lisa Charlotte Rost. What to consider when choosing colors for data visualization.

- If you have too much data in X and Y for a scatterplot, or you have a lot of categories in X or Y and continuous data in the other axis, you might try binning the continuous variable to create a heatmap.

Learn More

- Statistical Computing in R and Python chapters:
  - Data Visualization Basics
  - Exploratory Data Analysis
  - Data Visualization
  - Creating Good Charts

References

Broman, Karl. 2018. "The Top Ten Worst Scientific Graphs." Blog. *Karl Broman*. https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/.

Express Web Desk. 2018. "Twitter Abuzz over BJP's Graph on Fuel Prices, Congress 'Fixes' It." *The Indian Express*.

Pies, Eager. 2013. "Better Than Minard." Blog. *Eager Pies*.

Wikipedia contributors. 2023a. "Hertzsprung–Russell Diagram." *Wikipedia*, April.

———. 2023b. "Principles of Grouping." *Wikipedia*, April.

wtfViz. 2021. "Do You Think Nebraska Should Legalize Marijuana?" *Tumblr*. https://viz.wtf/image/646651837987061760.

———. 2022. "High Support." Tumblr. *Viz.wtf*. https://viz.wtf/post/143173587191/high-support.

Zorzi, Marco, Chiara Barbiero, Andrea Facoetti, Isabella Lonciari, Marco Carrozzi, Marcella Montico, Laura Bravar, Florence George, Catherine Pech-Georgel, and Johannes C. Ziegler. 2012. "Extra-Large Letter Spacing Improves Reading in Dyslexia." *Proceedings of the National Academy of Sciences* 109 (28): 11455–59. https://doi.org/10.1073/pnas.120556609.